Loss-free architectures in optical burst switched networks for a reliable and dynamic optical layer

Thomas Coutelen

LOSS-FREE ARCHITECTURES IN OPTICAL BURST SWITCHED NETWORKS FOR A RELIABLE AND DYNAMIC OPTICAL LAYER

by

Thomas Coutelen

A thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy

Concordia University, Montréal, Québec, Canada

And Télécom SudParis, Evry, France

In accordance with the co-tutelle agreement between Quebec and France

April 2010

THESE n°2010TELE0010

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mr. Thomas Coutelen**

Entitled: **Loss-free Architectures in Optical Burst Switched Networks for a Reliable and Dynamic Optical Layer**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

_____ Chair

George Rouskas _____ External Examiner

Annie Gravey _____ External Examiner

Monique Becker _____ Examiner

Dongyu Qiu _____ Examiner

Anjali Agarwal _____ Examiner

Clement Lam _____ Examiner

Brigitte Jaumard _____ Supervisor

Gerard Hebuterne _____ Co-supervisor

Approved _____

Chair of Department or Graduate Program Director

_____ 20 _____ _____

Rama Bhat, Ph.D.,ing., FEIC, FCSME, FASME, Interim Dean

Faculty of Engineering and Computer Science

Abstract

Loss-free Architectures in Optical Burst Switched Networks for a Reliable and Dynamic Optical Layer

Thomas Coutelen, Ph.D.

Concordia University

Télécom SudParis, 2010

For the last three decades, the optical fiber has been a quite systematic response to dimensioning issues in the Internet. Originally restricted to long haul networks, the optical network has gradually descended the hierarchy to discard the bottlenecks. In the 90's, Metropolitan networks have been enlightened. Now, optical fibers are deployed in access networks and reach the users.

In a near future, besides wireless access and local area networks, the whole network may be made of fibers, in order to support current services (HDTV) and new applications emergence (3D-TV newly commercialized in USA). The deployment of such greedy applications will initiate an upward upgrade. The first step may be the Metropolitan Area Networks (MANs), not only because of the traffic growth, but also because of the variety of applications served, each with a specific traffic profile. The resulting variability will not be affected through passive optical devices.

The current optical layer is of mitigated efficiency dealing with unforeseen events. The lack of reactivity is mainly due to the slow switching devices: any on-line decision of the optical layer is delayed by the devices configuration. At the MANs enlightenment, lots of efforts have been deployed to improve the reactivity of the optical layer. The Optical Circuit Switching paradigm (OCS) has been improved but it ultimately relies on off-line configuration of the optical devices. Optical Burst Switching (OBS) can be viewed as a highly flexible evolution of OCS, that operates five order of

magnitude faster. Within this architecture, the loss-free guaranty can be abandoned in order to improve the reactivity of the optical layer. Indeed, reliability and reactivity appear as antagonists properties and getting closer to either of them mitigates the other one.

This thesis aims at proposing a solution to achieve reliable transmission over a dynamic optical layer. Focusing on OBS networks, our objective is to solve the contention issue without mitigating the reactivity. After the consideration of contention avoidance mechanisms with routing constraints similar as in OCS networks, we investigate the reactive solutions that intend to solve the contentions. None of the available contention resolution scheme can ensure the 100% efficiency that leads to loss-free transmission. An attractive solution is the recourse to electrical buffering, but it is notoriously disregarded because (1) it may highly impact the delays and (2) loss can occur due to buffer overflows. The efficiency of translucent architectures thus highly depends on the buffer availability, that can be improved by reducing the time spent in the buffers and the contention rate.

We show that traffic grooming can highly reduce the emission delay, and consequently the buffer occupancy. In a first architecture, traffic grooming is enabled by a translucent core node architecture, capable to re-aggregate incoming bursts. The re-aggregation is mandatory to "de-groom" the bursts in the core network. On the one hand, the re-aggregation highly reduces the loss probability, but on the other hand, it absorbs the benefits of traffic grooming. Finally, dynamic access to re-aggregation for contention resolution, despite the significant reduction of the contention rate, dramatically impacts the end-to-end delay and the memory requirement.

We thus propose a second architecture, called CAROBS, that exploits traffic grooming in optical domain. This framework is fully dynamic and can be used jointly with our translucent architecture where re-aggregation. As the (de)grooming operations do not involve re-aggregation, the translucent module can be restricted to contention resolution. As a result, the volume of data submitted to re-aggregation is drastically reduced and loss-free transmission can be reached with the same reactivity, end-to-end delay and memory requirement as a native OBS network.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Motivation and Research Objectives

1.1.1 General Concerns

With the progressive expansion of the Internet and the upgrade of the data rate, the Internet is assigned an increasing position in common life and a crucial role in the global economy. The hierarchical deployment of the network allows gradual upgrades and considerably simplifies the management, both in administrative and technical points of view. Each level of the hierarchy can be seen as a collect network that transmits data from the lower level to the upper one. The choice of the equipment and technologies belongs to the organization in command and is ruled by, e.g., the type of applications, the volume of traffic and its profile and the radius of the network. A classical description of a hierarchy discloses four levels: Local Area Networks (LAN), access networks, Metropolitan Area Networks (MAN) and finally long haul backbones (Wide Area Networks, WAN). An access network connects several LANs to a MAN, that supplies an aggregate of the traffic to a backbone. At each level, the dimensions of the networks increase. The radius is typically enlarged and the traffic involved – an aggregate of an increasing number of connection – keeps on growing.

At the top level – the backbone – , transmissions involve very high data rate over long distances.

The optical fiber is particularly well suited to carry such a traffic and has become the medium of choice in long haul networks.

In the 90's, the progressive increase of the number of users and of the data rates disclosed the discontinuity between the copper MANs and the optical WANs: The MANs were not able to feed the WANs and the vast transport capacity at the top level was not exploited. This situation provoked the extension of the optical medium to the MANs, what pushed the discontinuity to the lower hierarchical level. Nowadays, this discontinuity between the access networks and the MANs is highlighted by the sudden increase of the penetration rates and the convergence of communication services through the Internet. Common communication services such as telephony or TV broadcast used to be handled by dedicated networks, but it is only a matter of time before they completely migrate on the Internet because the services may benefit of a significant enhancement. For instance, with VoIP systems, a user is not identified physically, but can be reached on any connected device, not to mention the addition of the video to the voice. Similarly, on the Internet, TV broadcast is only limited by the creativity of the users ... and the transport capacity. Once again, optical fibers have been elected to replace copper links. The deployment of optical fibers in the "first few miles" (referred to as FTTx) is well advanced in Asia and in European and North American metropolises where Internet Service Providers (ISPs) propose domestic solutions reaching up to 10 Gbps, i.e., largely enough for the future domestic applications as, e.g., HDTV, VoD, teleconferencing.

To reduce the capital (CAPEX) and the operational expenditure (OPEX) related to FTTx, passive equipment have been favored. With such architectures, the traffic profile is preserved through the access up to the MAN. Thus, with FTTx, MANs should be more loaded with variable and heterogeneous traffic (resulting from the wide variety of services).

In addition, some services have stringent quality of service constraints. Real time applications are delay constrained, whereas the connection-oriented applications are more impacted by loss rates. In addition, some applications require a strict availability (e-business, tele-work, cloud computing ...). The availability seems to gain in significance because the colossal economical stake raised from

the democratization of the network. For instance, the social networking site FaceBook might earn around 1 Billion$ from advertisement next year according to the "emarketer.com"; This amounts to 100,000$ each hour. Again according to the "emarketer.com", 22 G$ have been spent for advertising on the Internet in 2009 and this amount is expected to grow up to 34G$ until 2014. In this context, service perturbation can severely affect the e-market and the global economy.

1.1.2 The Optical Medium

In the 80's, the optical fiber reached the status of de-facto standard to support mass data transfer in long haul networks. In addition to attractive physical characteristics, it offers a huge transport capacity. During the past 30 years, a lot of efforts have been deployed to exploit this tremendous transport capacity, in spite of physical constraints imposed by the optical medium.

The first concern relies on the discontinuity between the electrical domain and the optical domain: The data rate is limited by the performances of the end-systems, which operate in electrical domain. The overall data rate is increased with Dense Wavelength Division Multiplexing (DWDM): Several flows are simultaneously emitted in the fiber, each of them on a dedicated wavelength. DWDM thus defines independent transmission channels, whose transport capacity depends on the performances of the end-systems. A key advantage is that the transport capacity can be increased from end-systems without requiring civil engineering expenses (up to a point). Another asset lies in the fact that the wavelengths can be demultiplexed and switched in the optical domain, with mirrors. Those optical switches preserve the characteristics of the signal but are slow to configure, and, consequently, imposed circuit oriented transmission. Nowadays, DWDM systems can involve 160 independent channels in a single optical fiber. The transport capacity of a wavelength is related to the number of wavelengths, but systems at 40 Gbps are available. Overall, DWDM systems can transmit up to terabits per seconds within a single fiber. Though DWDM reduces the transport granularity, it remains several times larger than the data rate requested by a flow. Nowadays, this granularity issue is solved with Time Domain Multiplexing (TDM) over DWDM, managed in

electrical domain by SONET/SDH. With recent enhancements, SONET/SDH completely solves the problem of granularity, but it relies on specific equipments that must be installed with caution.

Nevertheless, a more critical concern is related to the circuit establishment latency that severely compromises the reactivity of OCS networks. This poor flexibility has been addressed in WANs for fault management: Extra resources are reserved, possibly pre-configured so that in case of failure, the transmission instantaneously switches on the backup path. This solution avoids service interruption at the cost of extra capacity reservation.

When MANs were enlightened in the 90's, lots of efforts were deployed to improve the flexibility of the optical layer to deal with the variability of the traffic. In particular, the Link Capacity Adjustment Scheme (LCAS) can renegotiate the capacity of a channel, but it does not avoids re-routing, which may severely affect the service in OCS networks.

With the deployment of FTTx, the reactivity of the optical layer is of utmost importance. Packet oriented transmission is a legitimate alternative: Processing and configuration is done on demand. The resulting reactivity and flexibility is of precious help to deal with failures or traffic variations without redundancy. Optical Packet Switching (OPS) attracted a significant interest in the 90's. Transposing classical protocols used in electrical domains toward the optical domain is a challenging task. The main obstacle is the lack of optical memory, required at each node to store the payload during routing procedures. In [DH04], Dr. Hau demonstrates the feasibility of optical memory, but it may take decades to exploit those revolutionary results and deploy commercial optical buffers. Another promising solution can be envisioned with the optical processing. In spite of intensive research, this solution remains far from maturity. Finally, the most advanced prototypes rely on Fiber Delay Lines to delay the packet in the optical domain. However, such an architecture raises several issues, in particular, the dimensioning of FDLs is a crucial and complex task and restricts the flexibility of the traffic engineering.

In the late 90's, Optical Burst Switching has been proposed as a viable alternative to OPS. It is discussed in the next section.

1.1.3 Optical Burst Switching for a Dynamic and Reliable Optical Layer?

In OPS networks, buffering is required between the data arrival and the switch configuration. In Optical Burst Switched networks (OBS), the switches are configured prior the data arrival and core buffering is no more required.

Pre-configuration of the switches is initiated by a header sent prior the data. The header carries the information about the destination, the arrival date and the duration of the packet. At the header arrival, the node looks for the appropriate output port and is scheduled to switch to the corresponding configuration prior the data arrival. This way, the data cut through the nodes all-optically, provided they are preceded by their header. As only the header is delayed in core nodes, it is assigned a processing budget called Offset Time (OT). The OT is the gap between the header emission and the data emission. It has to be larger than the overall header processing time up to the destination, but it should be minimized in order to improve the reactivity of the network. Thus, the data are usually sent without acknowledgment from the control plane.

To reduce the transmission and processing overhead on the control plane, the data are aggregated into so-called bursts. Edge nodes manage several queues among which the ingress traffic is sorted using various criteria. The packets that belong to the same queue are identical from the core network point of view. They will be served the same way. Thus, the data are usually sorted according to their destination, and, possibly their class of service.

After a significant academic infatuation following the birth of OBS, a large part of the community went back to OCS. It may be due to the reluctance of industrials, or from a technical point of view, to the difficulty to exploit the reactivity of OBS while providing transfer guarantees: As the resource reservation is not acknowledged, contentions can occur if an output port is requested by two or more bursts for overlapping periods of time. In the electrical domain, this issue is solved by managing queues, but without buffering, "store and forward" is prohibited and achieving loss-less transfers is the most critical step toward OBS maturity.

Hybrid devices combining slow and fast switching fabrics have been envisioned to support circuit

and burst switching. Slow devices are cheaper and preserve the signal, but they severely compromise the reactivity. On the other hand, the fast devices enable a significant degree of reactivity, but they increase the CAPEX and damage the signal. With such hybrid devices, circuit and burst switching can cohabit to provide either reactive or guaranteed transmission, depending on the application to be served. Clearly, the dimensioning of the switching fabrics is critical since it dictates the capability of burst and circuit switching, while influencing the CAPEX of the network. An alternative to provide polymorphism exploits the high flexibility of OBS. OBS does not impose strict restriction on the size of the bursts or even the signaling protocol. As a result, the control plane deployed in OCS networks can be deployed in OBS networks as well, without discarding burst switching. Thus, loss-less can be guaranteed pro-actively in OBS networks by using loss-less routing configurations of OCS networks.

Our objective in this thesis is to envision a dynamic and reliable optical layer. A worthwhile solution must reach at least the same loss-less throughput as current architectures in stable scenarios, and be only just or not sensitive to traffic variations. It is worth to discuss the reactivity of the loss-free OBS architectures derived from OCS to evaluate the direct benefits of deploying fast switching devices. However, with those solutions, the transfer guarantee depends on routing restriction that mitigate the dynamism of OBS. An ideal solution should ensure loss-free transmission via reactive mechanisms, as independent as possible from off-line decisions and static configurations.

1.2 Contributions and Organization of the Thesis

During this PhD research, we considered two approaches. First, we used routing restrictions inspired from OCS networks. In [CHJ09a], we transpose the classical Routing and Wavelength Assignment problem (RWA) from OCS to OBS networks. The so-called RWA-OBS problem takes advantage of OBS specificities to improve the loss-free multiplexing and, as a consequence, the loss-free throughput of the network. The problem is solved via an ILP model that can be adjusted to describe either an OBS network or advanced OCS architectures. It outputs routing, wavelength and offset time (if relevant) assignment for the granted connections. In [CJH09b], the model is described with a

column generation formulation to improve the scalability and the resolution time. Our experiments demonstrate the benefits of using fast switching devices and OBS transmission. This architecture increases the number of granted connections as compared with the most advanced all-optical OCS solutions.

In practical deployment, the optical networks are not "all-optical". The multiplexing potential of OCS is improved by SONET/SDH, which enables sub-wavelength switching. Multiplexing is operated in electrical domain to ensure a collision-free convergence of several flows. In other words, in opposition to all-optical architectures, flow convergence is allowed in SONET/SDH networks. The converging flows are re-emitted from electrical domain so as to avoid contentions. In [CHJ09c], we proposed a translucent architecture for OBS networks, designed to perform similar tasks. The bursts of conflicting flows are converted into electrical domain and their re-emission is scheduled so as to prevent contention. The shortcoming of translucent architectures lies in their impact on the end-to-end delay and on the resulting processing overhead. The proposed architecture thwarts those effect by exploiting traffic grooming. In [CJH09a], we show that the traffic grooming enabled by our translucent architecture can compensate the increase of the delay caused by the electrical processing.

Thus, the architecture proposed in [CHJ09c] and [CJH09a] can significantly reduce the loss probability without increasing the end-to-end delay. Nevertheless, the access to electrical buffers is set at the emission and the electrical buffering is not a reactive solution to contention. In [CJH10a], we propose the CAROBS transmission scheme that enables dynamic traffic grooming in optical domain. With the resulting reduction of the delay and contention rate, the translucent architecture can be used as a reactive contention resolution mechanism without imposing additional delay or electrical buffering requirements. The framework is further improved in [CJH10b], where loss-less transmission is achieved with basic OBS equipment and with no impact on the end-to-end delay.

The thesis is organized as follows. Chapter 2 presents the current optical network and discusses the possible impact of the deployment of FTTx. Its conclusions address the need for a dynamic

optical layer. Chapter 3 overviews the OCS architecture and confirms the limited flexibility of OCS networks. Chapter 4 describes the fundamentals of OBS and the relevant contributions. In particular, the loss-less solutions for OBS are attentively considered, but all of them compromise the reactivity. In Chapter 5, the RWA-OBS problem is presented and used to compare loss-less OBS networks with the most advanced all-optical OCS architectures. The theoretical observation of OBS core and edge nodes proposed in Chapter 6 motivates the study of traffic grooming in OBS networks. In Chapter 7, we discuss the traffic grooming with translucent architectures and propose to involve the aggregation process in the demultiplexing operations. Experiments confirm the conclusions of Chapter 6. In Chapter 8, we further exploit traffic grooming with CAROBS, which enables all-optical and dynamic traffic grooming. We then trade the benefits of CAROBS with electrical buffering, accessed on demand by the contending bursts. We show that, in translucent mode, CAROBS meets the objectives of the thesis since it preserves the reactivity of OBS, and attains a sufficient reliability without extra expense. Conclusions and further works are finally proposed in Chapter 9

List of Contributions

[CHJ09a] T. Coutelen, G. Hébuterne, and B. Jaumard. An OBS RWA Formulation for Asynchronous Loss-less Transfer in OBS Networks. In *Proceedings of HPSR*, 2009.

[CHJ09b] T. Coutelen, G. Hébuterne, and B. Jaumard. Core OBS Traffic Properties and Behavior. *Les cahiers du Gerad*, Mai 2009.

[CHJ09c] T. Coutelen, G. Hébuterne, and B. Jaumard. Is It Worth to Keep an All Optical OBS Data Plane? In *Proceedings of CCECE*, 2009.

[CJH09a] T. Coutelen, B. Jaumard, and G. Hébuterne. A Translucent OBS Node Architecture to Improve Traffic Emission and Loss Probability. In *Proceedings of the IEEE WOBS*, 2009.

[CJH09b] T. Coutelen, B. Jaumard, and G. Hébuterne. Improving RWA-OBS Formulation and Solution. In *Proceedings of the IEEE WOBS*, 2009.

[CJH10a] T. Coutelen, B. Jaumard, and G. Hébuterne. A Viable Translucent Architecture for Lossless OBS Networks. In *to appear in Proceedings of the IEEE ICC*, 2010.

[CJH10b] T. Coutelen, B. Jaumard, and G. Hébuterne. An Enhanced Train Assembly Policy for Loss-less OBS with CAROBS. In *To appear in Proceedings of the IEEE CNSR*, 2010.

CHAPTER 2

Background and Motivation

The optical fiber becomes the medium of choice all over the Internet. After its deployment in long haul national backbones and metropolitan networks, operators are currently reaching the users with this medium in order to provide various services, including common ones such as telephony and HDTV, to a large number of users.

In this Chapter, we present the optical network as expected in the near future. First, we draw an overview of optical components and of fundamental multiplexing techniques. Then, we describe the traffic to highlight its evolution toward heterogeneousness, both in terms of profiles and requirements. The presentation of a typical optical network hierarchy leads to a discussion of the possible impact of the enlightenment of the access networks. Those conclusions pledge for the necessity of a dynamic optical layer. Optical switching is then discussed. We highlight the limitations of the current technology in dynamic scenarios and describe the potential of alternate paradigms. A qualitative comparison of optical switching paradigms finally leads to the explicit definition of the objectives of this thesis.

2.1 Optical Network Basics

2.1.1 Optical Fibers

The optical fiber won recognition as the best medium to carry high speed connections over long distances. In the optical domain, the information is coded by the signal strength of a time-slot: The light-wave is segmented in "slots", the length of which depends on the transceiver clock. Intensity multiplexing is performed at each "slot", e.g., if four levels of intensity can be distinguished, then two bits are coded by each time-slot. The upgrade of end-systems can increase the degree of multiplexing and consequently increase the transport capacity of a fiber without additional cost for civil engineering.

The optical fibers are either single-mode or multi-mode. In multi-mode fibers, the signal propagates through different paths at different velocities. The difference of velocities, called multi-mode dispersion, limits the effective transmission length and regenerators have to be installed every few tens of kilometers. Multi-mode fibers have been deployed in early telecommunications. They allow the use of cheaper amplifiers, but the multi-mode dispersion and the fact that they operate at low bit-rate (up to 140 Mbps [Kar00]) restrict their deployment to small areas such as buildings or schools.

Single-mode fibers (SMF) only have a single propagation mode and thus eliminate the multi-mode dispersion. The transmission length is thus significantly increased and single-mode fibers can operate at 40 Gbps or more and up to several hundreds of kilometers without regeneration [Kar00]. According to [OMKB09], the chromatic dispersion – the main impacting phenomenon – is significantly reduced by Non Zero Dispersion Shifted Fibers (NZDSF). To further struggle with chromatic dispersion, compensation modules can be installed along some fibers.

In [OMKB09], the quality of the signal is evaluated with regard to various equipment and physical phenomena. Figure 2.1 shows the quality of the signal versus the distance it traveled, both without (Figure 2.1(a)) and with (Figure 2.1(b)) chromatic compensation modules.

Operators quickly agreed to deploy the optical fiber in the backbone and then closer and closer to

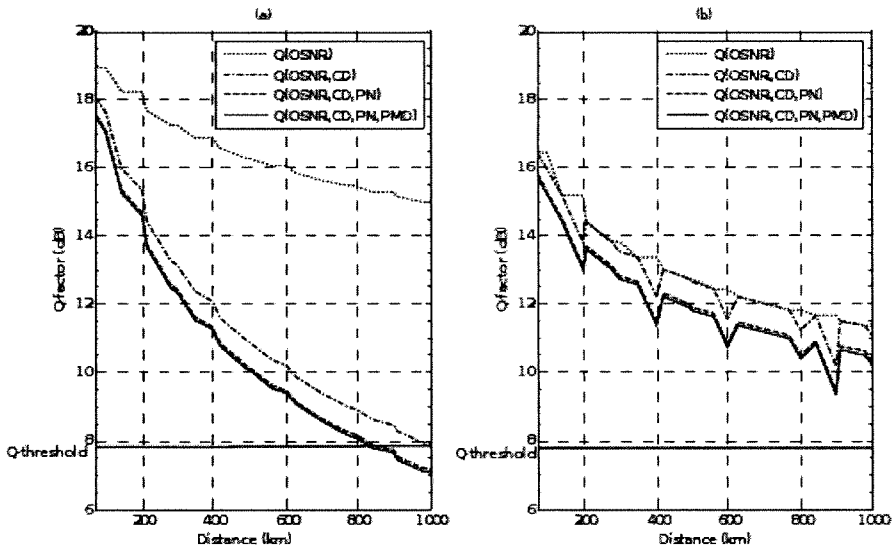


Figure 2.1: Signal Impairment (taken from [OMKB09])

the user. Such deployment favored the emergence of greedy applications, resulting in a considerable increase of load of each flow. Nevertheless, it remains far from the transport capacity. This addresses the need for multiplexing.

2.1.2 Multiplexing in Optical Fibers

The total capacity of a fiber is most likely much higher than the need of a connection between two nodes. So the way the available bit-rate is allocated is an important issue, and this introduces the concept of "granularity": The minimum amount of bandwidth which can be allocated to a single connection. In optical networks, making use of circuit switching paradigm, the granularity results of the way the different flows are multiplexed. This is performed in time domain (Time Division Multiplexing - TDM) and in spectral domain (Wavelength Division Multiplexing - WDM) .

Wavelength Division Multiplexing

Dense Wavelength Division Multiplexing (DWDM) splits the bandwidth of a fiber among numerous independent wavelengths. The number of wavelength depends on the precision of the emitters and receivers and on the distance to travel: Increasing the number of wavelengths decreases their signal

strength and thus their distance range (according to [Bat02], systems providing 82 wavelengths at 40Gbps cannot reach more than 300km).

74 THz of bandwidth can theoretically be used for multiplexing, but only a restricted range is used in practice: The C-band (4 THz wide: From 1530 nm to 1565 nm) has been favored since it offers the lowest attenuation ratio (between $0.17dB.km^{-1}$ and $0.4dB.km^{-1}$ at 2.5 Gbps). Such low attenuation ratio provides a large distance range limited by the receiver precision: They must be able to differentiate the slot intensity levels. This task is harder as the signal strength decreases. Consequently, increasing the time-slot capacity decreases the maximum distance reached by the signal (according to [SS06a], information on a single wavelength at $2.5Gbps$ (resp. $10Gbps$, resp. $40Gbps$) reaches $4000km$ (resp. $3000km$, resp. $2000km$) without repeater).

Today, commercial solutions can differentiate up to 160 wavelengths, whereas a wavelength typically operates at 2.5, 10 or 40 Gbps, for an overall fiber capacity close to terabits per seconds.

SONET and Time Division Multiplexing

With WDM, the transport granularity is reduced at the level of the wavelength capacity. It however remains pretty coarse and WDM is typically combined with Time Division Multiplexing (TDM). TDM is the most intuitive multiplexing flavor. It has been widely used in most operator networks during the past 30 years: The payload is segmented in constant-size blocks, sent at regular intervals. The major concerns are the framing of the payload and synchronization of the blocks. In the context of optical networks, TDM is constrained by the "electro-optical discontinuity": The signal carries the aggregate of multiple clients and the nodes must operate at the aggregate data rate, which must be kept lower than electronic emitting, receiving and processing technology. This is a major concern that led to the standardization of the Synchronous Digital Hierarchy (SDH). SDH has been defined by the European Telecommunications Standards Institute (ETSI) and is used worldwide but in North America, where the Synchronous Optical Networking (SONET, American National Standards Institute) is used. SONET has been defined soon before SDH, but the worldwide market penetration of SDH promotes it as the standard.

SDH Frame Format	SONET Frame Format	Optical Signal	Bit Rate (Mbps)
STM-0	STS-1	OC-1	51.84
STM-1	STS-3	OC-3	155.52
STM-3	STS-9	OC-9	466.56
STM-4	STS-12	OC-12	622.08
STM-6	STS-18	OC-18	933.12
STM-8	STS-24	OC-24	1244.16
STM-9	STS-36	OC-36	1866.24
STM-16	STS-48	OC-48	2488.32
STM-64	STS-192	OC-192	9953.28

Table 1: SONET Hierarchy

The synchronous transport signal level-1 (STS-1) is the basic building block (51.85 Mbps) and a higher-level signal in the hierarchy is obtained by combining signals from the lower-level component signals. It uses the term tributary to refer to the component streams that are multiplexed together. Table 1 shows the SONET hierarchy and Figure 2.2 shows how a SONET multiplexer can handle a wide range of tributary types. A slow-speed mapping function allows DS1, DS2, and CEPT-1 signals to be combined into an STS-1 signal. Mapping of ATM streams, DS-3 and CEPT-4 has also been defined. Once all incoming streams are mapped into STS-n signals, they can be combined into a higher-order STS-n' signal.

Although SONET considerably reduces the transport granularity, it still imposes a fairly coarse granularity. It can be further reduced with the Virtual Concatenation (VCAT [Mim02]). VCAT breaks the bandwidth into smaller individual containers grouped logically in a channel. It can create custom-sized SONET channels composed of STS-1 (51.84 Mbps), VT-1.5 (1.6 Mbps) or VT-2 (2.176 Mbps). The virtual tributary rate is designated by STS-m-nv, defining a rate of n STS-m. Lower order virtual concatenations are designated similarly by VT-m-nv. Members of a channel are routed transparently through the network and all the intelligence regarding the virtual concatenation is located at the endpoints of the connection. Thus, VCAT can be deployed on the existing SONET/SDH infrastructure with a simple upgrade of the endpoints and each connection is assigned the right amount of bandwidth, and the effect of the coarse granularity of circuit switching is decreased.

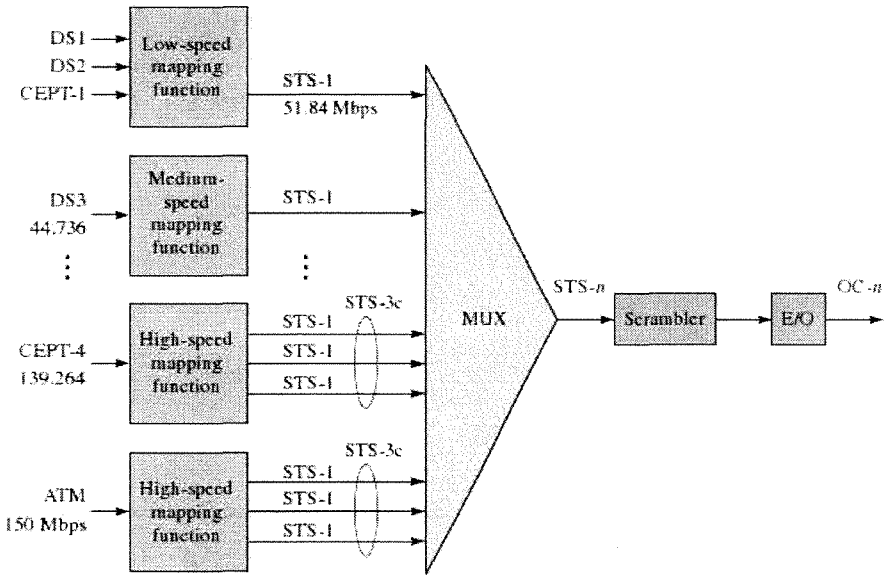


Figure 2.2: SONET Multiplexing (taken from [LGW01]

2.1.3 Basic Equipment

Transmitters

The optical signal is emitted by laser emitters, that translate an electrical information to an optical signal. Emitters can be either static (they are built to emit on a specific wavelength) or tunable (the wavelength can be arbitrarily chosen among a given set).

At the receiver side, photo-detectors perform the opposite function: They convert the optical signal to electrical domain.

Wavelength Multiplexer/Demultiplexer

A wavelength multiplexer combines different wavelengths into a composite signal. On the opposite, wavelength demultiplexers split a composite signal into its channel constituents. Multiplexers and demultiplexers can be seen as similar devices operating in opposite directions. They are build by combining couplers and optical filters.

A coupler can combine several signals into a single fiber (combiner), or inversely split a signal among several fiber (splitter). A 2x2 coupler is composed of a 1x2 splitter and a 2x1 combiner.

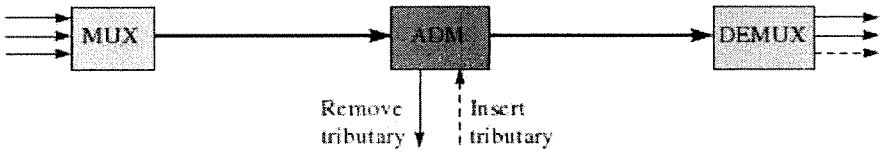


Figure 2.3: A SONET add/drop Multiplexer

Active couplers achieve their task via O/E/O regenerators whereas passive ones are all-optical.

Optical filters can select a specified wave-band by absorbing several specified wave-bands. They are either static or tunable. In the latter case, they can operate on variable wave-bands. In that case, important characteristics are the tuning time (wavelength processing time) and tuning range (the wavelength range the filter can access).

Some demultiplexers are able to extract a group of adjacent wavelengths. The resulting waveband switching is very attractive to reduce the cost of the switching devices. Nevertheless, grouping the wavelengths in the switching process affects the flexibility and the provisioning capacity (see [GWM02] for more details).

2.1.4 Switching Equipment

Optical Add/Drop Multiplexers

OADMs are used to discard wavelengths that terminate in a node (drop) or to insert wavelengths that start in a node (add), whereas other wavelengths only transit through the node. The OADMs can either be static (meaning that the add and drop operations are set once for all) or they can rely on tunable equipment to be reconfigurable (ROADM, [Nor06]).

Add/drop multiplexers operate at the wavelength level, but they can be combined with an electrical module in order to process the signal at the SONET level. Figure 2.3 illustrates a SONET add/drop multiplexer: The signal is converted into electrical domain and tributaries can be removed or inserted.

Optical Cross-Connect

In complex mesh topologies, (R)OADM are replaced with more flexible equipment: The Optical Cross Connects (OxC).

In OxCs, the signal is switched at the wavelength level: The wavelengths are demultiplexed, directed from input ports to output ports by a set of mirrors (Micro-Electro Mechanical Systems - MEMS), and then multiplexed in the outgoing fiber. The use of mirrors preserves the signal characteristics, but the mechanical nature of the system slows the re-configuration (50 ms according to [JV05]).

Such an architecture is transparent since the signal travels all optically up to its destination. This approach possibly mitigates the resource utilization because of the gap between request granularity and wavelength transport capacity: If a request does not fill a wavelength, then part of its capacity is wasted. Switching at a sub-wavelength granularity is mandatory to fill the gap between user request granularity and wavelength granularity by grooming several requests onto a single wavelength. Such multiplexing uses TDM and imposes the signal to be switched at a bit or packet level. Such treatment must be done in the electrical domain. For this flexibility, opaque architectures are attractive, but highly mitigating the optical transmission benefits.
Current networks are called translucent since opaque operations are performed if and only if sub-wavelength switching operations are required, while transparent switching is performed for wavelength (or higher) granularity.

Multi Service Provisioning Platform (MSPP)

The optical layer is independent of the semantic of the signal. A wavelength can transport several connections, typically multiplexed in TDM with SONET/SDH. Manipulating SONET/SDH streams mandates electrical treatments and is usually achieved via MSPPs [HJJ05]: The incoming signals are demultiplexed and a sub-set of wavelengths can cut-through the OxC, whereas the other wavelengths are converted to electrical domain. At that point, a connection can be inserted in the stream ("add")

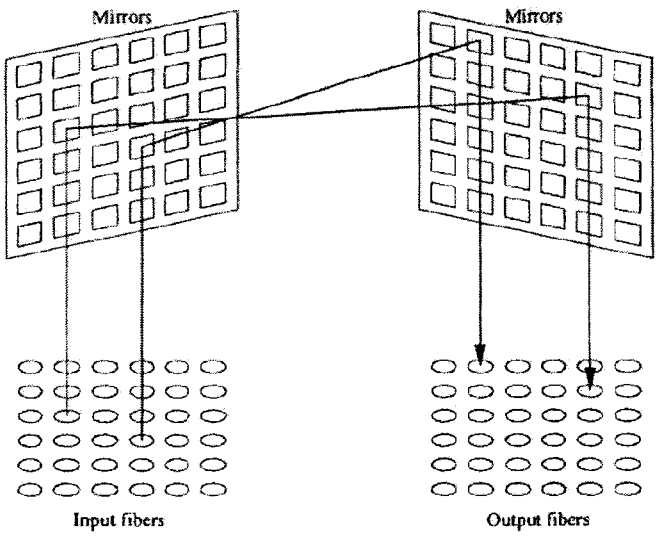


Figure 2.4: MEMS Switches (taken from [JV05])

or extracted from the stream ("drop").

The cost of an MSPP is mainly imposed by the number of wavelengths connected to the grooming factory and network design with such devices typically aims to minimize it.

Light-trail Access Unit

Under some circumstances, traffic grooming can be done in optical domain. This scheme is referred to as light-trails [GC03a]. A Light-trail Access Unit (LAU) has a similar function as an add/drop multiplexer, except that the incoming signal can "drop and continue". A possible architecture is described in [VBMS05] and represented in Figure 2.5. A fraction of the signal power is deflected to the local node, whereas the remaining signal cuts through the node or can be blocked, depending on the configuration of a traversed shutter. The LAU can also emit a signal on the medium, under the condition that it is free.

Due to the transparency of this architecture, it has a lower impact on the network cost as compared with opaque modules.

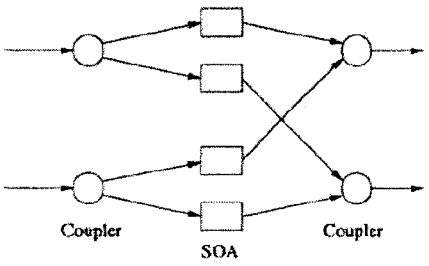


Figure 2.6: SOA Switches (taken from [JV05])

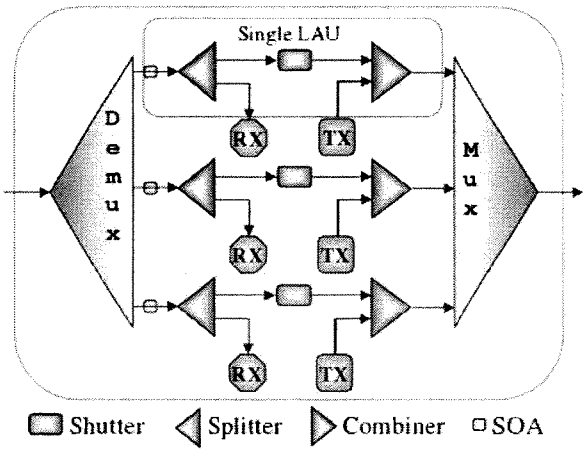


Figure 2.5: Light-trail Access Unit (taken from [VBMS05])

Semi-Optical Amplifiers Switches

A basic SOA switch is illustrated in Figure 2.6. An input port is connected to every output port through a tree of couplers. The signal is broadcast towards each output port, and a SOA is placed after each coupler. SOAs act as gates that can block the signal, or let it go. Such an architecture can switch configuration as fast as in one nano-second and enables multi-cast. A possible limitation is that the couplers reduce the strength of the signal.

2.2 Traffic Profile

High speed solutions progressively changed the relation between the users and the network. A user can now access various content and services from a single device via the network. The e-mail has replaced postal services ; websites supplant news-papers ; TV broadcast channels recourse to live streaming in order to increase the persistence and the diffusion of their content ; and the Internet provides a wide potential to improve the telephony, though VoIP solutions are not widely exploited yet.

In [EGH+09], the authors report the observation of the traffic of 100K users between 2001 and 2008. The most significant applications on the overall traffic are the web browsing, the web streaming and the peer-to-peer file sharing. The report confirms the overall increase of the traffic between 2001 and 2008.

The traffic profile – the emission process – and the user expectations are specific to each service. The heterogeneousness of the traffic addresses a serious challenge to the operators, which are doomed to fulfill various requirments to meet the expectations of their customers.

In this section, we describe the variety of service requirements and the heterogeneousness of the traffic involved by the multiplication of services.

2.2.1 Quality of Service Parameters

Most experts believe that the volume of traffic should evolve according to a kind of Moore's curve, firstly because of the number of users, secondly because of the increasing requirements of the applications. Beside this, each service requires a specific support: The success of a session is indeed conditional to the fulfillment of several constraints enumerated in the Service Level Agreement (SLA) between users and operators. The SLA mainly relies to the following parameters:

1. Loss Rate: Each application has its own sensibility to loss. For example, multimedia connection may allow a certain probability of loss. The session can either deduce the missing data or degrade the quality before being interrupted if the loss rate overtakes the ceiling limit.

2. Delay: Some applications are not sensitive to delay (up to a point). For example, the delay encountered by a file transfer must be minimized, but has no impact on the success of the transfer. On the opposite, for real time applications, the delay is critical. For instance, large delay in voice communications could imply talking overlaps, leading to misunderstandings. Real-time applications usually assign a delay budget to data transfer and the data are assumed to be lost once the delay expires, with the same impact as loss on the quality of the service.

3. Jitter: In the case of multimedia connections, the signal, if segmented into packets, must be easily reconstructed; out of order packets and large inter-packet delay variations must be minimized or even avoided to simplify the re-construction process. If the data arrive too late, it may not be included in the resulting signal, with the same consequences as packet loss from the user point of view.

4. Availability: The availability is the ability to satisfy the SLA over the time. Some applications may not tolerate any interruption (telemedecine, telework) while others may accommodate short duration interruption. For example, a VoIP connection is assumed to be more availability constrained than a file transfer. Note, however, that service perturbation of best effort sessions may highly impact businesses although it is not specifically included in the SLA.

2.2.2 Traffic Model

From the user point of view, a flow denotes the traffic the currently used applications emits. It is common to use simultaneously multiple services, e.g., file sharing, web browsing and VoIP. Each application has its own way to supply traffic. The traffic as seen in the network is a super-imposition of user flows, that can be aggregated at some points.

In this section, we will first describe the traffic from the application point of view, i.e., as emitted by the end user. Then, we will describe the impact of flow super-imposition and flow aggregation.

Application Traffic

The traffic emitted by an application can be characterized by many parameters including the data rate, the size of the packets and their emission distribution, the duration of the connection.

Depending on the applications, it is relevant to consider either the average, minimum, or peak (i.e., maximum) data-rate. The gap between the peak and the average rate suggests traffic variability. The variability can be caused by the emission process of the application, or by external events related, e.g., to congestion controllers or hardware malfunctions. that can be reproduced with ON/OFF processes. The process models a source that alternates between states active and idle. We denote α and β the average durations of active periods and idle periods. At any instant, the probability for a source to be active is $\mathcal{A} = \frac{\alpha}{\beta+\alpha}$, in which case it emits traffic at a rate $\widehat{a}$. Otherwise, the source is void. The average data-rate of the source is $a = \widehat{a} \times \mathcal{A}$. The burstiness factor, which reflects the variability of the traffic, can be defined by $\mathcal{B} = \widehat{a}/a = 1/\mathcal{A} \in [1, +\infty[$.

Flow Traffic

From the network point of view, the flow between two nodes is an aggregate of a number of connections. Therein, the traffic profile is defined by the joint activity of all involved connections.

An important factor is the aggregation degree of the flow, defined as the number of connections it multiplexes. If a small number of applications are multiplexed, then the behavior of each of them significantly impacts the profile of the overall traffic. Increasing the aggregation degree masks the behavior of individual flows. The variation of the aggregation degree is ruled by the duration of the sessions, which, in turn, depends on the type of applications involved. For instance, peer-to-peer transfers have a long life-time whereas streaming or web services are more sporadic. Thus, a flow can be modeled by a super-imposition of ON/OFF processes, active during the life-time of a their session.

Finally, the traffic can be modeled in different time-scales by similar processes. This property, called "self-similarity" [GS08], is largely observed in the network. For instance, the network is

stressed during business hours (though the peak is reached around 8 PM [MFPA09]) and reflects the activity of a population over a day. Monitoring the traffic on a lower time-scale discloses similar variations reflecting the behavior of a set of multiplexed sessions, each handling a number of applications.

A much simpler model follows the Poisson distribution. A Poisson process describes the traffic emitted by a theoretical system composed of an infinity of sources, each having an infinitesimal contribution. Though that can reflect fairly accurately some real systems ([WM00],[CCL+02]), they are often used because they drastically simplify theoretical approaches, or when the specific behavior of the applications is neglected.

2.2.3 Heterogeneousness Illustration

Let us illustrate the heterogeneousness of the traffic through the description found in [Mai08] of two antagonist applications which become increasingly popular: Online gaming and peer-to-peer file exchanges.

Online Gaming

Popular online games operate in client-sever mode, where a server keeps a global view of the state of the game and periodically broadcasts this "picture" to the clients. Between broadcasts, the clients notify the server of the actions of the user so that the server keeps a consistent state of the game. The fluidity of the game calls for a low latency and a minimum loss probability. The user tolerance to state refresh frequency (30 updates per second would ensure visual fluidity) imposes periodical update every 50 to 100 ms (10 to 20 frames per second). This constraint dictates the use of small packets in order to avoid aggregation latency (nearly constant packet of 40 bytes downstream and more variable packets between 0 and 300 bytes upstream) and bans retransmissions (and consequently data drop). Finally, the service requires a high reliability since, in some cases, even short interruption – few seconds – may irreversibly impact the outcome of the game.

Peer-to-peer File Sharing

As opposed to online gaming, peer-to-peer file sharing (P2P) involves high data-rate of nearly constant traffic. With peer-to-peer applications, the content is duplicated and shared by several users. The protocols first discover "servents" that share the requested file. Then, each of them transfers part of the file to the servent that requests the file.

P2P is used to transfer large files, the transfer can take hours, or even days. As a result, it entails long duration transfers of high-data rate, independent of the time or weekday. P2P transfers often involve nearby peers, because popular requests are largely duplicated, and because people from the same area – for cultural or linguistic reasons – more likely share similar contents. Here, the only performance metric is the duration of the transfer, which should be minimized, but is not critical. Lost packet can thus be retransmitted and, in case of interruption, the transfer can continue from reachable hosts, or, in the worst case, it can resume once the service is restored.

Traffic Taxonomy

A traffic taxonomy describes the traffic of applications via a set of characteristics. Table 2.2.3 reports a possible taxonomy for online gaming, VoIP and file transfer. In practice, handling each service specifically would considerably complicate the network architecture and management. Therefore, operators rely to service taxonomies to define a reduced number of classes of service and to map the applications to the most appropriate class. Classes of service definition raised many and long debates. It is a widely open and a complex issue far away from the scope of this study.

We will mainly focus on the loss probability and consider the end-to-end delay as a secondary metric. Indeed, data losses have a similar impact as delay (or jitter) violations. In delay sensitive applications, late (or out of order) data are usually considered as lost, whereas a dropped packet cannot be retransmitted and fulfill delay and jitter requirements. Thus, delay sensitive applications are also loss sensitive (with a given tolerance).

The service availability is also an important metric. The network must be endowed with failure

management mechanisms, not only to ensure the service of constrained applications such as telework or online gaming, but also because link failure may entail significant side-effects and turn the network into congestion, jeopardizing the service for a large number of users.

Application	Burstiness	Latency	Real-time	Loss Tolerance	Persistence	Impact of Interruption
Online gaming	High	Low	Yes	Very low	Hours	Fatal
VoIP	Medium	Low	Yes	Low	Minutes	Fatal
File Transfer	Low	High	No	High	Days	Negleagible

Table 2: An Example of Application Taxonomy

2.2.4 Conclusion on the Traffic Profile

The characterization of the traffic in the network is a difficult task, and it is getting more complicated with the increased number of applications. The flow, as seen by the network, is composed of traffic generated by numerous applications, each with different emission process and service requirements. The variability of the traffic increases when a flow multiplexes such heterogeneous processes, but this behavior is reduced when the number of multiplexed processes increases.

The number of processes multiplexed in a given flow can widely fluctuate depending on the location of the flow and the type of applications. Nevertheless, at some places in the network, several flows can be aggregated, i.e., re-arranged (in electrical domain) and re-emitted as a single flow. This resulting flow has a completely redefined profile, only slightly linked with the original profile, but rather defined by the emission process after aggregation. The network is composed of many interconnected sub-networks, independent from each other and aggregation is usually performed in interconnection nodes in order to fit in each sub-network architecture. Thus, the traffic goes through several aggregation processes. In the next section, we describe the hierarchical structure of the network and the evolution of the traffic at each level.

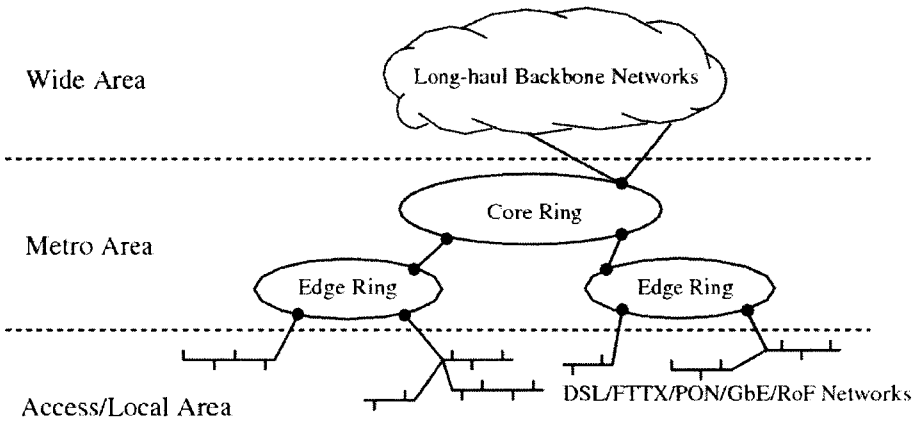


Figure 2.7: Illustration of the Optical Network Hierarchy (taken from [Mai08])

2.3 Optical Network Hierarchy

Communication networks are managed independently by numerous of organizations. Nodal equipment, transport medium, topology and protocol stack are the operators decisions, but those decisions are ruled by several parameters such as the types of service supported, the volume of traffic involved, its repartition, geographical and economical issues to name a few.

The networks are hierarchically interconnected. This architecture is motivated by various reasons. From a technical point of view, it reduces the management complexity (e.g., the size of the routing tables) and improves the robustness of the network. In addition, technological migrations or updates can be done gradually, what drastically reduces the probability of collapse. Finally, from an administrative point of view, it avoids interference between governments or operators.

At each level, the traffic is an aggregate of traffic from several lower level networks. The optical fiber was first deployed in long haul backbone networks to accommodate the massive capacity requirement emanating from successive aggregation of traffic. However, the discontinuity between copper and optical networks is a clear bottleneck that limits the access to optical networks. Over the past 20 years, the operators progressively pushed this discontinuity toward users and, from now on, the optical fiber is progressively deployed in the access networks and is reasonably expected to reach a large number of users in the near future. The resulting topology is illustrated in Figure 2.7:

A user emits its traffic on optical fibers up to the MAN, where several connections are aggregated and transmitted to the backbone, so as to reach another MAN, up to the end users. Let us describe each level of the hierarchy.

2.3.1 Optical Access Networks

The access network represents the first few miles of the transmission (FTTx). It links up to hundreds of users to the MAN. The access network is probably the most heterogeneous part of the network. Currently, Digital Subscriber Lines (DSL) and cable modems are still used, but they are unable to handle the next generation of services such as HDTV. Indeed, access networks slow down the complete migration of some services and limit the access to the optical network.

This bottleneck is progressively removed by the deployment of optical fibers between the MAN and the user. To reduce the deployment cost, passive components are preferred to prevent the access from sophisticated switching procedures and often implies a tree based topology. This is the current solution and the assumption for our work. Note that mainly three options are available to the providers, based on Ethernet (EPON), ATM (BPON), or circuits (WDM-PON). The last option appears as the most promising in the near future, mainly because of the transparency of the data plane that avoids processing in the end nodes.

As the access network does not involve more than hundreds of users per access point, the traffic profile is marked by the behavior of each user. Thus, the traffic injected in the access networks is expected to become more and more heterogeneous and variable. Because of the passive nature of the optical access networks, this heterogeneousness may be preserved along the access networks, up to the MANs.

2.3.2 Metropolitan Area Network

Current MANs have been deployed as collect networks to feed WANs with the traffic of the access networks. However, some of the emerging applications may use MANs as top level networks. For

example, local news or local advertising in the case of IPTV, local phone calls in the case of VoIP ... do not need backbone resources. Despite the democratization of the communication, the social ring of people remains in a limited radius, leading to balance the utilization of the MANs between collect networks and top level networks. The part of this "local" traffic depends on several parameters, mostly the considered applications and size of MANs.

Although the ring topology has first been favored, it discloses shortcomings. First, it is unable to recover from several simultaneous failures. Second, it mandates all nodes to be upgraded simultaneously. In addition, the MANs currently needs more bandwidth, more flexibility and an "any to any" connectivity. Consequently, alternate topologies have been considered. As the MAN radius is relatively small, civil engineering costs for redesigning the topology, though significant, remain affordable. The RINGOSTAR appears as particularly attractive: Some nodes of the ring are upgraded and interconnected by a star subnetwork.

Although the traffic is more aggregated in a MAN than in an access network, it remains bursty and dynamic because of its proximity to the users. This is particularly true with a passive optical access network since, as discussed in previous section, the first few miles do not mix the sessions. In addition, it reflects the human activity (working hours, sleep hours and days off can be identified in the MAN) and thus follows daily patterns. Another particularity of the MANs is the fact that, since they operate as collect networks, their WAN access nodes are expected to be more loaded than the other nodes.

2.3.3 Backbone Networks

The optical fiber was initially restricted to long haul transmission of large volumes. The deployment and upgrade is expensive because of the civil engineering costs and thus, the dimensioning is crucial. Radius and transport capacity are large and must be designed carefully. Newest technologies are quickly included in the WAN to continuously increase the transport capacity (up to 150 OC-48 wavelength, or 40 OC-192 wavelengths). The role of the WAN is to interconnect MANs over a

country or a continent and thus irregular mesh topology is unavoidable.

The aggregation degree is very large in the WAN since it serves a large number of users, whom contribution to the overall traffic is low. In such a context, the human activity is less representative of the traffic profile. Stability over different time frames is expected (low burstiness, low dynamism) and long term planning strategies are relevant is this context.

2.3.4 Conclusion on Hierarchical Optical Network

As described in this thesis, network upgrades appear to be performed sequentially and independently. At each step of this evolution, the bottleneck hierarchical level of the network is enlightened and moved toward adjacent levels. In the 90's, the optical network has been extended in MANs and today, the bottleneck is located in the access networks. The current deployment of Passive Optical Network (PON) between the user and the MAN provides a tremendous capacity to each user, so that its connection can support various applications with specific requirements. With passive equipment the first few miles will supply a none aggregated traffic the MAN. This traffic will be a super-imposition of heterogeneous emission processes, each constrained by specific QoS criteria.

The need for a dynamic optical layer in the MAN has already been raised a decade ago, following observations that the resource utilization was strongly limited by the dynamism of MANs. In retrospect, despite the low resource utilization, the enlightenment of the MANs was appropriate. But the load was limited by the access networks, the traffic was somehow shaped, less services were involved and MANs were highly over-dimensioned. The current deployment of passive optical access networks will reduce the traffic shaping, while the multiplication of the services will increase the variability of the native traffic and the overall load. Within this context, the deployment of a dynamic optical layer appears more critical than 15 years ago.

In the next section, we briefly describe three optical switching paradigms and their potential in dynamic scenarios.

2.4 Optical Switching

Concluding remarks of the previous section justified the design of a dynamic optical layer, at least for the MANs, where the dynamism will highly increase with the massive deployment of FTTx. In this section, we will describe the fundamentals of the Optical Circuit Switching and highlight its limitations in dynamic scenarios. Following the description of two alternate switching paradigms (Optical Burst Switching and Optical Packet Switching), a qualitative comparison of the three paradigms will argue in favor of OBS for the design of a dynamic optical layer.

2.4.1 Optical Circuit Switching

Fundamentals

In Optical Circuit Switched networks (OCS), the resources are reserved to serve a whole connection, i.e., they are reserved from source to destination for the whole duration of the transfer. For multiple reasons, the light-path is not all-optical from the source to the destination, but is rather composed by a sequence of optical segments (OLSP) defined by a couple (λ, R), where λ is a wavelength and R is a set of connected links. The signal remains in the optical domain along an optical segment. It is switched at the wavelength level. As processing and switch configuration is performed off-line, no real-time processing actions or decision are required. Consequently, the network can be deployed with slow reconfiguration MEMS-based devices as, e.g., OxCs.

Because a single flow hardly exploits a whole wavelength, WDM is combined with TDM to allow several slow flows to share a high-speed wavelength. Synchronous Optical Network (SONET) has been standardized in 1989. It took five years to design this protocol so that it can carry multiple various flows within a single structure. Flows are groomed/degroomed in electrical domain, between two optical segments. The flows are re-arranged and the signal, composed of multiple flows, is re-generated on a pre-defined output port (the wavelength continuity is relaxed between two segments). With SONET, the grooming improves the provisioning flexibility, the resource utilization and the throughput.

Such functionality is offered by SONET add/drop multiplexers, or by MSPPs that interface various protocols. Those equipment significantly increase the capital expenditure (CAPEX) and they should be installed carefully so as to reduce their number. Then, once installed, they should be economically accessed in order to limit the impact of electrical processing on the end-to-end delay. A widely investigated problem is the design of SONET networks with the objective to reduce the number of translucent devices for a given traffic matrix (e.g., [JJ05]).

The main advantage of OCS comes from the pre-configuration of the network. As everything is done off-line, it can operate with slow switches. The typical equipment (OxC) are based on MEMS and are configured in milliseconds. In addition, the routing and wavelength assignment (RWA) is designed to avoid conflicts and OCS can respect stringent loss and delay constraints.

Limitations

However, off-line configuration also entails the main shortcomings of OCS. In OCS networks, any variation of a given configuration imposes transport and processing overhead and re-configuration is subject to a non-negligible latency. For example, establishing a connection takes at least the round-trip delay and the milliseconds required for the configuration of the switches, not taking into account the route computation. Thus, on demand configurations should be avoided. To ensure the stability of the service, the operator must over-provision. Two classical scenarios must be considered: Traffic variation and failure recovery.

In case of failure, a reactive approach, consisting in re-routing the connections touched by the failure, does not ensure the service survivability and however perturbs the service during the re-configuration phase. A more reliable technique consists in pre-allocating backup resources so that an alternative light-path can ensure the service instantaneously when the failure occurs. This solution relies on provisioning redundancy and a wide set of contributions can be found to ensure full protection with a minimum redundancy.

The reconfiguration may also be requested in case of dynamic scenarios, to establish a light-path between two end-nodes that were not logically connected. In practice, however, this task

usually boils down to re-negotiate the capacity of an existing light-path. Link Capacity Adjustment (LCAS,[LCA06]) has been designed to be integrated in the SONET frames to allow this task to be achieved transparently with a limited impact on the services. However, it remains a non-negligible latency that makes the technique inefficient for fast traffic variation. In that case, the connections should be served at their peak rate, though it limits the resource utilization.

A promising alternative, the light-trails, has been proposed in [GC03a]. The light-trails are similar to light-paths: They are established exactly the same way. Indeed, they differ in the access to the medium. Whereas SONET defines carriers between two nodes, whose access is ruled by a synchronous access to dedicated time-slots, a light-trail can be accessed by any of its node, and the signal can be received by any downstream node. Thus, the establishment of a light-trail logically connects all its nodes to all its downstream nodes. The connections time share the resources asynchronously. As a result, any node can emit on the medium as long as it is free. This architecture can handle traffic variations without any type of reconfigurations, as long as light-trail establishment is not required. The only limitations rely on the medium access control that must avoid contentions and light-trail establishment latency. The main restriction of light-trails is that only local "add" ports can insert a signal on a transiting flow. Consequently, merging two signals in transit still involves SONET, but with an appropriate routing strategy, the number of electrical devices and conversion can be drastically reduced, as well as the network cost (e.g., [SCD+07, RS07]).

2.4.2 Optical Packet Switching

Optical Packet Switching (OPS) is somehow the extreme opposite case of OCS. Whereas OCS relies on off-line configuration, in OPS, processing and switch configurations are done at the payload arrival. The data are arranged in packets sent independently with a control header that contains relevant information on the packet. Typically, the header will contain the size of the packet, its destination and possibly its class of services. The packets are routed from one node to the next up to their destination. When receiving a packet, an intermediate router retrieves information

from the header so as to compute the appropriate output port and the packet is forwarded on this port. To extract the information from the header, the router must convert the optical signal into electrical domain. Encouraging experiments suggest that optical processing of the header may become effective. In [DHL+03], the header can be optically extracted from the payload and its contents can be used to configure a 1×2 switch. In [MLC+07] a label can be read to perform label switching. Those results are encouraging but still, the prototypes are far from maturity.

Without optical processing, the header processing time is significant and the main challenge of OPS architectures is the storage of the payload during the routing procedure. On the one hand, the entire conversion of the packet is not reasonable, because it would overload the electrical plane. On the other hand, there are currently no reliable alternatives in the optical domain. Optical buffering with the same level of flexibility as electrical RAM is hypothetical. Optical buffering has been achieved by Dr. Hau ([DH04]) in laboratory, but it is not reasonable to expect commercial solutions for optical memory before several decades from now. Nowadays, prototypes (e.g., [MZV+02]) use Fiber Delay Lines (FDL) in order to delay the payload during the header treatment. A FDL is a looping optical fiber in order to increase the distance traveled by the signal. FDLs simulate optical memory, but they can store the data for a fixed amount of time. The critical issue is thus the length of the FDL: It must be long enough to cover the routing procedure, but its length is limited by space and temperature constraints. With this solution, the optical infrastructure lacks flexibility since a FDL is linked to header processing time and packet length.

OPS has earned a significant interest in the 90's, when optical fibers began to be used in dynamic contexts. It has been seen as a credible solution to build a dynamic optical layer because of its fine granularity and the absence of signaling. Those characteristics highly increase the reactivity of the optical layer as compared with OCS.

2.4.3 Optical Burst Switching

Optical Burst Switching (OBS) appeared in the late 90's ([QY99]). The idea behind this paradigm is to offer the same granularity and reactivity as OPS, but, similarly to OCS, to rely on edge memory to side step the lack of memory in core networks. Instead of using FDLs to delay the payload during control plane operations, the overall amount of processing time is spent at the edge. In other words, the payload, grouped in so-called bursts is sent some time after the header. The interval between the header departure and the burst departure is called offset time. This is a processing budget allocated to the header. The header is in charge to signal the burst by providing its arrival date and duration so that each intermediate node can set the switch appropriately just before its arrival and for its exact duration. The burst thus cut-through all-optically, progressively catching up the header, that is delayed at each node for conversion and processing issues. To reduce the emission latency, OBS is usually combined with a one-way signalling protocol (e.g., JET). This signaling. combined with the good granularity of OBS offers a good reactivity.

Nevertheless, OBS faces several difficulties. From a feasibility point of view, OBS requires fast switching equipment because each burst must be separated by the switch reconfiguration time. Recourse to OxC (1 ms to reconfigure) is thus prohibited. A viable solution is the use of combined Semi Optical Amplifiers. With such equipment, the reconfiguration time is reduced to few nanoseconds. The problem with those equipment is that they impair the signal strength, and accentuate the problem of signal impairment inherent in optical transmissions. Now, in terms of performances, the contention problem is commonly seen as the most important issue in OBS: The one-way signaling cannot guarantee the resource availability and contention can occur, leading to payload loss. To combat the contention, proactive mechanisms (namely, flow control and load balancing) and reactive mechanisms (namely, Deflection routing, FDL, wavelength conversion) have been considered. Although those mechanisms can be combined to increase the throughput, no guarantee can be given on the loss rate. Consequently, performances for QoS constrained services cannot "a-priori" outperform circuit switching.

Loss-less solutions either simulate circuit switching or rely on acknowledged signaling scheme. Both those approaches mitigate the reactivity of OBS. In [QWL06], a framework (PATON) is proposed to switch circuits transparently at a wavelength or a sub-wavelength level with OBS technology. Switching transparently at the sub-wavelength level is particularly attractive. Bandwidth is shared with Time Division Multiplexing (TDM), but several issues must be addressed. In PATON, a header can reserve resources for several periodic bursts, thus providing circuit oriented transfer at a sub wavelength granularity. To allow the share of several circuits on a common wavelength, the bursts of a given connection must be successfully inserted between busy time slots. In [QWL06], there is a particularly interesting direction where the authors envision the possibility to provide circuit oriented and packet oriented transfers transparently with a single equipment set. Several questions remain open, especially regarding to the sub-wavelength circuit switching: Investigations must be made on the size of the bursts and their inter-arrival in order to maximize the grade of services ; scheduling and routing must be designed to increase the use of resources let idle by previous TDM connections.

OBS enjoyed a large infatuation at its birth. A wide community addressed several issues, but the interest seemed to have reduced in the last few years, mainly because of the contention problem. Recently, the OBS earned a second breath with the announcement of the deployment of a real-size OBS network in Ireland MANs [INT10].

2.4.4 Qualitative Comparison of Optical Switching Paradigms

The conceptual comparison of the optical switching paradigms boils down to the opposition between packets (OBS and OPS) and circuits (OCS). Generally speaking, the circuits are simpler to manage and they offer strict service guarantees at the cost of set-up latency that reduces the reactivity. On the opposite, packet switching provides fine granularity, robustness, flexibility and efficient resource utilization ; but the delay and the loss probability cannot be guaranteed. The assets of circuit and packet switching have been largely discussed in the 70's (resulting in the de-facto standard Internet

Protocol, IP). The choice of either of packet or circuit is driven by the compromise between the reliability and the reactivity of the network.

This compromise has been re-addressed in the 90's in the context of optical MANs. The efficiency of circuit switching was threatened by the variability of the traffic. Part of the community focused on improving the reactivity of OCS (see LCAS+VCAT and light-trails). An alternative approach was the migration towards optical packet switching. OPS has been widely investigated, in spite of the challenging question of storage capability in the core network, mandatory during the routing procedures.

OBS appeared in this context as an intermediate solution between packet and circuit switching. As in OPS, the payload is emitted in "small" autonomous units (burst in this context), each requesting for its necessary resources. The switches are re-configured between any two successive bursts. The main advantage of OBS over OPS is that instead of requiring buffering in core nodes – to store the payload during the routing procedures –, the payload is buffered at the edge of the network. Each burst is signaled "in-advance" so that the switching devices are configured prior to its arrival, similarly as in OCS networks.

As compared with OCS, OBS presents two fundamental differences. Firstly, as core nodes are reconfigured between two successive bursts, fast switching devices are mandatory. Such equipment are more expensive than MEMs based switches, but they are required to reduce the transport granularity. Secondly, the circuit set-up latency of OCS, which prevents fast reconfiguration, is reduced in OBS by discarding signaling acknowledgement. The counter-part is that such "one-way" signaling raises the problem of contentions and prevents transmission guarantee. Finally, the improvement of the reactivity sacrifices the transmission guarantee, and vice-versa. As a result, OBS and OCS are not evaluated according to the same metrics and it is complicated to propose a fair comparison between OBS with OCS.

Instead of focusing on performance comparison, a fairly recent tendency consists in taking advantage of OBS and OCS assets. Indeed, the antagonism of OBS and OCS assets suggests that they

are not suited for the same scenarios. This observation has motivated a number of proposals for hybrid architectures. A first set of solutions completely relies on the OBS framework and enables circuit switching from the control plane (PATON [QWL06]). This architecture is highly dynamic and flexible since the number of circuits can be adjusted on demand. In order to reduce the equipment cost, hybrid hardware has been envisioned, combining cheap and slow MEMS with fast and more expensive SOAs. The MEMS part is assigned to circuit switching and the SOA part handle OBS traffic. the shortcoming of this solution is that the OBS switching capacity is statically set by the equipment dimensioning, whereas in pure OBS architecture, the switching capacity is spread arbitrarily between circuit and bursts.

A promising alternative consists to transmit bursts over an OCS framework (light-trails). With light-trails the routing is OCS-based, whereas the access to the resources is done in OBS fashion. This transform a light-path into a directed bus. Though the rigidity of OCS persists at the routing level (optical switching device configurations), the granularity and reactivity issues are solved. This solution is less flexible than a pure OBS network, but it presents attractive advantages. Firstly, the equipment cost is reduced. Secondly, the contention problem is seriously simplified due to OCS routing restrictions. In [GC03b], experiments report that the loss probability is lower with light-trails than with OBS. Nevertheless, whatever the performances achieved with light-trails, it can be achieved with OBS, assuming it is subject to the same routing restrictions. A fair comparison should, at similar performance, discuss the flexibility issue, critical while, e.g., dealing with failures.

Switching Paradigm	Transport Granularity	Emission Latency	Equipment Bottleneck	Switch Configuration	Failure Recovery	Control Processing
OPS	Arbitrary	Low (aggregation)	Data storage in the core nodes	On-demand	Re-active	On-demand
OBS	Arbitrary	Medium (aggregation + signaling)	Fast switching devices	In-advance	Re-active	In-advance
OCS	Mandates SONET VCAT and LCAS	High (routing process + establishment)	SONET processing	Off-line	Pro-active	Off-line
Light-trails	Arbitrary	Medium (aggregation + + signaling)	SONET processing	Off-line	Pro-active	In-advance

Table 3: OCS, OBS and OPS comparison

2.5 Objectives of the Thesis

The current optical network and the traffic expected in the near future clearly suggest increasing dynamism. Various applications with various requirements will cohabit. As the recent deployment of passive optical access networks will have a very limited impact on the traffic, the heterogeneous and variable nature of the traffic emitted by the users should be preserved up to the MANs. With such traffic, the current OCS optical layer is seriously penalized by its poor reactivity. As a result, the integrity of the MANs is seriously compromised and the reactivity of the optical layer becomes an utmost concern, mandatory in order to improve the resource utilization.

The need for a reactive optical layer suggests the consideration of alternate paradigms. Currently, OBS appears as a natural evolution in order to build a reactive optical layer since it is viable with current technologies. It has several assets in terms of reactivity, but it also addresses the critical issue of contention. As for today, the architectures that preserve OBS reactivity cannot provide delivery guarantee, as opposed to OCS solutions that can guarantee the transmissions but sacrifices the reactivity.

The antagonism of OBS and OCS assets led to the proposal of hybrid architectures where the traffic benefits from either the reactivity of OBS or the reliability of OCS. This thesis aims at breaking the compromise between reliability and reactivity ; it aims at proposing a reliable and reactive solution for the optical layer. Whereas part of the community focuses on the improvement of the reactivity of loss-less OCS solutions, we propose here to start from reactive OBS solutions and endow them with mechanisms that sidestep or solve the contention issue. OBS loss-less architectures will be evaluated with regard to their reactivity, our goal being to preserve the asynchronism, the low latency and the dynamism of OBS.

Our loss-less OBS solutions will be qualitatively compared with OCS solutions in terms of reactivity, in order to evaluate their meaningfulness. A fair comparison in terms of performances mandates a wide set of experiments and is out of the scope of this thesis. However, the proposal of loss-less OBS solutions is a crucial step toward a fair comparison since it allows performance

evaluation with similar service (loss-free).

CHAPTER 3

Optical Circuit Switching (OCS)

In the previous chapter, we described the baseline of the Optical Circuit Switching. In this chapter, we recall the basics of OCS and propose an overview of the topical issues in OCS networks. OCS is the current technology used in MANs and backbone networks. It relies on the pre-configuration of the switching equipment to avoid conflicts in the data plane.

3.1 OCS Basics: Illustration in an All-Optical Network

3.1.1 Fundamentals

In OCS, processing and switch configurations are still mostly performed off-line. The switching devices are configured from the source to the destination for the entire duration of the connection. Usual equipment for optical switching is based on MEMS that are placed between wavelength demultiplexers and multiplexers to operate at the wavelength level.

Such a transport granularity is too coarse and sub-wavelength granularity is required to use the transport capacity of the optical fibers efficiently. On the example depicted on Figure 3.1, all optical provisioning requires one wavelength for each connection, though they may only request a fraction of the transport capacity. With current technology, the capacity of a single wavelength ranges between 2.5 Gbps and 10 Gbps in the MAN and between 10 Gbps and 40 Gbps in the WAN. As a connection

rarely requests beyond OC-3 (155 Mbps), assigning a single request to each wavelength entails a waste of transport capacity. Switching at sub-wavelength granularity is crucial because of the gap between the requested bandwidths and the wavelength capacity.

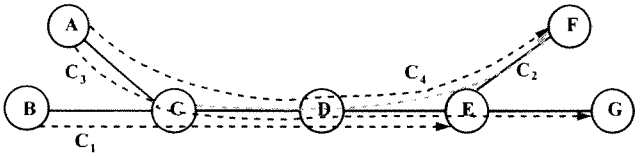


Figure 3.1: An Illustrative Scenario

3.1.2 Routing and Wavelength Assignment (RWA)

The RWA problem describes OCS in all-optical mode. Optical switches operate at the wavelength level and, as TDM (implemented with SONET/SDH, see Section 3.2) cannot be deployed in all-optical networks, no two connections can share a wavelength in a fiber. As the request granularity is usually far from the transport capacity of a wavelength, the resources in all-optical networks would not be efficient used, and such scenarios are only theoretical fields of investigation. Nevertheless, the related contributions are are worth to be mentioned because they are baselines for more elaborated models that include grooming, failure or signal impairment.

In static scenario, the traffic matrix is constant over the time and the problem can be formulated as an integer linear model ([BM00a, JMY06b, JMT06, JMT07b, JMT07c, TMP02]). The design flavor of RWA aims to minimize the infrastructure costs, i.e., the number of ports (or wavelengths) to be used to grant a given matrix of traffic (e.g., [Agg94, BH94]). The provisioning variant maximizes the grade of service (i.e., the ratio of granted connections) for a fixed number of wavelengths, see, e.g., [JMY06b, RS95].

3.2 SONET/SDH over WDM

3.2.1 Overview

Currently, multiplexing several flows within a single wavelength is performed by SONET/SDH that combines TDM with WDM. SONET/SDH is built around a signal hierarchy. Various protocols are interfaced with the STS-1 signal, that can be combined in STS-n streams. The SONET/SDH hierarchy (reported in Table 1 in Section 2.1.2) defines the granularities that can be handled by combining several STS-n streams. The resulting signal is emitted in the WDM network where optical switching at the wavelength level is achieved. The transmission remains all-optical as long as no multiplexing or demultiplexing operation is required. Those operations consist adding or dropping STS-n streams to/from the wavelength. Thus, the SONET/SDH hierarchy encapsulates traffic to facilitate the grooming procedures. However, manipulating sub-wavelength streams must be done in electrical domain. In a SONET over WDM network, a light-path is composed by a sequence of optical segments (OLSP), defined by a couple (λ, R), where λ is a wavelength and R is a set of connected links. The data are sent and travel all optically along each optical hop, with switches (typically OxC or OADM) operating at a wavelength level, the fiber level or in between (wave-band level). The node that interconnects OLSPs combines two switching fabrics. The OxC is used by wavelengths that do not require sub-wavelength manipulation, whereas those that must be manipulated at the SONET level are handled by a DxC, in electrical domain. Thus, the transmission is thus all optical along each OLSP and grooming operations are performed between each OLSP in electrical domain.

In our example (Figure 3.2), node C can multiplex all the connections and emit all the traffic on a single wavelength (assuming it offers enough transport capacity). Then, the signal is converted into the electrical domain in node E that can demultiplex the connections and re-emit each of them on the appropriate output port.

The granularity handled by the SONET/SDH hierarchy however remains too coarse for an efficient resource utilization [Mim02]. Virtual Concatenation (VCAT, [VCA02]) breaks the STS signals

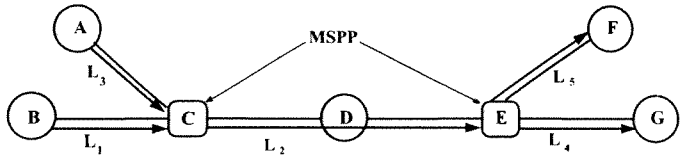


Figure 3.2: Provisioning Scenario of Figure 3.1 with SONET

into smaller individual containers grouped logically in a channel. A panel of granularities is proposed to better map the connection granularity. Members of a channel are routed independently through the network and all the intelligence regarding the virtual concatenation is located at the endpoints of the connection. Thus, VCAT can be deployed on the existing SONET/SDH infrastructure with a simple upgrade of the endpoints and the capacity assigned to each connection better maps its data-rate and the over-provisioning due to transport granularity is reduced.

3.2.2 Grooming, Routing and Wavelength Assignment (GRWA)

With grooming facilities, the design problem is much more challenging. In that case, a light-path is composed of several optical segments with grooming devices at the end-points. The grooming devices must be carefully placed in order to minimize their impact on the CAPEX. In addition, they should be economically accessed to reduce the impact of O/E/O conversions on the end-to-end delay.

The static version of the problem is adapted in WAN networks since it is assumed that the traffic does not change drastically over the time. It can be solved through integer linear programming ([ZM02b, RCR05] in mesh networks, [JJ05] in ring networks) for small networks. Approximate solutions ([VJV09, HL04, JBCB07]) and heuristics ([HJS05]) are considered for larger instances. In [HL04, JBCB07], the problem is solved in two steps. First the grooming problem is a design formulation that optimizes the virtual topology (i.e., the logical connectivity in the network). The virtual topology depends on the definition of the OLSPs, itself ruled by the placement of the SONET-ADMs. Once the grooming problem is solved, the RWA problem is solved on the virtual topology. The sequential resolution of the two problems highly improves the scalability (large instances – up

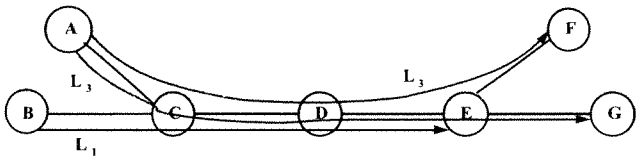


Figure 3.3: Provisioning Scenario of Figure 3.1 with Light-trails

to 30 nodes and thousands of requests – are solved in [JBCB07]) but no optimality is guaranteed.

3.3 Light-trails

A light-trail is similar to a light-path, in the sense that it also relies on the off-line establishment of a transmission channel between two nodes. However, the signal can be transmitted between any two nodes of the light-trail, transparently and without any optical switching operation. In other words, light-trails open the possibility for optical traffic grooming. In this section, we describe the fundamentals of light-trail architectures. Then, we will discuss the Medium Access Control, and finally, we will provide an overview of provisioning techniques.

3.3.1 Principle

With light-trails, switch configurations are performed off-line. From the optical switch point of view, there is no difference between light-paths and light-trails. They are set up and teared down in the same way. The difference lies in the access to the medium. Whereas SONET imposes a synchronous transmission between the two end-nodes of the OLSP, a light-trail defines an unidirectional bus that can be accessed by any node to communicate with any downstream node (see Figure 3.4). Back on the example of Figure 3.1, instead of establishing one OLSP for each source-destination pair, a single light-trail is established and provides a full logical connectivity with downstream nodes (Figure 3.3. All those connections time share the bus to improve its utilization.

3.3.2 Equipment

The optical switching of a light-trail can be supported by MEMS based devices. Those devices are configured at the light-trail establishment and this setup will persist for the whole life-time of the light-trail.

Prior to the optical switching fabric, the signal is treated by the Light-trail Access Unit (LAU) described in [VBMS05] and reported in previous chapter on Figure 2.5. The LAU is the key component to provide multi-point to multi-point connectivity within the light-trail: It transmits part of the power of the signal to the local node and can either block or transmit the remaining of the signal to the MEMS. If the shutter is closed (it blocks the signal), then the local node can use the bus and emit signal.

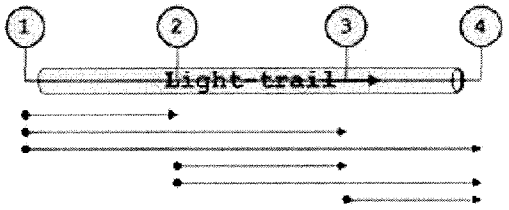


Figure 3.4: Logical Connectivity with Light-trails (taken from [FHS04])

3.3.3 Medium Access Control in Light-trails

Once the light-trail is established, its nodes can transmit to their downstream nodes with no need for optical reconfiguration. However, access to the bus must be controlled to avoid simultaneous accesses, resulting in data loss.

Light-bus Access

The first approach, proposed in [FVS04] consists in observing the incoming signal to identify free time-slots, i.e., periods during which the light-trail is free and emission is allowed. The deflected part of the signal informs the nodes on the activity of the light-trail: If a given node does not receive

any signal, it deduces that it can use it.

To let the node synchronize the emission with the free time-slot, the incoming signal is delayed by a fix amount of time in a FDL [FVS04] (see Figure 3.5). The "duration" of the FDL should at least be equal to the maximum frame size. This architecture reminds the one of OPS, where the data plane is also delayed to let the node take some decisions. However, with light-trails, no routing procedure is run and much smaller FDLs are required.

The main drawback of this architecture is that an intermediate node on the light-trail cannot block the optical signal, though it may be the final destination. Thus, any data emitted on the light-trail will be transmitted up to the destination. It results that the last links will be highly loaded with parasite signal.

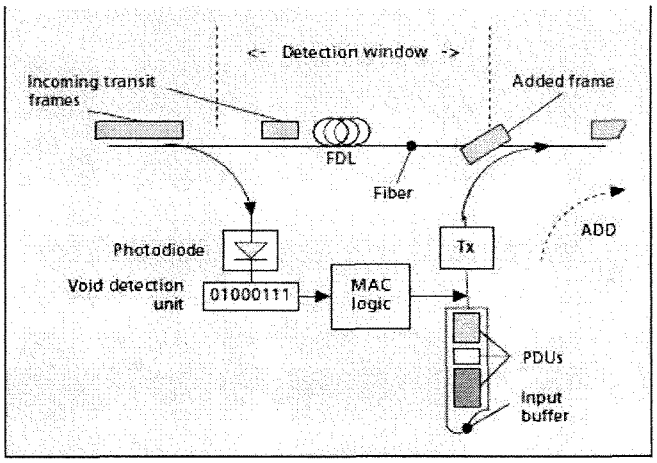


Figure 3.5: Void Detection Based MAC (taken from [Bou07])

"OBS" Light-trails

Connection signaling appears unavoidable to sidestep the shortcoming of light-bus. The approach described in [GC03a] is highly inspired by the signaling in optical burst switching. The sender emits a control packet (out-of-band) to request the bus. A node that receives this packet interrupts its emission (if it was actually emitting) and opens the shutter, unless it is the last destination, in which case it blocks the signal.

The control packet must be converted toward electrical domain to let the node read its content. As, on the opposite, the payload cut-through the switches all-optically, it is emitted some time units after the control packet, to guarantee that all nodes are configured properly at its arrival. With such an "in-advance" signaling, the contention is observed in the control plane while the ingress traffic is still in the electrical domain. As a result, it can be rescheduled on a further available time-slot.

This protocol is not contention-free but as the traffic in transit preempts the resources over the ingress traffic, any contention can be solved by rescheduling the emission of the ingress traffic. The counter-part of preemption is that some resources may be unnecessarily reserved and downstream nodes have a better access to the bus [VBMS05].

To solve those shortcomings, an alternative have been envisioned in [GC03a]. It is proposed to break the transmission of the transit traffic in the contention node and to re-emit it from there, once the bus becomes available. This way, a node can actually decide which signal should be served and it will reschedule the other one. Consequently, it improves the control on the medium access process and can adjust the fairness. The drawback that arrises when the transit signal in stored and forwarded is that it wastes some Offset Time and experiences additional delay.

3.3.4 Provisioning with Light-trails

The light-trails opens the possibility of multiplexing several connections in time domain. The multiplexing is ruled by the light-trail configuration. The light-trail configuration is a crucial concern: As the set-up of the carrier imposes a large delay, the configuration must be carefully elaborated in order to minimize the reconfiguration. It is also important to avoid bottlenecks in the network, because highly loaded links could severely increase the delay.

The objective of the ILP model proposed in [FHS04] is to minimize the total number of required light-trails to serve a given matrix of traffic. This objective tends to prioritize long light-trails. It has been previously considered in [GKC03] (in the case of clustered light-trails – CLT) as a way to reduce the number of required wavelength (and consequently to reduce the network costs), but also

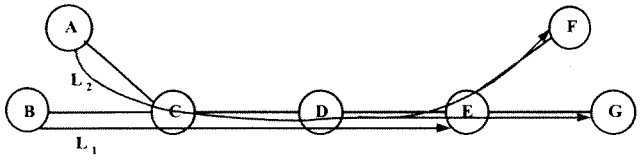


Figure 3.6: Provisioning Scenario of Figure 3.1 with Clustered Light-trails

because long light-trails can support simultaneous transmissions of non-overlapping connections.

In [ZXTT05], the routing problem is considered in a dynamic context, with and without protection. Comparison with light-paths shows, in both cases, a better grade of service with a lower wavelength requirement when light-trails are used.

3.3.5 Clustered Light-trails

A light-trail traditionally describes a path between two nodes. In that case, an input port is connected to one output port and the signal either goes to the output port or is dropped in the local node, or both. The "drop and continue" capability enables multi-casting among a set of downstream nodes.

An extension have been envisioned to increase the multiplexing and the multi-cast potential. A clustered light-trail defines a tree: At each node, only one optical input port is involved, but the signal can be split and transmitted among several output ports. Thus a clustered light-trail connects a larger number of nodes and can reduce the overall number of light-trails to be deployed. Figure 3.6 illustrates the enhancement: Although the destination of C_3 and C_4 are not on the same path, they can be carried by the same light-trail and node E will switch the signal either toward node F or node G, according to the information previously supplied by the control plane.

This architecture requires a much complex switch architecture that cannot rely to classical OCS equipment (MEMS). Figure 3.7 represents a node of connectivity 4. Each input port is connected to each output port and, at a control packet arrival, the corresponding input port is connected to a set of output ports. With this architecture, the number of required blockers required is highly increased (N^2, but it can be reduced to $N\log_2(N)$ with a tree arrangement). The splitters and the blockers

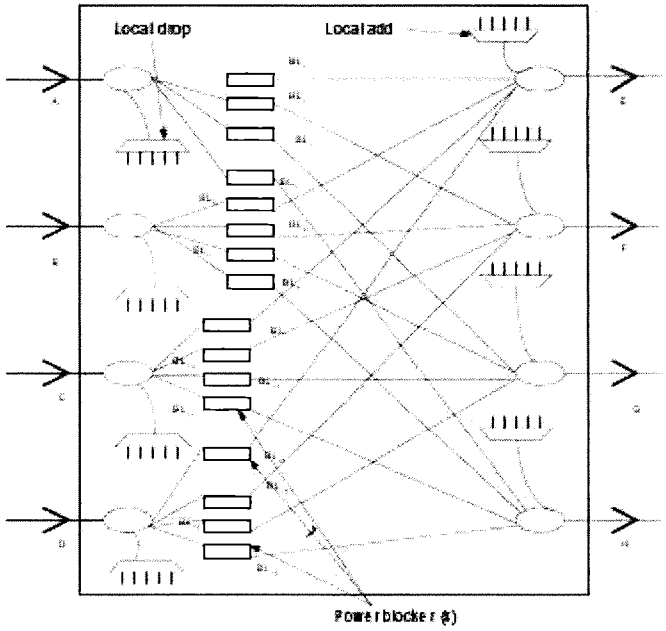


Figure 3.7: A Light-trail Node Architecture for Mesh Networks (taken from [GC03b])

completely define the internal connectivity of the switch, thus replacing the MEMS. Therefore, the switch configuration time is reduced from few milliseconds to few micro-seconds, or even nanoseconds with Semi Optical Amplifiers (SOAs), and off-line configuration is no more required. In the next chapter, we present the Optical Burst Switching, that is based on similar equipment and signaling protocol. Indeed, the architecture presented here should definitely be labeled as an OBS solution. This will be further discussed in the next chapter, after the description of the OBS architecture.

3.4 Fault Management

Considering the huge data rate offered by optical fibers, and taking into account the failure cut probability and the average time to repair [ZM04a], a failure leads to the loss of a large volume of data [SSS01] and is a critical issue in optical networks. In case of failure, several connections must be re-routed and the objective is to minimize the latency before the service is restored. In OCS networks, restoration mechanisms sequentially notify the failure, look for available resources

and establish a new channel. Those reactive mechanisms cannot guaranty the availability of backup resources and impose a non-negligible recovery time.

Protection mechanisms can ensure survivability and reduce the recovery time by reserving backup resources at the connection establishment. If the backup path is dedicated to a single working path (1+1 protection), the signal can be sent simultaneously on both paths to annihilate the recovery latency. However, such solutions allocate a large capacity to protection and consequently reduces the provisioning potential of the network. The resource redundancy can be reduced with shared protection: several connections share some backup capacity. The trade off is that the recovery time increases due to the notification delay and the possible configuration of some switches along the backup path.

To balance the recovery time and the resource redundancy, the protection can be designed at different levels: Link, path or segment ([WSB02]). With link protection, each link is assigned a backup path. The notification is instantaneous, but the redundancy is high, though it can be reduced by resource sharing [SR04].
Path-oriented schemes reserve a complete alternative path from the source to the destination. The notification delay increases, but the redundancy is usually reduced as compared with link protection. The segment-oriented schemes are in between path and lik protection and balance their advantages and shortcomings. The path is divided into a number of segments to be protected independently. Thus the restoration time is smaller than with the path protection and redundancy is reduced as compared with link protection [PHNL06, HHK06].

Nowadays, the p-cycles appears as the most promising solution [GS98]. It transposes classical Bi-directional Line Switched Ring protection BLSR to mesh networks. The backup capacity is deployed on cycles. The switches can be pre-configured and the cycle protects any segment between any two of its nodes. In case of failure, the two end-nodes just have to deflect their traffic on the cycle. The high sharing degree of the backup capacity leads to a significant reduction of the redundancy. The problem has first been approached with heuristics (e.g., [DHGY03]), but recently, Sebbah *et*

al. proposed a column generation formulation that provides near optimal solutions as fast as in few seconds [SJ09]. Then, they extended the p-cycles to so-called p-structures, where protection is not restricted to cycles. This is currently the most efficient mechanism, with a redundancy ranging around 40%.

Although the major part of the literature considers single link failure, an emerging concern is the recovery in more tragic scenario such as natural disasters or bombing, which may affect a large area. The most important concern in that case is the survivability of the network, before considering the service perturbation. The protection/restoration is designed Shared Risk Link Groups (SRLGs) composed of links likely to be affected simultaneously in the considered scenario.

3.5 OCS for a Dynamic Optical Layer

The lack of reactivity of OCS has been addressed in the 90's, when optical MANs have been deployed. As discussed in Section 2.2, this issue becomes critical since the deployment of FTTx. How to handle the traffic variations depends on the amplitude and the persistence of the variations. In some cases, the variation can be handled by adjusting the transport capacity of existing light-paths. In case of large and persistent variations, it may be preferable (or mandatory) to establish new ligth-paths or/and to re-route existing ones. On the opposite, the negotiation may be inefficient to deal with burstiness (i.e., fast traffic variations).

In this section, we will discuss the potential of OCS to deal with those situations.

3.5.1 Light-trails versus SONET/SDH

With VCAT, each flow is allocated more or less the appropriate data-rate to improve the multiplexing and the resource utilization. The allocation consists in the assignment of a set of time slot that can be used exclusively by the considered flow. Restriction to a single flow allows the destination node to easily discriminate between the streams.

The problem appears in case of traffic variation. If the data-rate of a stream decreases, then

the time-slots not filled by the stream are wasted, even if another stream experiences the opposite variation: a stream whose data-rate exceeds the transport capacity will see its data buffered so as to conform to the data-rate negotiated at the channel establishment.

With light-trails, TDM is exploited asynchronously: each flow uses the wavelength on demand, as soon as it is available. Consequently, if the load of a flow decreases, the resource availability instantaneously increases for the other flows, whereas with SONET, the idle time-slot cannot be used by other flows. As a result, traffic variations can be efficiently and transparently handled by light-trails [GKC03].

However, if the variation is persistent, the transport capacity of the channel can be re-negotiated with the Link Capacity Adjustment Scheme (LCAS). LCAS has been designed to re-negotiate the capacity of a channel "on the fly". The source sends a message embedded in the SONET/SDH overhead that a virtual tributary is about to be added or remove. Then, a synchronization message is transmitted to give notice of the date the change will occur. As only end-systems manage the multiplexing, the intermediate nodes are not concerned by this modification.

With larger variations, the requested capacity may not be granted, in which case, re-routing or ligth-path establishment may be required. In the next section, we discuss the dynamic provisioning.

3.5.2 Dynamic Provisioning

It could happen that large traffic demands cannot be served by existing light-trails or OLSPs, in which case, new carrier establishment or existing carrier re-routing are required. As we already discussed, re-routing entails significant perturbation of the service (and involves a large number of connections). This is a challenging issue since some applications do not tolerate service perturbation. Then, though other applications may be fairly tolerant to short term interruptions, the perturbations should be avoided as much as possible.

Thus, a dynamic provisioning process must quickly propose a set of actions that allows the service of the new traffic with a minimum number of re-configurations. The configuration should also

anticipate future requests. This problem has been intensively treated by the community (as reflected by the survey proposed in [HD07]). To speed up the computation time, dynamic provisioning is mostly solved by heuristics. Few ILP formulations have been proposed (e.g., in [XWCL05b]) and used on small topologies (5 nodes). The poor scalability of exact resolution scheme restrict their use to heuristic evaluation and configuration (e.g., in [XWCL05a]). The heuristics can then be used for larger networks. The objective in dynamic scenarios is to grant the largest number of connections by re-routing the minimum number of connections, without any assumption on next connection arrival.

Dynamic provisioning with grooming devices mainly consists in choosing between the use of existing light-paths and the creation of new light-paths, see, e.g., [SSS06, CZB04, YR04b, WHLW03]. The provisioning is computed on a virtual topology that describes the logical connectivity. To reduce the size of the virtual topology, an incremental approach is proposed in [HL05]: Only nodes on shortest paths are considered and their neighbors are added sequentially until a solution is found or all nodes have been added. In [TOZ+05, TBZ+07], the resource utilization is improved by taking into account the remaining holding time of the connections in place.

Note that efficient solutions of the static problem (e.g., [JBCB07]) can be be helpful for periodical re-configurations as in [GM03], where re-routing is performed periodically to achieve load balancing. Thus, the computation time is not critical because it is processed off-line. The objective is to minimize the utilization of the most congested link. The idea is to tear down under utilized light-paths by re-routing their connections through existing OLSPs. Then, a new light-path is established to reroute the largest connection of the most congested light-paths. This strategy reduces the provisioning delay of the future requests.

In [YR04c], the authors propose to re-route connections instead of light-paths. At the connection level, the computation is more complex, but the operator has a finer control. Especially, a re-routed connection may use an existing light-path, with the objective to balance the load on the virtual topology. In that case, the service perturbation is reduced. On the other side, if a connection is rerouted on a new lightpath, then the connectivity of the virtual topology is increased, and this may

help to serve future connections.

The capability to serve a new request without carrier set-up is highly related to the transport capacity and the connectivity of the virtual topology. The limitation of SONET systems is that the virtual topology is less flexible than the one involving light-trails. The first reason is that light-trails improve the connectivity of the virtual topology (OLSP only connects the end-nodes whereas a light-trail connects each node with all the downstream ones. The second reason is that the light-trail establishment is not subject to any design issue, whereas the definition of the OLSPs is ruled by the location of the electrical modules. The flexibility of the grooming operations leads to a better provisioning with light-trails in dynamic scenario ([SCD+07, Bou07]).

3.5.3 Dynamic Fault Management

Provisioning becomes more complex in a dynamic scenario since request arrivals are not known in advance. A promising approach is the deployment of a protection envelop. Instead of defining backup path jointly with the working path, a backup network is set-up off-line, and provisioning is done accordingly. In other words, the backup network defines a protected network on which the provisioning can be done on demand.

In [LNJ+05], the ant algorithm is modified to look for a population of cycles instead of paths. A cycle includes both primary and backup paths and can be used by several requests. The population of cycles so generated is then improved by a genetic algorithm In [RZG06], dynamic provisioning with protection is addressed with regard to the signaling overhead and the blocking probability.

3.6 Conclusion

The rigidity of the current optical layer is largely attributable to the slow configuration of the switches. It imposes off-line and long-term resource reservation, which complicates sub-wavelength multiplexing and disables fast reconfiguration to react to unforeseen events.

Nowadays, the shortcomings of the circuit granularity are sidestepped with SONET/SDH, that

operates multiplexing/demultiplexing in time domain. Flows are manipulated in electrical domain so as to be emitted on a pre-established set of time slots. The source and the destination must however exchange synchronization information so that the source can transmit "packets" over the light-path and the destination can discriminate each flows without any signaling or overhead, but only based on their location in the stream (i.e., time slot). Thus, SONET/SDH systems still relies on pre-planned transfers, and the resource allocation remains static. However, the resource allocation within the SONET stream only concerns the end node and it is managed in electrical domain. Thus, it can be modified without involving the optical layer (LCAS).

To avoids pre-configuration and synchronism, the source node must provide an alternate way to discriminate the flows. With light-trails, the payload is signaled in advance by an out-of-band header. This way, every node can be set to forward the data, except the destination node that retrieves the signal. Thanks to the in-advance signaling, any burst along the light-trail can emit data, provided the light-trail is free. Thus, light-trail opens a large multiplexing potential and can efficiently deal with traffic variations.

Nevertheless, both light-trails and SONET/SDH intend to improve the reactivity over an established light-path, but those mechanisms are still subject to the service perturbation is case of optical switch reconfiguration. Re-routing is typically initiated to recover from a failure or for provisioning purpose.

The failure recovery received lot of attention, resulting is quite competitive mechanisms that enable fast recovery with a reduced redundancy (p-structures, see [SJ09]). The redundancy might prevent from online circuit establishment, but there is no way to sidestep online establishment requested by an unexpected connection.

Enabling fast and dynamic re-configuration of the optical layer imposes to reduce the switch configuration time. In Section 3.3.5, fast switches are considered to extend light-trails from linear structures to trees and enhance the multiplexing and the resource utilization. Nevertheless, in order to achieve loss-less transmission, the routing is ruled by several restrictions that impose off-line

configuration computation. Though clustered light-trails are derived from an OCS architecture, it is based on similar equipment and protocols as Optical Burst Switched networks (OBS), previously proposed in [YQ99] as an alternative to OCS, and presented in the next chapter.

CHAPTER 4

Optical Burst Switching (OBS)

The lack of reactivity of OCS is due to the slow configuration of the switching equipment. The unfortunate consequence is that on-line resource allocation entails latency and service perturbation. Online allocation can be prevented at the expense of extra capacity allocation, either to deal with failure or to handle bursty traffic. Burstiness can also be handled quite naturally by light-trails. In that context, redundancy can be reduced. The provisioning redundancy reduces the resource utilization and cannot completely avoid online re-configuration in the case of dynamic traffic.

The need for optical reactivity has been seriously considered at the time the MANs have been enlightened. Optical Packet Switching attracted a deep attention because it removes the transport granularity and reduces the emission latency. The major limitation to OPS deployment is the lack of a reliable solution for the data plane buffering. A storage solution is mandatory at each node and for each packet during the setup of the switch. The Optical Burst Switching (OBS) proposes a solution to sidestep data-plane buffering in core networks, based on the appropriate setup of the switching equipment for each burst, prior to its arrival. OBS preserves the packet-oriented nature of OPS (and consequently, discards the transport granularity) but the pre-configuration of the switching equipment is closer to OCS and imposes emission latency, that compromises the reactivity. The emission latency is reduced by neglecting resource reservation acknowledgment. Such a "one-way" signaling protocol improves the reactivity of the network, but also raises the problem of contention:

The data are sent without guarantee that resources are available up to the destination and data losses can occur.

In this chapter, we provide an overview of OBS fundamentals (equipment, the aggregation process, the signaling protocols) in Section 4.1. Then, the contention problem is presented and loss approximation models are discussed in Section 4.2. Section 4.3 presents the major contributions related to the contention problem. The contributions that achieve loss-less transfers are presented in Section 4.4, where their impact on the reactivity of OBS is discussed. Finally, Section 4.6 concludes this chapter.

4.1 Fundamentals of OBS

The Optical Burst Switching (OBS) has first been proposed in [YQ99] as a solution for all-optical, packet-oriented communication. This section overviews the fundamentals of OBS. For a more detailed description of OBS basis, the reader is referred to [JV05].

4.1.1 OBS Signaling

The signaling in OBS accommodates the lack of optical memory. Buffering in the core network is the major obstacle to packet-oriented transmission since it is required to delay the data plane during control plane operations. In OBS networks, the overall duration of control plane operation – i.e., the overall delay to be imposed to the data plane for routing purposes – is spent in the edge node and switches are configured appropriately for each burst, prior its arrival.

In the original proposal, OBS was signaled with the Just Enough Time (JET, [QY99]) protocol illustrated in Figure 4.1. A control header is sent ahead, on a wavelength dedicated to the control plane, followed by the burst. The gap between the header and the burst is called Offset Time (OT).

At each node, the control header is converted toward electrical domain and supply the information on the burst. The information includes the arrival date of the burst, its duration and its destination. The router then determines the appropriate switch configuration to handle the burst, updates the header and forwards it to the next node. Thus, at each node, the header is delayed (the duration of

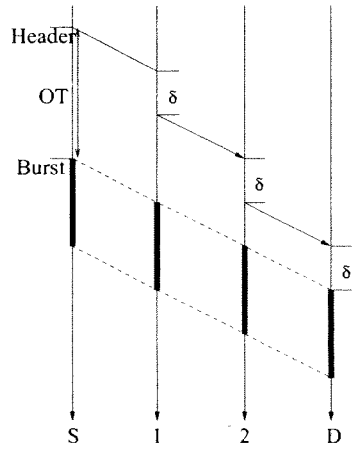


Figure 4.1: OBS-JET Signaling

the header processing is denoted by δ), whereas the burst is not.

As a result, the gap between the header and the burst is reduced by δ at each node. To ensure that each switch is configured appropriately at the burst arrival, the bust must always be preceded by the header, which boils down to keep a non-negative OT. Thus, the OT must be larger or equal to the overall processing time from the source to the destination. Formally, at each node v, the OT must satisfy:

$$OT > \ell_v \times \delta \tag{1}$$

where ℓ_v is the number of nodes between v and the destination of the burst and δ is the processing time of the header at each switch (assuming it is the same processing time at each node).

The JET protocol proposes to set the switch just before the burst arrival. In that situation, the resources are reserved for the exact duration of the burst. Several variations of JET have been proposed. For instance, with Just In Time (JIT, [WM00]) the switch is configured as soon as the header is processed. This alternative simplifies the processing but mitigates the resource utilization since the resources cannot be used between the header and the burst. More sophisticated protocols have then been designed for particular purposes. Those of interest for our project will be presented later in this chapter.

4.1.2 Equipment

Figure 4.2 depicts the architecture of an OBS core router. Over the W wavelengths multiplexed in a fiber, N are dedicated to the control plane. The N control channels are directed to the control packet processor and the $(W - N)$ data channels are connected to the optical switch fabric. On the control plane, the header is converted toward electrical domain and its information on the associated burst are used for routing procedure. The data channel scheduler keeps the resulting configuration in memory and the header is updated, then converted back to the optical domain and forwarded toward the next node. Just before the burst arrival, the optical switch scheduler sets the appropriate configuration so that the burst is switched properly in the optical domain.

With this architecture, a wavelength can transport several bursts from different connections, i.e., to be switched differently. The resulting statistical multiplexing is expected to improve the resource utilization. Nevertheless, the switching fabric must be reconfigured between two successive bursts, thus, they must be separated by a guard time, equal to the reconfiguration time of the switch (δ^s). This last constraint somehow drives the design of the switches and the length of the bursts: The switching technology must be fast and the bursts must be orders of magnitude longer than δ^s to absorb the guard time and improve the resource utilization.

Thus, MEMS is not suited because of their slow switching time (around 50 ms according to [JV05]). The best potential architecture is based on Semi-conductor Optical Amplifiers (SOAs). The signal is broadcast to multiple SOAs by an optical coupler. At each SOA, the signal either transits toward the output port, or is blocked. The SOA can be switched from a state to the other as fast as in nanoseconds ([JV05]). According to [OMPR03], a realistic switching time is upper bounded by 100 ns. Note that this architecture offers a natural multiplexing capability: In the extreme case, the signal reaches all output ports if all SOAs are open. Note that splitting the signal results in a reduction of the signal power. This aspect can limit the distance traveled by the signal. On the architecture presented on Figure 4.2, the signal is boosted after being switched.

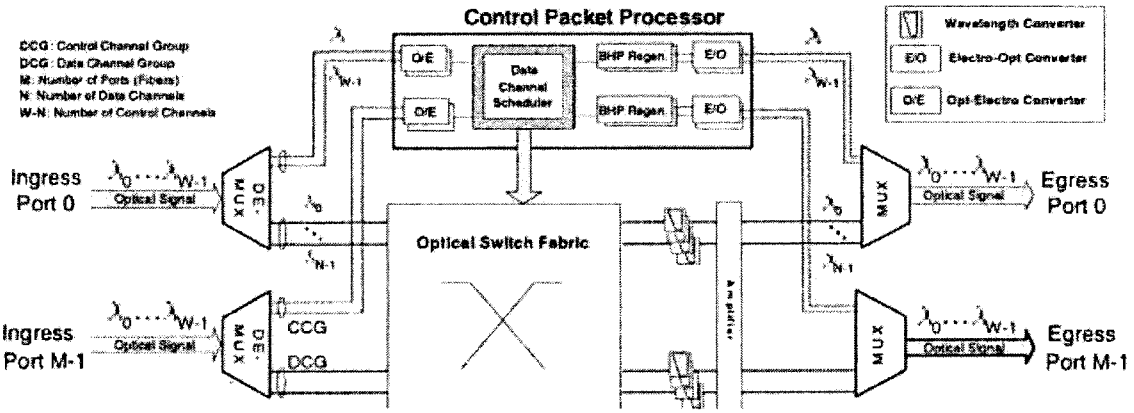


Figure 4.2: Architecture of OBS Core Router

4.1.3 Burst Aggregation

To reduce the fragmentation of the resource and the signaling overhead, several packets are aggregated into a burst, signaled by a single header. Thus, from the core network point of view, the packets of a given burst are identical and will be handled the same way. Therefore, in the edge nodes, ingress packets are assigned a label, defined by the operator according to, e.g., its destination or/and its class of services. Packets with the same label are stored in a common aggregation queue (AQ) and sent together as a burst at the occurrence of a triggering criterion. The triggering criterion is usually either *time-driven* (the burst is triggered after a given aggregation duration) or *size-driven* (the burst is triggered once it reaches a given size). The shortcoming of both criteria is reflected with low loaded AQs: The time criterion triggers small bursts, whereas the data will spend more time in the AQ before the size criterion is met. The former mitigates the benefits of aggregation, whereas the latter reduces the reactivity.

A hybrid solution is presented in [CLCQ02], where both criteria are considered: The burst is triggered once the timer expired or the burst reached the required size. This way, the assembly process time is bounded by the time criterion and reduced in the case of highly loaded AQ by the size criterion.

The burst grooming is an alternative to avoid small bursts. In [FZJ05], the burst triggered by a timer-criterion is complemented with the payload of other AQs. The resulting burst contains

several sub-bursts, separated somewhere along the path to the destination. Demultiplexing the sub-bursts involves electrical processing and thus mandates signal conversion. In [LQ04], a burst can be triggered in order to follow another burst to be handled the same way in a number of nodes. This can be viewed as traffic grooming from the data plane perspective. In that case, demultiplexing can be achoieved in optical domain. The traffic grooming offers additional benefits and will be deeply described in Chapter 7.

4.2 Contentions in OBS Networks

The reactivity of OBS relies on the "one-way" signaling. The expense of this reactivity is that the burst is sent with no explicit guarantee to reach the destination. In this section, we describe the contention and present an overview of pro-active and reactive mechanisms that have been proposed to face this issue.

4.2.1 Contention Definition

In this section, a flow denotes the traffic between an input port and an output port of a core node. It is composed of bursts from various connections, potentially with different sources and destinations. For instance, on the scenario depicted in Figure 4.3, assuming each link has only one wavelength, in node C, the flow from A to D is composed of burst of connection C_3 and connection C_4, though they have different destinations.

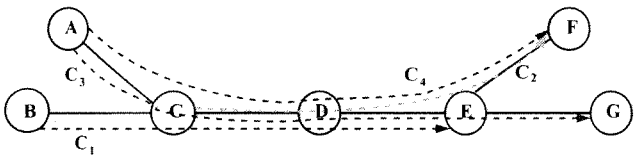


Figure 4.3: Contention Illustration

A contention is observed when two headers request for the same output port (i.e., in WDM networks, the same wavelength on a given link), for overlapping periods of time. Back on the

example of Figure 4.3, Figure 4.4 represents the traffic transit through node C and node E. In node C, three flows compete for the same output port and contention occurs if two or more bursts overlap. In that case, only one burst is served and the others are lost.

The resulting traffic on link $C \rightarrow D$ is an incoming flow of node D. The bursts of this flow did not overlap while exiting node C and, as they travel at the same speed, they do not overlap while reaching node D. As no other flow request for the output port, the traversal of node D is contention-free. In node E, the incoming flow is divided into three flows, each forwarded to a different output port. As each output port is requested by only one flow node E is contention free as well.

Finally, a contention only occurs between bursts arriving from different input ports and sent toward the same output port (node C in our exemple).

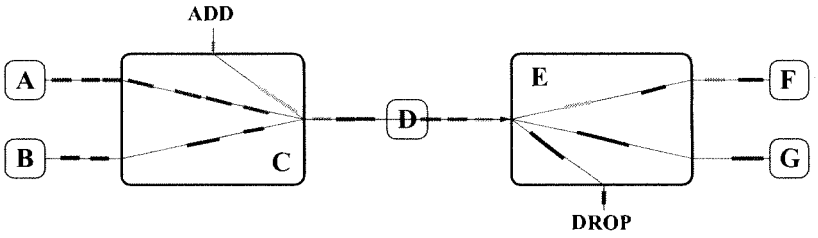


Figure 4.4: Flow Interaction

4.2.2 Loss Approximation

The loss probability is among the most important performance metric of a network. The loss probability highly impact the revenue of the operators and it is a key aspect in network planing and dimensioning. In that case, it is important to predict the loss probability obtained with a given configuration in order to setup an efficient network (network design, load balancing). Nevertheless, the network are too complex and involve too many protocols and parameters to characterize exactly. To get a estimation of the loss probability with reasonable computational effort, the loss process is described in the simplified model of the network. The quality of the approximation depends on the accuracy of the model. Sophisticated models result in a better description of the system, but strongly complicate the analysis and lead to computation-greedy mechanisms. The models are

usually limited to the description of the traffic arrival and of the server behavior.

Poisson Systems

The Poisson assumption has been widely used in many situations. Under the Poisson assumption, a source of load A is described as an infinity of sources of infinitesimal load, summing to A. With Poisson distributed arrivals, the loss probability can be approximated with the Erlang-B formula (2).

$$Erl_B = \frac{\lambda^W / W!}{\sum_{i=0}^{i=W} \lambda^i / i!} \tag{2}$$

The Poisson distribution is very popular, and, though it is relevant in some situations, it is often used as the default distribution because it simplifies the analysis, and because the Erlang-B formula can be easily integrated in online heuristics (e.g., for routing computation, as in [YR06]).

Unfortunately, the Poisson assumption is irrelevant in an OBS core network. In Section 4.2.1, we illustrated the fact that bursts of the same flow cannot contend. A Poisson traffic does not conform to this property. Indeed, since an OBS core node multiplexes a finite (and usually small) number of flow the Poisson assumptions must be definitely banned. It must be replaced by models that take advantage of the finite number of source such as the Engset scheme [FH04, Sys86].

Engset Systems

The Engset distribution is derived from the binomial distribution. It describes a system with a limited number of sources that switch between the state "idle" and the state "busy". Only the sources in state "idle" can submit a client. If the client is served, the source switchs to state "busy" during the service and then switches back to the state "idle". Thus, a source can submit only one client at a time. The client is rejected if its arrival occurs while all servers are busy. In that case, the corresponding source immediately switches to the "idle" state. The dropping probability can be expressed with (3).

$$Eng(A,S,W) = \frac{C(S,W).M^W}{\sum_{X=1}^{W} (C(S,X).M^X)} \quad (3)$$

$C(S,W)$ is the number of possible assignments of S bursts to W wavelength.

A = total offered intensity in Erlang,

S = number of incoming streams,

W = number of wavelength,

P_b = blocking probability

$M = \frac{A}{S-A.(1-P_b)}$ represents the average load of a source that are free to ask for a server.

In OBS networks, the Engset model may be accurate in the edge nodes since a rejected burst is immediately rescheduled. Nevertheless, it is not the case in the core network because a dropped burst cannot be resubmitted locally, and the next burst, as it travels on the same wavelength, cannot overlap. In other words, the behavior of an input port in a core node is independent of the performances of the node (in [CHJ09b], we called this behavior the Loss Independent Arrival – LIA). It results that the input port should remain in state "busy" for the whole duration of the burst, regardless if it is served or dropped.

The Streamline Effect

The streamline effect has been disclosed in [PCG+05]. It reflects the LIA and the fact that bursts of the same flow cannot contend. In [PSCG06], a loss approximation is proposed. The idea is to consider incoming flows as Poisson on one side (Figure 4.5-A), and the super-imposition of all flows on the other side (Figure 4.5-B). The approximation – expressed by (4) – is obtained by removing the loss probability of each flow is system A from the overall loss probability of system B (each time, the loss probability is estimated with the Erlang-B formula).

$$P_A = \frac{\sum_{i=1}^{N} \lambda_i (1 - P_i) P_i^A}{\sum_{i=1}^{N} \lambda_i (1 - P_i)} = \frac{\lambda P_B - \sum_{i=1}^{N} \lambda_i P_i}{1 - \sum_{i=1}^{N} \lambda_i P_i} \tag{4}$$

Figure 4.5: Two Models of Input Ports in an OBS Core Node

The formula however assumes Poisson arrival and quickly converges toward the Erlang-B formula when the number flow increases. Though this behavior is relevant, the convergence is too fast and the accuracy of the formula is reduced for a moderated number of flows (see Section 6.1.4.

4.2.3 The Offset Time Priority

All the models presented in this section describe the behavior of the data plane. They provide the contention probability of a system, which can be directly transposed to the loss probability. Nevertheless, in OBS, the contention is noted in the control plane, at the arrival of the header of the contending burst. In case of contention, the burst that survives the contention is the first signaled one, regardless of the actual arrival date of the burst. It results that the OT has a strong influence on the dropping process, as disclosed in [YJQ98].

Illustration of the Phenomenon

A contention reveals a conflict between two bursts in the data plane. Nevertheless, it is detected on the control plane. More precisely, a contention is detected when a header requests for resources that have already been reserved by a previous header. Hence, the burst associated with the latest header will be dropped and the first signaled burst will survive. Thus, the offset time (OT) plays a crucial role in the dropping process as illustrated in Figure 4.6. Two bursts B_1 and B_2 (of respective

duration b_1 and b_2, signaled at t_1 and t_2 by headers H_1 and H_2) compete for the same output port in node v. OT_i denotes the OT assigned to burst B_i.

We denote by OT_i^v the remaining OT of burst B_i at node v. For given bursts B_i and B_j, the difference of their OT is defined by $\Delta_{i,j}^v = OT_i^v - OT_j^v$. We define t_2^E (resp. t_2^L) as the earliest (resp. latest) arrival date of H_2 at node v, so that B_2 contends with B_1. We have:

$$t_2^E = t_1 + OT_1^v - b_2 - OT_2^v \tag{5}$$

$$t_2^L = t_1 + OT_1^v + b_1 - OT_2^v \tag{6}$$

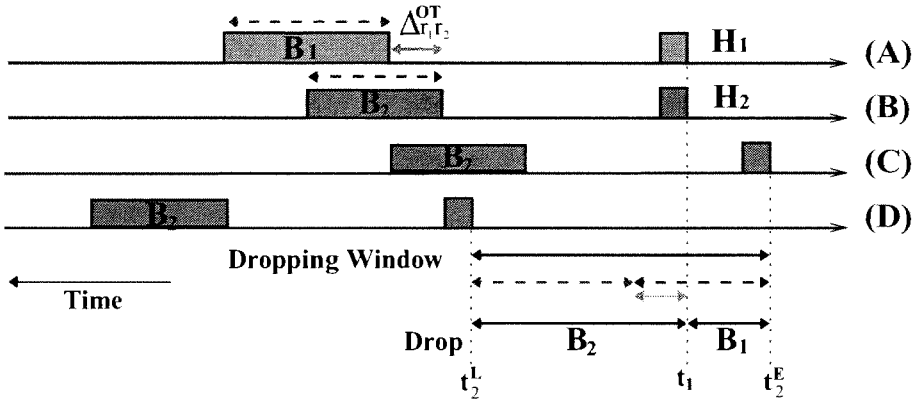


Figure 4.6: OT priority

A contention occurs if $t_E \leq t_2 \leq t_L$, so the length of the contention window is $CW = t_L - t_E = b_1 + b_2$. B_1 is dropped if H_2 arrives before H_1 within the contention window, i.e., $t_E \leq t_2 \leq t_1$ and the length of the "B_1 dropping window" is $CW_1 = b_2 - \Delta_{1,2}^v$, which decreases when $\Delta_{1,2}^v$ increases. Conversely, B_2 is dropped if $t_1 \leq t_2 \leq t_L$ and the length of the "B_2 dropping window" is $CW_2 = b_1 + \Delta_{1,2}^v$.

In case of contention, one of the bursts is dropped. The ratio CW_i/CW is the probability that burst B_i is the latest signaled, while a contention occurs. The expression of the contention windows reflects that the burst with the largest OT has more chance to survive, and this probability increases if the difference of the OT increases. Based on those observations, we derived the Offset Time Aware

(OTA) loss repartition adjustment $\mathrm{OTA}^{b}_{r_1,r_2}$ between r_1 and r_2 considering bursts of size b:

$$\mathrm{OTA}^{b}_{r_1,r_2} = \begin{cases} 1 - \dfrac{\Delta^{OT}_{r_1,r_2}}{b} & \text{if } \Delta^{OT}_{r_1,r_2} < b \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

We denote by p_{C_i,C_j} the contention probability of a burst of connection C_i with a burst of connection C_j. This probability is given by any appropriate model taking account of the specific behavior of OBS traffic (e.g., (3) or (4)). Those formulas reflect the behavior of the data plane. Thus, p_{C_i,C_j} is the probability that a burst B_i (of connection C_i) contends with an earlier burst B_j (of connection C_j). From the contention probability on the data plane, the OTA loss adjustment (7), derived from the relative size of the dropping windows, can be used the evaluate the probability $\mathrm{loss}_{r_1 r_2}$ that a burst on r_1 is dropped due to a burst on r_2, taking into account the impact of the offset time:

$$\mathrm{loss}_{r_i,r_j} = \begin{cases} p_{C_i,C_j} \times \mathrm{OTA}^{b_j}_{r_i,r_j} & \text{if } \Delta^{OT}_{r_i,r_j} > 0 \\ p_{C_i,C_j} + p_{C_j,C_i} \times \mathrm{OTA}^{b_i}_{r_i,r_j} & \text{if } \Delta^{OT}_{r_i,r_j} < 0. \end{cases} \tag{8}$$

Let us evaluate the OTA approximation on the simple example. In a given node v, two connections C_1 and C_2, having the same destination, compete for a link ℓ. Node v is the merging node of routes r_1 and r_2. In this illustration, we choose, arbitrarily $b_1 = b_2 = 5 \times \delta$. All bursts have the same basic OT, but the bursts of C_2 are assigned an extra offset time α_2. Figure 4.7 presents the loss probability for each connection using: (*i*) simulation, (*ii*) the streamline formula of [PCG+05] (see Section 4.2.2) and (*iii*) the streamline formula with the OTA correction. The x-axis represents α_2/δ.

The streamline effect formula accurately approximates the global loss rate of the simulated system, but it does not reflect the impact of the OT. According to the simulation results, when the

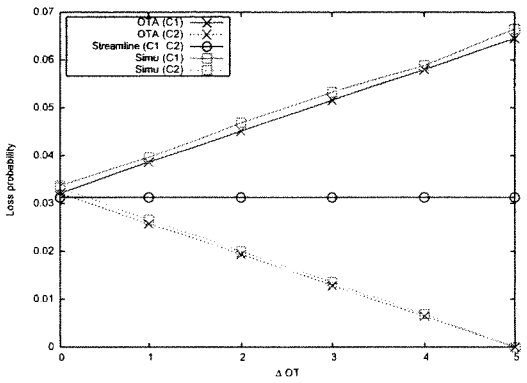


Figure 4.7: Impact of the OT on the Loss Repartition

OT of C_2 increases, the global loss rate remains constant, but the loss probability of C_2 reduces. The OTA, applied to the streamline formula, reflects this behavior: It preserves the same overall loss rate, but spreads it among the two connections, leading to the same loss repartition as measured by simulation. Note that with $\alpha_2 \leq 5\delta$, i.e., $\Delta^{OT} \geq b_1$, each contention entails the loss of B_1, and consequently, no burst of connection C_2 is dropped. This configuration is described in the next section.

Exploiting of the OT Priority

The OT priority [YQ99] can be exploited to protect a connection from the burst of another connection. If $\Delta^v_{1,2} > b_2$, then in case of contention occurs, B_1 is necessarily signaled first and cannot be dropped because of B_2. Inversely, if H_1 arrives after H_2, then the two bursts cannot contend (see Figure 4.8). In [YQ99], this configuration is exploited to achieve classes of service differentiation: The bursts with high priority are assigned an extra OT to get priority on resource reservation.

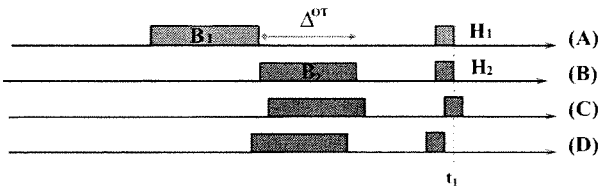


Figure 4.8: OT Isolation

Disabling the OT priority

In [LQXX04], Li *et al.* investigated the impact of the OT and the size of the bursts on the contention probability. The study disclosed the negative impact of OT priority on the worst case performances of on-line scheduling policies. The authors recommend to reduce the variance of the size and OT of the bursts. This is somehow intuitive since the OT decreases along its journey and the bursts that are close to their destination are more likely dropped.

The Virtual Fixed Offset Time (VFO, [LQXX04]) protocol is proposed to remove the effect of offset time. In classical OBS-JET, the resources are reserved as soon as it supplies the burst information. With VFO, the routing is processed and the header is forwarded to the next node, but the resources are reserved F time units before the burst arrival. This way, the resource reservation is always done F time units before the burst arrival, regardless of the OT of the burst. In [BS05], the OT priority is disabled in a similar way, but the implementation relies on a dual header: The first header initiates the routing procedures whereas the second one actually requests for the resources.

4.3 Facing Contention

Lot of efforts have been deployed in the last decade to deal with burst contentions. In this section, we propose an overview of the resulting proposals. Facing contention has been approached from two points of views: Proactive mechanisms have been proposed to avoid the contentions, whereas reactive mechanisms intend to solve the contentions.

4.3.1 Reactive Mechanisms

Reactive mechanisms tend to solve the contention by unbinding the bursts either in space, spectral or time domain.

Deflection Routing

Deflection routing solves the contention in the space domain. In case of contention, the requested output channel is not available. The deflection routing consists in deviating the burst to an available output channel on an alternate link. This solution can be implemented with basic OBS equipment, but addresses the OT decision problem: If the deflection path is longer than the one originally assumed at the source, the burst may run out of OT and reach an intermediate node before being signaled (IOT). The OT decision is a crucial compromise: If it is too short, the alternate paths are rare, but if it is too large, a burst may use more resource and contributes to congestion. Adaptive solutions have been proposed in, e.g., [CEJ05a, BH07].

The second issue is the choice of the alternate route. In [Met05], several deflection criteria have been compared. Selecting the shortest path before any other criterion gives the best performances. Firstly, this is due to the fact that shortest paths spare resources. Secondly, longer paths may involve IOT loss. In [WH97], the authors proposed to deflect on the link that provides the largest number of shortest paths. Their study is limited to Manhattan networks and is within the context of OPS networks, but is clearly adapted to OBS constraints.

The efficiency of deflection routing is limited under congested scenarios. In that case, contending bursts will likely be deflected on longer paths and consequently, the deflection routing accentuates the congestion and is counter-efficient.

Fiber Delay Lines (FDLs)

If the light cannot be stored (yet), an FDL aims to delay it by extending its journey. A FDL is an optical fiber looping inside a node ([KKK02]) until the output port becomes available. Indeed, FDLs can be seen as deflection routing inside a node. With such equipment however, the travel of the header is not extended and the IOT problem is sidestepped. Moreover, delaying the burst increases the OT.

FDLs offer a sort of optical buffering since the signal can be delayed. However, it is much less

flexible than electrical buffering because the signal can only be delayed by a fixed amount of time equal to the time it takes to traverse the FDL. The design of the node is thus crucial and depends on the range of the size of the bursts. Thus, the main problem with this solution is that it relies on hardware installations that must be physically designed for a specific traffic configuration and limit the flexibility of further traffic engineering. In addition, the relevance of FDLs is mitigated by physical shortcomings (overheat, space, signal attenuation).

Wavelength Conversion

The wide multiplexing capability offered by DWDM can be used for contention resolution. With DWDM, an optical fiber can be accessed by many output ports, each emitting on a different wavelength. Thus, if the output port requested for a burst is not available, wavelength converters allow the use of an alternate output port on the requested fiber.

From a hardware point of view, three techniques are available: Signal re-emission (via opto-electro-optical – O/E/O – conversion), cross-gain modulation or four-wave mixing. Note that O/E/O conversion is the simplest solution, but its efficiency is compromised by the loss of transparency. For details on all optical techniques and their limitation, the reader can refer to Section 2.2.3 of [JV05].

Once the wavelength converters are installed, an interesting issue is the selection of one wavelength among the available ones. Several strategies are compared in [XQLX03], leading to the conclusion that the wavelength conversion should minimize the void interval between the converted burst and the preceding one on the chosen wavelength. Although wavelength converters highly improve the throughput of OBS networks, they are rarely considered due to their cost and the prototype state of all-optical wavelength converters.

Electrical Buffering

"Store and forward" can indeed be done in the electrical domain after signal conversion. The converted bursts can then be re-emitted on any wavelength at any time. This approach is clearly the most reliable, but it is often disregarded since it can dramatically impact the delay of transmission,

and it may require a tremendous buffering capacity to handle the contention: Burst may be dropped in case of buffer overflow. The loss probability here is directly related to the memory availability. If signal conversion is performed, the loss only occurs when a burst contends while the buffers are full (buffer overflow). Those limitations are pointed out in [SR09], where the authors propose to balance the effects of translucence with wavelength converters.

Conclusion on Contention Resolution

Contention resolution mechanisms operate on demand once the contention occurs. As they do not rely on pre-configuration, they preserve the reactivity of OBS. Unfortunately, none of them can guarantee 100% of contention resolution. In addition, each of them presents pros and cons. The deflection routing can be implemented without additional equipment, but it addresses the OT decision problem [CEJM05] and it is of limited efficiency in a congested state [CEJM05]. FDLs and wavelength converters increase the cost of the network and rely on experimental equipment. The burst segmentation reduces the volume of payload involved in the contention resolution by chopping the contending burst so that the contention-free part of the burst is not dropped. Recourse to electrical buffering is probably the most reliable scheme, but it can mitigate the end-to-end delay and increase the equipment cost because of the buffer requirement. In that case, loss would occur because of buffer overflow.

The contention resolution schemes can be combined to mitigate their shortcoming and increase the resolution rate [GKS04, SR09], but it remains that no guarantee can be provided.

4.3.2 Pro-active Mechanisms

Pro-active mechanisms attempt to reduce the contention probability. They usually operate in edge nodes, since they take the major decisions (burst length, departure date, wavelength, routing, offset time). Major pro-active mechanisms rely on routing (load balancing) and congestion control, but we will also mention original proposals relying on scheduling and signaling.

Load Balancing

Load balancing consists in routing the traffic so as to minimize a given metric. It is often decomposed into two phases: First, a route computation phase which determines one or more routes for each source/destination pair. Then, a route selection phase which picks up a route among the pre-computed ones. GMPLS is the protocol of choice to describe the routes in the control plane. With this framework, the entire route is chosen at the source. In [Qia00], the OBS implementation of GMPLS is described. The so-called Labeled Optical Burst Switching (LOBS) is considered as a default component of OBS in the major part of the OBS literature.

The routes can be recomputed periodically in order to adapt traffic variations ([HHM05, TVJ03, WSLW05]), whereas the route selection (if several routes are available) is usually done on demand for each burst, depending on resource availability and network feedbacks ([YR06]). The drawback of online operations is that they are based on feedbacks of the network state (usually the loss probability [LGC03a] or resource utilization [PAM+05]) as measured over a previous time interval. In addition to traffic and processing overhead, the network feedbacks entail inertia in the on-line decisions, that limits their efficiency to deal with fast variations. In [PAM+05], the routing is decided locally at each node. The link status is broadcast in a limited area so that the information of each node is more likely up-to-date.

In stable environments, off-line solutions may be accurate. The problem is that they cannot rely to actual measurements and are driven by performance metric approximations. The efficiency of the load balancing relies on (1) the accuracy of the approximation and (2) the quality of the solution (i.e., the efficiency of the resolution). For instance, the problem can be modeled accurately by a non Linear Program, but the resolution of such problems is complicated and hardly leads to the optimal solution. The resolution can be strongly simplified if the approximation formula is linear, but in that case, the model is less faithful. For example, the loss probability can be assumed to follow the resource utilization [Cou05] or the number of competing flows [BM09]. To improve the accuracy of the approximation, one can use a linear interpolation of the metric approximation [TR05]

(this option discloses a trade-off between the size of the problem and the approximation accuracy). Accurate non-linear loss approximation formula can be used in heuristics (e.g., [PCG+05]). in that case, each solution is a better evaluation of the real performances, but the process may not reach the optimal solution.

Intermediate Node Initiated Signaling (INI)

The contentions result from the one-way signaling protocols. Neglecting the reservation acknowledgment allows the improvement of the reactivity. In [KVJ03], the Intermediate Node Initiated signaling (INI) is proposed. With INI, the reservation is acknowledged by an intermediate node. This solution is a tradeoff between one-way and acknowledged protocols: Any burst is ensured to reach the initiator node, but does not have to wait for the whole round trip delay. The farther the initiator is and the more reliable the transmission is, but the longer the burst waits for the acknowledgment. An important advantage of this technique comes from the fact that the reservation up to the destination is done while returning to the source. This way, the header can collect information on the availability along the path to adjust the scheduling of the burst.

Congestion Control

OBS does not provide any delivery control, so the reliability of the transmission is the responsibility of the transport layer (mostly the TCP protocol). In TCP, reliability is achieved through retransmissions that needs to be coupled with congestion control in order to prevent the network to turn into a congestion state. In IP networks, TCP segment losses are due to buffer overflows and reflect a congestion state, but in OBS networks, transmission is triggered without acknowledgment so that burst losses can occur due to contention, even if the network is far from being congested. Thus, a loss does not give a hint on the congestion occurring in the network. Since a burst contains several TCP segments, several sources may be strongly affected by a burst loss because their window (which specifies the transmission rate) may be unnecessarily reduced, affecting the global throughput of the network. Moreover, as a burst may contain several segments from the same TCP connection, the

impact of a burst loss is even stronger. In [GK09], authors showed that the impact can be reduced by increasing the number of Aggregation Queues. This way, the number of segments of a given connection within a burst is reduced and so is the impact on the TCP flows. But this does not solve the problem. We want TCP to be aware of the congestion state of the network thanks to finer feedbacks. This is called the "false congestion detection problem" [YQL04]. In [YQL04], the authors compare three common TCP implementations (RENO, new RENO and SACK) and demonstrate the impact of OBS on TCP throughput. As a response, a new TCP, called Burst TCP (BTCP), is proposed to avoid burst losses in non-congested state to affect the TCP-window.

In [SHZ07], TCP-ENG consists in explicitly notifying the source when a loss occurs under low load. The source behavior will act as for a triple acknowledgment in TCP. This solution is efficient and also easy to implement.

- OBS layer: Edge nodes monitor the loss rate of paths. A given threshold allows distinguishing between congested states: If a loss occurs, BEN (Burst Explicit Notification) is sent to the TCP sender.

- TCP layer: At each RTT, the sender knows about the loss. If no acknowledgment is received, then a loss occurred. If instead, a BEN is received, then the sender knows that the loss is not due to congestion and will react according to the triple acknowledgment equation. The elegance of the method lies in the fact that TCP is only slightly modified.

Finally, the BAIMD protocol [SHZ07] proposes to adjust dynamically the congestion control rates to negate the side-effects of burst losses on TCP. This approach has been shown to be efficient and does not involve any specific communication between OBS nodes and the TCP senders.

Scheduling

In native OBS, the burst are sent based on local information. The edge node has a particular influence on the performances in an OBS network. As we discussed in [CHJ09b], the transparency of the OBS data plane entails that the traffic profile in the core network is completely defined at the

burst emission. In particular, the assembly process, the choice of the burst emission date and the choice of the wavelength are the responsibility of the edge nodes.

In [LQ04], the authors propose to limit the burst overlapping in the core network from the Medium Access Controller. The protocol reduces the number of wavelengths used simultaneously by bursts of a given connection. As a result, the wavelength availability increases and the contention probability decreases. This solution, however, impacts the emission delay.

4.4 Loss-less OBS

A crucial step toward OBS maturity is the capability to provide loss-less transmissions with a mechanism that preserves the reactivity of OBS. In this section, we discuss the solutions found in the literature. As we already mentioned in this chapter, no reactive solution can offer strict guarantee of transmission yet. Loss-less solutions thus rely on pro-active decisions. Indeed, each of them can be transposed to the OCS world. We will thus discuss the potential benefits of deploying OCS-like solutions with OBS equipment in terms of throughput and reactivity. The discussion will be illustrated on the scenario depicted by Figure 4.3.

4.4.1 Wavelength-Routed OBS (WR-OBS)

A major difficulty in OBS networks is that the edge nodes take important decisions based on local information, e.g., scheduling and wavelength assignment. In WR-OBS networks, a particular node – the "Central Control Node" (CCN) – maintains a global view of the network. For each burst, edge nodes request resources to the CCN, which computes the routing and wavelength assignment, and the departure date of the burst, so that it will not contend in the core network. The signaling is initiated by the CCN which notifies the involved nodes of the arrival date of the burst.

In WR-OBS networks, any operation on the data plane is initiated by the CCN. This is mandatory to ensure that it keeps a correct view of the network and computes valid solutions. This architecture avoids the contentions, but the systematic recourse to the CCN imposes a latency to any operation.

This latency can be significant, not only because of the transmission delay on the control plane, but also because of processing overhead in the CCN.

This architecture is particularly unsuited to manage link failures. Though the failure notification only consists in the notification of the CCN, nodes are not allowed to take local decision and reactive mechanisms are unusable. Moreover, the resources dedicated to the control plane and the CCN must provide a strict availability and provide efficient transmission. At the end, considering the strict availability requirement, the sensitivity to the delay and to the large traffic overhead may significantly increase the capacity dedicated to the control plane.

Finally, it is worth mentioning that such centralized architectures open serious security breaches and compromise the robustness of the network.

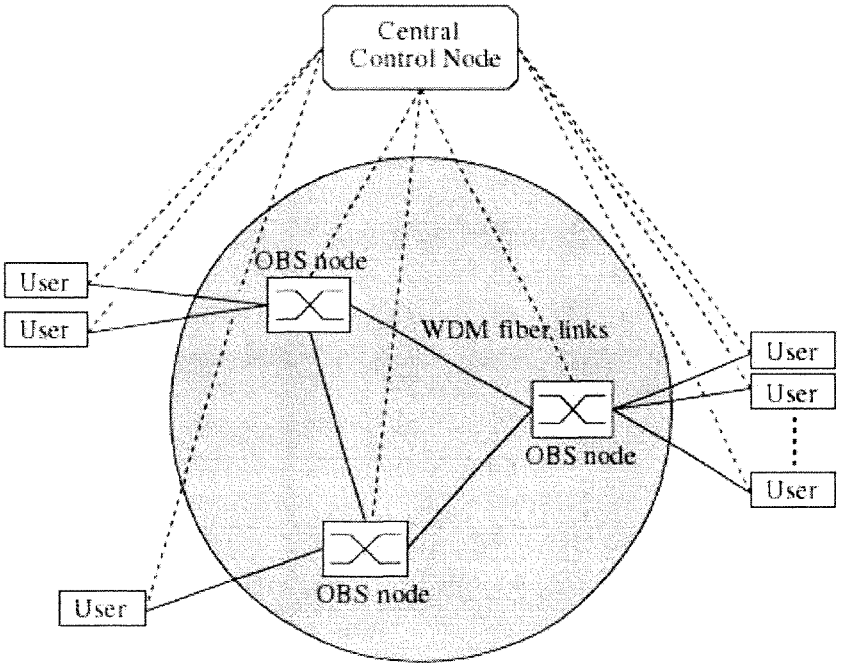


Figure 4.9: WR-OBS network Architecture

The next two solutions reduce the control plane overhead by allocating long term resources.

4.4.2 Circuit-Oriented Transmission at a Wavelength Level

Wavelength switching in the OBS framework amounts to reserve a wavelength for the duration of the considered connection, just as light-paths are established in OCS networks. The OBS framework can handle the signaling with slight modifications in the control plane [QWL06]. Once the circuit is established, the data are transmitted the same way as in OCS networks (no need for OT or aggregation). Thus, in a static scenario, this solution is equivalent to an all-optical OCS network.

Nevertheless, with the OBS equipment, the circuit establishment latency ($RTT/2+\max\{RTT/2, \delta\}$) may be reduced, in which case the reactivity would be enhanced. In addition, the control plane described in [QWL06] opens the possibility of polymorphism: Circuit switching and burst switching can be handled simultaneously with a single set of equipment. This is valuable to deal with failure since the restoration mechanisms are much more efficient in OBS networks (see Section 4.5).

The shortcoming of wavelength provisioning in OBS is the introduction of a coarse transport granularity. This solution indeed restricts a wavelength to a single flow and consequently wastes some resources. In Section 3.2, we showed that the transparency can be sacrificed to solve this issue in OCS networks: At some nodes, several incoming signals are converted to the electrical domain and re-emitted on the same output port. This solution can be transposed in OBS networks, with the difference that the information required for multiplexing is supplied by the control plane, whereas in OCS networks, it is induced by the SONET hierarchy, and imposes framing processing and synchronous transmission.

4.4.3 Circuit Oriented Transmission at a Sub-wavelength Level: Synchronous OBS

The framework proposed in [QWL06] describes the signaling of a circuit of sub-wavelength capacity. TDM is achieved via the reservation of a sequence of bursts during which the connection can emit. This solution preserves the transparency of the data plane: No framing process is required as opposed as in SONET transmission.

The transparency removes the SONET hierarchy and the circuit can be assigned any capacity. In contrast to SONET, multiplexed bursts can be separated in the optical domain, that avoids signal conversion. This is not only attractive for cost and delay consideration, but also for the flexibility, since the performances do not rely on equipment location (as opposed to SONET/SDH architectures, see Section 3.2.2).

Nevertheless, in the electrical domain, the streams can be re-organized, what highly simplify the framing and synchronization procedure. With SONET, the stream can be setup with local information. In an all-optical network, it is much more complicated. In [YLC05, YLC06], the benefits of periodic reservation are experienced, but data losses are observed. Indeed, loss-free transmission with SOBS must be supported by a centralized architecture similar to WR-OBS. It results that the same drawbacks re-emerge, except that SOBS reduces the traffic and the processing overhead on the control plane.

4.4.4 Optical Multiplexing Avoidance

Circuit switching at the wavelength level avoids sub-wavelength multiplexing to prevent contention. Nevertheless, the description of the contention in Section 4.2.1 discloses loss-free multiplexing patterns. In our example depicted on Figure 4.3, connections C_3 and C_4 can be safely multiplexed. Note that in that case the configuration is not contention-free. However, at the time a contention is discovered (in the control plane), the bursts are still in electrical domain and can be rescheduled.

In addition, connection C_2 can be multiplexed with C_4, provided bursts of C_4 have a strict priority over those of C_2. This way, the contending bursts of C_2 can be rescheduled to let the output port available for bursts in transit. This configuration is similar to the multiplexing achieved with light-trails (see Section 3.3, [GC03a]). Nevertheless, with OBS equipment, the switches do not have to be pre-configured and two successive bursts can be switched on different output ports (provided they are separated by the few nanoseconds required to reconfigure the switch). Thus, with OBS equipment, loss-free multiplexing is not restricted to linear structures (paths), but can be extended

to trees. In other words, clustered light-trails (see Section 3.3.5) can be deployed within an OBS network. Indeed, as this solution involves switch reconfigurations – and consequently fast switching equipment –, it must be seen as an OBS solution.

4.5 Restoration in OBS networks

The reactivity of OBS offers the potential to deal with failures in a reactive way with a limited impact on the service ([KB03]). The restoration process is divided into three phases (illustrated by Figure 4.10): (1) the steady state before failure, (2) the recovery phase during which the path is re-computed, and the steady state after failure. The second step is the most critical. In [SME06], the authors discuss those three phases and investigate basic deflection routing effects. In [KKY+06], routes are computed at the failed link extreme points. This way, the re-computation process begins earlier and the recovery time is decreased. Note that the source cannot choose a completely different path in that case. In [XTKE+04], it is suggested that local deflection offers the lowest blocking probability, but the alternate paths are computed in advance.

All restoration schemes involve deflection during the recovery step. In that situation, the traffic is unbalanced and part of the network may become congested. Spare networks can be set to improve the survivability, but they are, however, not mandatory ([SME06, XTKE+04, KKY+06]).

Restoration is highly reactive, but it modifies the routing configuration and consequently, it may break the routing restrictions imposed by the loss-less solutions presented in Section 4.4. An efficient restoration scheme thus must be complemented with a reliable contention resolution scheme.

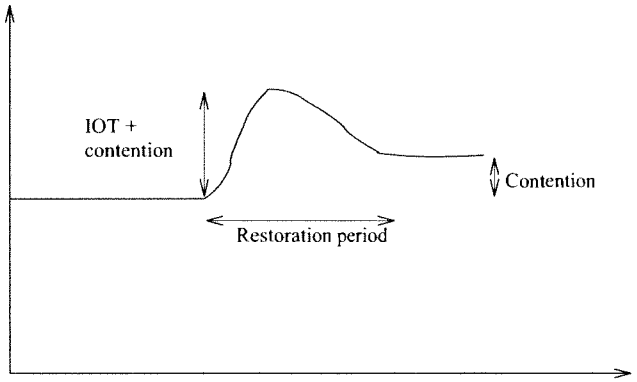


Figure 4.10: OBS restoration phases

4.6 Conclusion: Open Issues in OBS

OBS exhibits attractive features to design a dynamic optical layer, but the reactivity compromises the guarantee of successful transmissions. Preserving the reactivity of OBS imposes recourse to contention resolution mechanisms, which rely on additional immature equipment or are of mitigated efficiency. However, they cannot provide any guarantee. The loss probability can be further reduced with pro-active mechanisms, but they mitigate the OBS reactivity.

Achieving loss-less transmission in OBS networks is a very challenging issue. The challenging nature of the problem led to the consideration of hybrid systems where bursts cohabit with circuits. Polymorphism has been envisioned from the hardware point of view. Hybrid switches are endowed with two switching fabrics, a cheap one based on MEMS, handling stable connection-oriented traffic, and another one dedicated to bursty loss-tolerant applications.

Another solution consists in avoiding contentions with routing restrictions similar as those that ensure loss-free transfers in OCS networks. In [QWL06], circuit switching is achieved jointly with burst switching within a single basic OBS equipment. One can reasonably expect this architecture to be more expensive than hybrid hardwares, but it enhances the flexibility of the network since an internal link – connecting an input port to an output port – can handle either a circuit or a burst, without any specific management. This solution is close to the light-trails, except that

they are deployed over an OCS network. Indeed, light-trails perform OBS over OCS and share several attractive characteristics with OBS, but they partially sacrifice its flexibility. This hybrid architecture is circuit-oriented at the Layer 3, and burst-oriented at the layer 2. Here, loss-free transmission can be ensured by routing restrictions. We showed in Section 4.4.4 that, beside cost consideration, the fast switching capabiity of OBS devices can improve the resource utilization. In Chapter 5, we evaluate the benefits of OBS equipment versus pure light-trails.

All the loss-less solutions found in the literature rely on the off-line computation of the routing configuration, similarly as in OCS networks. Those solutions are not fully reactive since they face the same shortcoming as OCS networks regarding re-routing, which is requested in case of significant traffic variation or in case of link failure. To preserve the reactivity of OBS, the transmission guaranty must not depend on pro-active mechanisms. The reactive contention resolution mechanisms can not ensure loss-free transmissions. In Chapter 6, a deep observation of the OBS networks demonstrate that the emission process can significantly reduce the contention probability (and consequently the loss probability). Our observations point out the benefits of traffic grooming. Static grooming solutions are investigated in Chapter 7 to confirm the conclusions of Chapter 6. The solutions considered are enabled by recourse to electrical processing, which reduces the loss probability, but mitigates the benefits of traffic grooming on the delay. In Chapter 8, we propose to use traffic grooming to compensate the effect of electrical buffering, accessed dynamically by contending bursts. We propose the CAROBS framework that dynamically exploits traffic grooming in optical domain. With this framework, the effects of traffic grooming are accentuated and, due to its dynamic nature, it can be configured to solve contentions in electrical domain with no or with a limited impact on the equipment cost and on the delay.

CHAPTER 5

Loss-less transfers in All Optical Networks

The rigidity of the current optical layer is largely caused by its coarse granularity that imposes the use of synchronous transfers, unsuited to bursty traffic. In Section 3.3, we described how asynchronous all-optical multiplexing can be achieved in OCS networks. From the equipment and switching point of view, this solution, named light-trails, can be deployed over an OCS network. Indeed, a light-trail is deployed over a light-path featuring Light-trail Access Unit (LAU – see Section (see Section 2.1.4). LAUs enable "drop and continue" switching so that the sender is connected to any downstream node of the light-path. In addition, assuming adapted signaling protocols, the light-path can be accessed by any node along the light-path. Finally, a light-trail defines a directed bus, to which simultaneous access is avoided with in-advance signaling similar to the JET protocol used in OBS networks.

Therefore, a light-trail can be viewed as an hybrid architecture providing OBS over OCS transmission. Indeed, transmission over light-trails can be deployed within an OBS network. Thus, previous contributions that demonstrate the benefits of light-trails over native light-paths (e.g., regarding the throughput and the reactivity of the network) also demonstrate the benefits of OBS over an all-optical OCS network.

In Chapter 4, we illustrated the benefits that can be earned by transposing the light-trails

into an OBS network. The objective of this chapter is to measure those benefits on large scale scenarios. In Section 5.1, we present the loss-free multiplexing patterns for various architectures. In Section 5.2, we describe a generic formulation of the loss-less provisioning problem. The model – published in [CHJ09a] – can easily include alternate isolation patterns so as to describe various architectures (e.g., wavelength switching, SOBS, light-trails, OBS). The scalability of this first ILP formulation has been improved with advanced resolution schemes published in [CJH09b], and also reported in Section 5.2. In Section 5.3, our resolution schemes are evaluated and used to compare the provisioning capability of loss-less solutions in all-optical networks. Conclusions are drawn in Section 5.4, where the limitations of transparency are discussed.

5.1 Loss-free Multiplexing

5.1.1 Flow Isolation: General Case

Route r_i is isolated from route r_j, denoted by $r_i \rhd r_j$, if no burst of r_i can be lost due to bursts of r_j. By extension, denoting by C_i the connection supported by a route r_i, and B_i one of its bursts, $r_i \rhd r_j$ implies $C_i \rhd C_j$ and $B_i \rhd B_j$. They are mutually isolated, denoted by $r_i \diamond r_j$, if $r_i \rhd r_j$ and $r_i \lhd r_j$ (the negation is denoted by $r_i \bowtie r_j$).

In Section 4.2, we described the contention in OBS networks and observed that a contention only involves bursts arriving on different input ports and requesting the same output port. Thus, contentions between r_i and r_j can only occur in the nodes where r_i and r_j converge. If r_i and r_j are link disjoint or are carried on different wavelengths, then they never converge and consequently $r_i \diamond r_j$. In addition, two routes that merge at their source node are mutually isolated since both can access electrical buffers.

If only r_j starts in the merging node, then $r_j \rhd r_i$, because the bursts on r_j that are to be lost (i.e., that "lost" the contention) will be rescheduled instead. In this configuration, mutual isolation can be achieved by ensuring absolute priority to r_i. In that case, any contention results in the rescheduling on r_j and all contentions are solved. The absolute priority can be ensured with the OT

isolation or with preemptive reservation. Those two mechanisms are described in the remaining of this section.

5.1.2 Offset Time Isolation

The offset time can be exploited to disclose an additional isolation pattern, as described in Section 4.2.3.

We denote by $OT_{r_i}^v$ the remaining OT value of a burst on route r_i at node v. For given routes r_i and r_j, we define $\Delta_{r_i,r_j}^{OT} = \min_{v \in V} (OT_{r_i}^v - OT_{r_j}^v)$, where V is the set of nodes where routes r_i (carrying B_i) and r_j (carrying B_j) merge. By convention, $\Delta_{r_i,r_j}^{OT} = \infty$ if r_i and r_j are link-disjoint or if V is reduced to the first node of r_i.

In Section 4.2.3, we have illustrated the isolation via the OT priority: $r_i \rhd r_j$ if $\Delta_{r_i,r_j}^{OT} > b_j$, where b_j is the size of the bursts on route r_j. Figure 5.1 illustrates this case: C_1 competes with C_2 at node 2. Denoting by ℓ_i^2 the number of hops on route r_i from node 2 to the destination of the associated connection C_i, and assuming the OT is computed following the basic formula (see (1) in Section 4.1.1, $\Delta_{r_2,r_1}^{OT} = (\ell_2^2 - \ell_1^2) \times \delta$ (where δ is the header processing time at each node) and C_2 is isolated from C_1 if $\Delta_{r_2,r_1}^{OT} \geq b_1$.

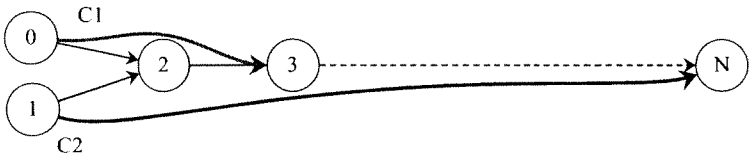


Figure 5.1: General Isolation Case: Simple Bus Topology

The isolation of a connection from another one thus depends on the length of their routes and on the length of the bursts. The use of small bursts favors the isolation. Indeed, if $\delta > b_1$, then $C_2 \rhd C_1$ if the remaining path on r_2 is at least one hop longer than the remaining path on r_1 (as measured starting at the node where the contention occurs). Now, if $2\delta > b_1 > \delta$, then the remaining distance on r_2 must be two hop longer than the remaining distance on r_1.

The assumption of short bursts, although it conforms to the suggestion of [YQ99], is worth to

be discussed at the light of recent contributions. The use of small bursts reduces the aggregation delay and improves the efficiency of some contention resolution mechanisms (e.g., with FDLs, see [Gau03]). In addition, with small bursts, the operator might exploit the observation reported in [GK09]: Spreading the incoming load among a larger number of aggregation queues reduces the impact of the synchronized loss on the goodput of TCP sources.

Opposite arguments can be set forth. In particular, longer bursts reduce the signaling overhead and smooth the traffic (see [CHJ09b, CHJ09c, CJH09a]). In addition, as any two successive bursts must be separated by a guard time to let the switch change its configuration, increasing the size of the bursts reduces the number of guard time insertion and can improve the resource utilization. In [SHM+05], the header processing time δ is around $50\mu s$, whereas the reconfiguration time is evaluated around 50 ns in [GLW+05]. Thus, if $b = \delta$, the guard time amounts 0.1% of the burst size. With 10 Gbps ports, the burst size would be 0.5 Mbit (around 150 IP packets as measured in [MC00]).

5.1.3 ALAP Reservation Protocol

If C_i benefits of the offset time priority over C_j, its bursts always "win" the contentions. However, if the contentions occur at the source of C_j, they are solved by rescheduling B_j. This solution assumes a careful and strict control of the OT of each burst and postponing the burst must be done carefully.

In the example on Figure 5.2, three bursts compete for the same output port. Bursts B_1 and B_3 are streamlined bursts in transit, whereas B_2 is an ingress burst (i.e., still in the electrical domain). We denote by b_i the duration of B_i, by OT_i its offset time and by r_i its route. In addition, the header of B_i is denoted by H_i. We assume that OTs are configured in such a way that B_1 and B_3 are isolated from B_2. Burst B_2 is aggregated at time t_2. The output port is requested from time $t'_2 = t_2 + OT_2$ to time $t'_2 + b_2$. As the output port has already been reserved by H_1 for an overlapping time slot, B_2 must be delayed in electrical buffers. Possible decisions are as follows:

- No Electrical buffering: B_2 is dropped (see Figure 5.2(A)).

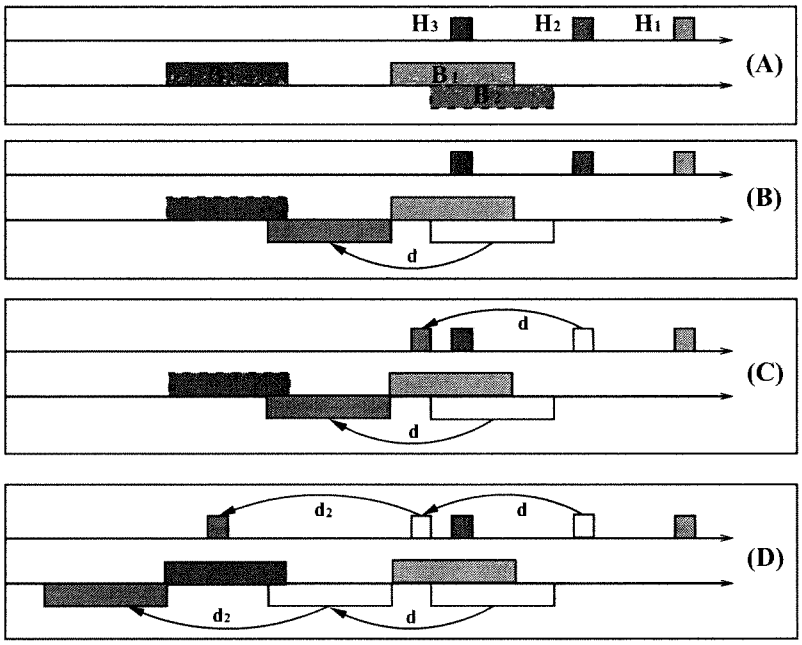


Figure 5.2: Burst insertion in OBS

- Electrical buffering with immediate reservation (As Soon As Possible (ASAP) Reservation Protocol): At time t_2, the contention is observed and the next available time slot is computed (it begins at time $t'2 + d$). If the reservation is instantaneously done, then the actual OT is increased by d (see Figure 5.2(B)). If the emission of the header is postponed as well, then the original OT is restored for the transmission (see Figure 5.2(C)). However, it remains that on the first link, the reservation has been done $OT_2 + d$ time units before the emission, although it was expected to be done only OT time units in-advance. This local favor breaks the isolation of B_3: At H_3 arrival, the resources are already reserved for B_2 and B_3 is dropped despite $B_3 \rhd B_2$.

- Electrical buffering with postponed reservation (As Late As Possible (ALAP) Reservation Protocol): In case of contention, the resource reservation is postponed by the same duration as the burst (Figure 5.2(D)). In our example, B_2 is delayed by d time units and the resource reservation is re-attempted d time units later. Consequently, the resources are still requested OT_2 time units before the scheduled burst emission date and the isolation of B_3 is respected:

H_3 arrives before the next reservation attempt of B_2 and the resources are available for B_3. Then, B_2 will contend again and will be postponed again by d_2 units of time. Finally, no burst is lost.

The ALAP burst insertion protocol reserves the resources for an ingress burst as late as possible (see Algorithm 1). The benefit is that an ingress burst, involved in a contention, will be more likely the latest signaled and electrical buffering can solve the contention. In particular, it preserves the predefined Offset Time and the possible isolation configurations of bursts in transit.

Algorithm 1 ALAP reservation protocol

Burst B of size b is ready at time t_{ready}
Compute OT; $t_{\text{res}} \leftarrow t_{\text{ready}}$; $t_{\text{send}} \leftarrow t_{\text{ready}} + OT$
repeat
  if at time t_{res}, resource is available on $[t_{\text{send}}, t_{\text{send}} + b]$ **then**
    reserve the resource
  else
    look for the next available interval $[t'_{\text{send}}, t'_{\text{send}} + b]$
    $t_{\text{send}} \leftarrow t'_{\text{send}}$; $t_{\text{res}} \leftarrow t_{\text{send}} - OT$; sleep until t_{res}
  end if
until reservation done.

The ALAP reservation protocol avoids undeserved privilege to ingress bursts, with the consequence that a burst might need several attempts before the resources are granted. This extra delay is the expense to be paid in order to reduce the loss of bursts in transit. Indeed, the postponed reservation can be viewed as a retransmission of the ingress burst, which should be preferred to the retransmission of the burst in transit. The ingress burst did not use any transport capacity and has not waited for its offset time yet. In opposition, the delay imposed by the retransmission of the bursts in transit includes the OT and the loss notification. In addition, the resources they used before the contention are wasted.

We propose to evaluate the impact of the ALAP protocol on the emission delay. Our simulations involve a source of traffic whom bursts contend with transit bursts. The ingress node runs ALAP and the measured average insertion delay of ingress bursts is reported on Figure 5.3. The x-axis represents the proportion of ingress traffic among the overall load. The extend of insertion delay due to the ALAP protocol is negligible if the ingress traffic dominates the transit traffic. Otherwise,

the increase of the extension delay remains reasonable for an overall load above 0.6 Erlang.

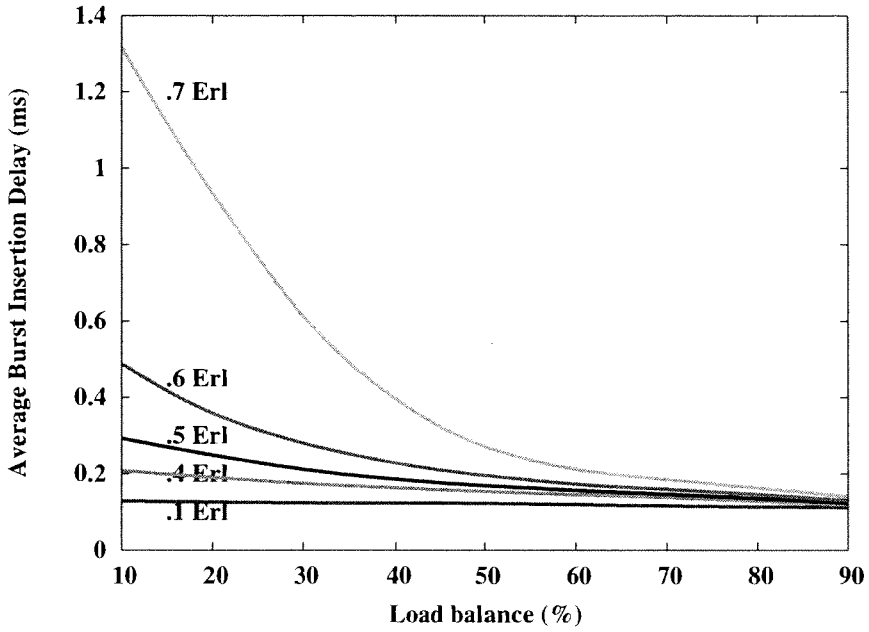


Figure 5.3: Impact of ALAP on burst insertion delay

The ALAP protocol allows the preservation of the OT priority. Thus, combining OT priority with the ALAP reservation protocol can ensure mutual isolation between two connections.

5.1.4 Preemptive Access

An alternative to OT priority is the resource preemption. This solution has been envisioned to access light-trails in OCS networks (see [GC03a] or Section 3.3.3). It can be directly transposed to OBS networks since it appeals to the same signaling protocol. Instead of ensuring that the latest signaled burst is the ingress one with the OT priority, preemption consists in forcing the retransmission of the ingress burst though it may have already been signaled. This way, the emission is interrupted so as to free the resources for the burst in transit. A possible shortcoming is that if the ingress burst has already been signaled, its header may have reserved some resources further on the path. Some of the resources will be used by the contending burst, but part of them are wasted, especially if the preempted burst goes farther than the transit one. Trailer mechanisms proposed in the context of

burst segmentation [VJ03] can be used to cancel the reservation, but they do not avoid resource wastage in the mean time.

5.2 Loss-less Provisioning: The RWA-OBS Problem

The RWA-OBS problem consists in the computation of routes, wavelengths and possibly OTs to serve the maximum number of connections without loss, in an OBS network. In this section, we first state the problem and then, we propose a generic ILP formulation that can be tuned to describe various isolation scheme.

5.2.1 Statement of the Problem and Notations

RWA-OBS is a provisioning problem. For a given traffic matrix, it aims to compute the routing configuration that maximizes either the number of granted requests or the overall amount of served traffic without loss, i.e., so that two routes r_i and r_j cannot be used simultaneously, unless $r_i \diamond r_j$.

L is the set of links and TC_ℓ is the transport capacity of the wavelengths on ℓ (assumed to be equal). We denote by F the set of flows indexed by f. For a given f, let o_f and d_f be its the origin and destination, b_f its bandwidth. W is the set of wavelengths and A is the set of possible values for the OT extension factor. An OBS route r is associated with a wavelength λ_r and an OT extension factor α_r. f_r refers to the flow that has the same source and destination as route r. $R^k_{f,\lambda,\alpha}$ is a set of k routes from o_f to d_f on wavelength λ with EOT factor α. We note R^k_f the union of all such sets related to connection f. As a result, the number of OBS routes considered is $|W| \times |\Lambda| \times |F| \times k$.

5.2.2 ILP Model

$$z^{\text{OBJ}} = \max \sum_{f \in F} s_f b_f \tag{9}$$

subject to:

$$U_{\ell,w} = \sum_{r \in R_w, p_r \ni \ell} b_{f(r)}\, x_r \leq \mathrm{TC}_\ell \qquad \ell \in L, w \in W \tag{10}$$

$$x_r \leq y_r \qquad r \in R \tag{11}$$

$$y_r + y_{r'} \leq 1 \qquad r, r' \in R\ ,\ r \bowtie r' \tag{12}$$

$$\sum_{r \in c\ell} y_r \leq 1 \qquad c\ell \in C\ell \tag{13}$$

$$\sum_{r \in R_f} x_r \geq s_f \qquad f \in F \tag{14}$$

$$x_r \in [0,1] \tag{15}$$

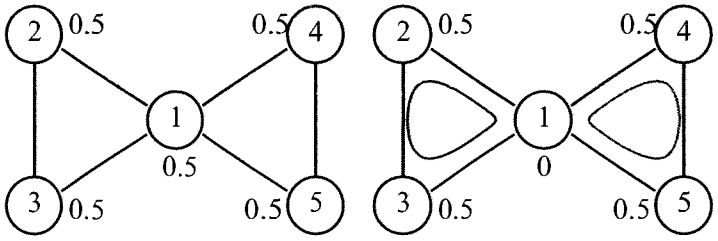
$$y_r \in \{0,1\} \tag{16}$$

$$s_f \in [0,1] \tag{17}$$

where x_r represents the fraction of flow f_r routed on route r. y_r is linked to x_r by (11) so that y_r equals one if route r is used, zero otherwise. Constraints (14) set the variables s_f to the part of flow f served with loss-less guarantee. The capacity constraints (10) prevent the wavelength overload. Constraints (12) prevent non isolated routes from being used simultaneously.

Constraints 13 are added to improve the LP relaxation. We consider the route conflict graph where a node is associated with a route and two nodes are linked if the associated routes are not isolated. A clique $c\ell$ is a complete sub-graph (any two nodes of the sub-graph are connected) and $C\ell$ is the set of all maximum cliques. The isolation constraints consist in allowing the use of no more than one route per clique. Figure 5.4 illustrates the improvement of the LP relaxation earned by the clique formulation. This conflict graph exhibits two cliques. As a sequel, no more than two connections can be served, by using one route of each clique (2 or 3 and 4 or 5). In that example, route 1 should not be used since it belongs to both the cliques. Without (13), the LP relaxation affects 0.5 to each route (Figure 4(a)) whereas (13) avoids the use of route 1 (Figure 4(b)).

The problem can be tuned via the definition of the isolation operator $\diamond$ to describe various



(a) Formulation without (13) (b) Formulation with (13)

Figure 5.4: Isolation Formulation and LP Relaxation

architectures. Wavelength switching imposes that no two connections use the same wavelength of a link. Thus, two routes are isolated if they are link disjoint or if they use different wavelengths. OCS light-trails are described by defining isolated routes to those which never converge nor diverge. With OBS equipment, divergence is allowed. Light-trails can be described with the OT isolation, or without, assuming preemptive access to the medium.

Note that the grade of service depends on the multiplexing possibilities. Thus, an upper bound of the problem can be reached by considering that all routes are isolated, whereas wavelength switching architecture provide a lower bound (no multiplexing within a wavelength is allowed).

5.2.3 Iterative Resolution

The number of considered routes directly impact the optimal value of the problem: Considering more paths and more EOT values improves the solution, but severely increases the resolution time. To sidestep this compromise, we propose an iterative greedy heuristic (IGH) that consists in iteratively increasing the set of considered paths (Algorithm 2). At each iteration, one new route per flow is added in the model, the problem is solved starting from previous iteration solution and the variables related to the used routes are frozen. Note that the symmetrical nature of the problem can be broken by iterating the resolution along the wavelength, i.e., solving over an increasing of wavelengths.

After the last iteration, one can either unfreeze all the variables and solve the problem (which means solving the initial problem with a good initial solution) or solve the problem only considering the route that has been used at a given iteration.

Algorithm 2 Iterative Resolution of RWA-OBS

Compute R_f^k for each connection f.
repeat
  insert a new route on λ into $R_{f,W'}$ for each flow f
  solve RWA-OBS
  freeze and tag the routes used
until No route can be added OR all flows are served
unfreeze all tagged routes
freeze all untagged routes (suboptimal solution)
solve RWA-OBS from the current solution

5.2.4 Column Generation Formulation

The iterative resolution described in Algo 2 accommodates the scalability issue due to the number of paths. However, the RWA-OBS problem suffers from its symmetrical structure: There is a huge number of equivalent solutions, all deductible from each other up to a wavelength permutation. It is known that such problems are getting quickly not scalable when, e.g., the number of wavelengths increases in the case of our study. Nevertheless, turning to the large scale optimization formulation, there is a way to set a model where the number of solutions is greatly reduced, and such that they are not equivalent up to a permutation of some indexing.

Master Problem

In order to set such a model, let us define a RWA-OBS configuration as a set of routes that can be established on the same wavelength, i.e., routes that are mutually isolated. Let $\mathcal{C}$ be the set of all possible configurations, indexed by $c \in \mathcal{C}$. Note that, for scalability reasons, the set R of considered routes will be limited to the first k shortest routes, for each pair of source and destination nodes with positive demand. The model can be written as follows:

$$z^{\text{OBJ}} = \max \sum_{f \in F} s_f b_f + \epsilon \sum_{c \in C} \varphi_c$$

subject to:

$$\sum_{c \in \mathcal{C}} z_c \leq W \tag{18}$$

$$s_f - \sum_{c \in \mathcal{C}} \sum_{r \in R} f_r^c x_r^c \leq 0 \qquad f \in F \tag{19}$$

$$\sum_{r} x_r^c f_r^c b_f = \varphi_c \qquad c \in C \tag{20}$$

$$x_r^c \leq z_c \qquad c \in C, r \in R \tag{21}$$

$$z_c \in \{0,1\} \qquad c \in C \tag{22}$$

$$s_f \in [0,1] \qquad f \in F \tag{23}$$

$$x_r^c \in [0,1] \qquad r \in R \tag{24}$$

where f_r^c is the maximum fraction of demand b_f served by route r, on configuration c. We call x_r^c the fraction of f_r^c actually activated. s_f is the fraction of flow f that is accommodated (19). Configuration c is activated if $\exists r$ such that $x_r^c \times f_r^c > 0$ (21). In that case, $z_c = 1$. As each configuration is activated on a different wavelength, the number of activated configurations must remain lower than W (18).

Pricing Problem

The number of valid configurations is the number of stable sets in the conflict graph. To accommodate this scalability problem, we will solve the LP relaxation of the master problem with a subset of configurations. Two approaches can be envisioned to build the subset: Either we proceed with an off-line enumeration of the configurations, or we generate them on-line. If we go ahead with off-line enumeration, only a subset of configurations can be generated, among the most promising ones, meaning we need a criterion to identify those. The on-line generation – that is the direction we will move along – relies on the use the column generation techniques to govern an efficient on-line

configuration generation. The auxiliary problem (pricing) inserts a new configuration to the incumbent set of configuration used by the master problem only if it improves the current value of the objective function, i.e., if it is a so-called augmenting configuration. The pricing problem, in charge of the generation of "augmenting" configurations, aims to maximize the sum of the reduced costs of each connection. The objective can be written as follows:

$$Z = \max -u_0 + \sum_f u_f s_f + \varepsilon(\sum_{f \in F} \sum_{r \in R_f} x_r b_f)$$

where u_0 and u_f are the dual variables associated with constraint (18) and (19) respectively. The last term operates as a secondary objective that adds as much traffic as possible in the best configuration. For example, a route isolated from every route should appear in every configuration, regardless of the associated reduced cost.

The set of constraints is the same as in the original model described in Section 5.2.2, but the set of routes is restricted to a single wavelength (its cardinality thus decreases down to $|F| \times |\Lambda| \times k$). Expression of the constraints is as follows:

$$x_r \leq y_r \qquad r \in R \tag{25}$$

$$\sum_f \sum_{r \in R_f, r \ni \ell} b_f x_r \geq \mathrm{TC}_\ell \qquad \ell \in L \tag{26}$$

$$\sum_{r \in c\ell} y_r \leq 1 \qquad c\ell \in C\ell \tag{27}$$

$$\sum_{r \in R_f} x_r = s_f \qquad f \in F \tag{28}$$

$$s_f \in [0,1] \qquad f \in F \tag{29}$$

$$y_r \in \{0,1\} \qquad r \in R \tag{30}$$

$$x_r \in [0,1] \qquad r \in R \tag{31}$$

Branching

The column generation stops when the pricing problem cannot generate a column that improves the LP relaxation of the master problem. The problem is converted into a MIP, i.e., z_c are converted to integer variables and solved with the current subset of configurations. Nevertheless, reaching the optimality of the LP relaxation does not ensure that the subset of configurations leads to the optimal solution of the MIP version of the problem. This phenomenon can be illustrated by Figure 5.4. Let us assume that routes 2 and 4 serve the flow f and routes 3 and 5 serve the flow f'. Only one wavelength is available. The optimal solution is 2 in both LP and ILP versions of the problem. However, the LP optimality can be reached with configurations $c = [0, 1, 0, 1, 0]$ and $c' = [0, 0, 1, 0, 1]$ if $z_c = z_{c'} = 0.5$. With this set of configurations, however, the ILP version of the master problem imposes the use of only one configuration and consequently, only one flow will be served. The gap between the optimal value of the LP relaxation and the optimal value of the ILP problem may be unavoidable. In our example, nevertheless, $c = [0, 1, 1, 0, 0]$ leads to the same optimal solution as the LP relaxation, which proves that the optimality is attained.

To force the master problem to remain integer feasible in spite of the relaxation of the integer constraints, we propose to freeze some configurations in the master problem to integer value. The management of the frozen configurations is done via a Tabu list. The size of the Tabu list is set to $2 \times |W|$. If the LP relaxation solution involves more than $|W|$ wavelengths, then the configuration c that handles the highest amount of traffic is added to the Tabu list and z_c is set to 1. The content of the Tabu list is swapped. The $W - 1$ latest inserted configurations are frozen in state "active" ($z_c = 1$), whereas the $W + 1$ earliest inserted ones are frozen in state "unused" ($z_c = 0$) until they are discarded from the Tabu list.

Table 4 illustrates the Tabu list management with $|W| = 3$. z_a and z_b are frozen in state "active". When z_c is inserted, it is set "active". The Tabu list content is swapped and z_a is frozen in state "unused". z_a cannot be used until it is discarded from the Tabu list (at the insertion of z_g).

-	-	-	-	$z_a = 1$	$z_b = 1$
-	-	-	$z_a = 0$	$z_b = 1$	$z_c = 1$
...					
$z_a = 0$	$z_b = 0$	$z_c = 0$	$z_d = 0$	$z_e = 1$	$z_f = 1$
$z_b = 0$	$z_c = 0$	$z_d = 0$	$z_e = 0$	$z_f = 1$	$z_g = 1$

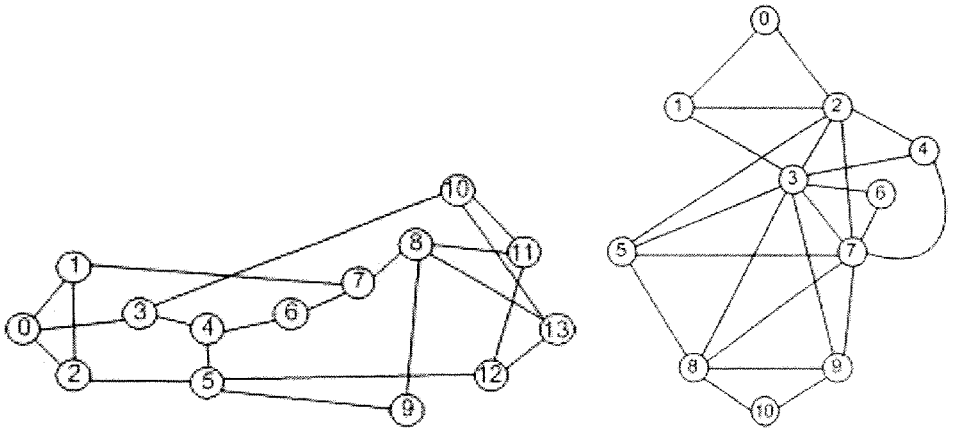
Table 4: Tabu illustration with 3 wavelengths

Algorithm 3 Column Generation Tuning

repeat
  Solve the master LP relaxation
  if number of configuration used > W **then**
    add the most useful configuration in the Tabu list
    Solve the master LP relaxation
  end if
  Solve the pricing
  Add the new configuration in the master problem
until Time limit exceeds
Unfreeze all configurations
Convert master problem to ILP
Solve the master ILP

5.3 Experimental Results

In this section, we report numerical results obtained on the NSF network (Figure 5(a)) and the New Jersey LATA network (Figure 5(b)) . We assume $b < \delta$ as discussed in Section 5.1.2. The load is equal between each pair of nodes and expressed as a ratio of the wavelength transport capacity (10 Gbps).



(a) The NSF Network Topology (b) The New Jersey LATA Topology

Figure 5.5: Network Topologies

5.3.1 Multiplexing and Grade of Service

The grade of guaranteed service – i.e., the rate of connection granted with loss-less guarantee – is directly related to the degree of multiplexing – the number of routes that can be used simultaneously. For a given topology, the degree of multiplexing decreases when the load of the connections increases, because of the capacity constraint. In particular, if the data-rate of the connections is equal to the transport capacity of the wavelengths, then multiplexing within a wavelength is impossible, and the grade of service will be the same for all architectures.

The grade of service obtained in that case is equal to the grade of service of wavelength switched networks, since in those networks, the contentions are avoided by discarding multiplexing beyond the wavelength. In that case, no two routes sharing a wavelength on a link can be used simultaneously, regardless of the load of the connections. However, with soft flows, the capacity constraint is less restrictive and the loss-free multiplexing patterns presented in this chapter can be exploited.

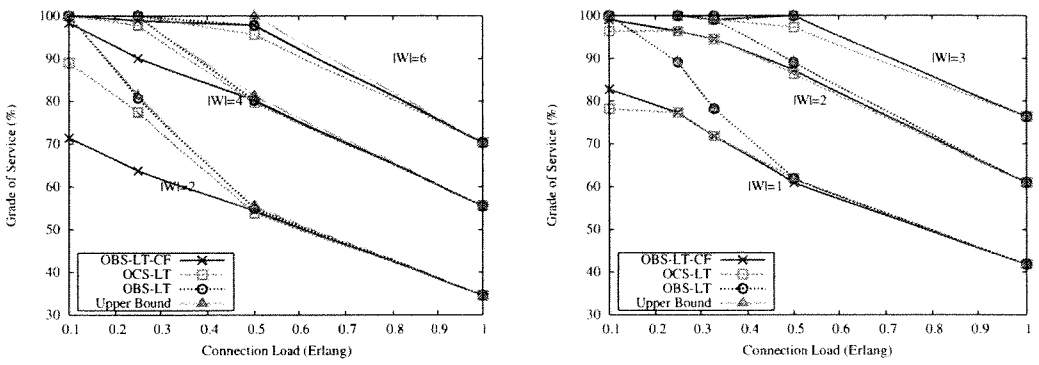
In an all-optical OCS network, multiplexing can be achieved over linear light-trails (i.e., over paths). With the fast switching capability of the OBS equipment, the multiplexing is extended to clustered light-trails (i.e., over trees). It is unavoidable that clustered light-trails outperform linear light-trails, but they require more expensive equipment. We propose here to explore the performance aspect of the question by comparing the grade of guaranteed service achieved in both cases. We also report an upper bound, obtained by relaxing the isolation constraints. Thus, any two routes are mutually isolated and the multiplexing degree is maximum.

Figure 6(a) and Figure 6(b) reports the grade of service obtained respectively on the NSF network and on the LATA network, depicted on Figure 5.5. With high load, the capacity constraint reduces the multiplexing possibilities and the upper bound is reached by both OBS and OCS. With soft connections, the multiplexing is improved because the capacity constraint is less restrictive. In this situation, the difference between the architectures can be observed, especially with a small number of wavelengths.

Let us first assume that the isolation over the light-trails is ensured by a reactive mechanism such

as the premptive access described in Section 5.1.4. With the possibility to deploy light-trails over trees, OBS enhances the grade of service of OCS by 10% in the NSF network. This improvement is probably not imperssive enough to justify the investment on fast switching equipment. However, in the LATA network, the improvement increases up to 20%. This is due to the higher connectivity of the LATA network, that consequently exhibits much more tree structures, so that clustered light-trails can be exploited more intensively.

If the preemptive access is disabled, then the isolation must be ensured via the Offset Time. With this OT restriction, the multiplexing possibilities – i.e., the number of isolated routes – is reduced. In addition, to prevent unacceptably long OTs, we set $A = \{1, 2\}$ (the multiplicative extension factor of the OT is not larger than 2). The degradation of the performances is observed in the same instances where OCS is outperformed bu OBS. In the NSF network, the OT restriction strongly affects the provisioning capability, with up to 30% of reduction. The worst case on the LATA network is less dramatic (20%), because of the stronger connectivity of the network. In the LATA network, the OT restriction puts OBS networks at the level of OCS networks with preemptive access.



(a) Grade of Service on the NSF Network

(b) Grade of Service on the LATE Network

Figure 5.6: Grade of Loss-less Service

5.3.2 Resolution Scheme Comparison

The exact model presented in Section 5.2.2 is heavy and requires large computation time. In this section, we compare the computation time and the quality of the solutions obtained with the iterative greedy heuristic (IGH) and the column generation approach (CG-RWA-OBS). Table 5 reports the computing time and the gap between the heuristic values and the optimal value obtained with the exact ILP formulation. From the results, we can identify "easy instances" and "hard instances". Instances with a connection load of 0.5 Erlang are more difficult to solve because the symmetry of the problem is increased. If the load is equal to the wavelength transport capacity, then the wavelength capacity constraints prevent multiplexing and constraints (12) and (13) can be relaxed. On the opposite, the capacity constraints are useless if the load is equal to 0.1 Erlang. The instances with $W = 4$ are also difficult because they are the largest instances where the full provisioning is impossible. The exact solution of the ILP formulation is highly dependent on the number of wavelengths. Indeed, the optimal solution of the difficult instances could not be obtained in reasonable time. IGH clearly outperforms the ILP formulation in terms of computing time but the quality of the IGH solution may not be satisfactory, especially for "hard instances" (with a 5% to 7.5% gap). CG-RWA-OBS improves the quality of the solution, often up to the optimal solution. Computing times are comparable with easy instances, and although it requires longer computing times for "hard instances", they remain reasonable, especially when compared with the exact ILP model.

5.4 Conclusion

In this chapter, we have presented and compared loss-less all-optical architectures. In OCS networks, synchronous transfer is impossible all-optically. Thus, we considered the deployment of light-trails. We showed that the transposition of light-trails to OBS networks allows the enhancement of the grade of service close to the upper bound. Those architecture assume that contentions are solved re-actively by giving absolute priority to the optical burst. With this solution, some resources may be reserved in vain. This can be avoided by exploiting the OT priority. Nevertheless, a very large

	$\|W\|$	Load (Erlang)		
		0.1	0.5	1
ILP (s)	2	36	$>$ 2 weeks	1
	4	89,161	$>$ 2 weeks	174
	6	5,274	$>$ 2 weeks	6,820
	8	9,094	123,300	5,703
	10	20,133	106,531	8,236
IGH (s)	2	3	458	1
	4	7	683	3
	6	22	6,628	5
	8	9	2,200	7
	10	12	752	19
$\text{Gap} = \frac{z^{\text{ILP}} - z^{\text{IGH}}}{z^{\text{ILP}}}$	2	2.4 %	5.0 %	0
	4	7.5 %	6.1 %	1.0 %
	6	0.6 %	5.1 %	0.8 %
	8	0	1.1	3.4 %
	10	0	0	3.1 %
CG (s)	2	6	24,290	1
	4	5,349	67,878	2
	6	319	76,822	2
	8	6	447	3
	10	6	56	3 , 223
$\text{Gap} = \frac{z^{\text{ILP}} - z^{\text{CG}}}{z^{\text{ILP}}}$	2	0	1.2 %	0
	4	1.2 %	0	0
	6	0.5 %	0	0
	8	0	1.1 %	0.75 %
	10	0	0	0.70 % , 0

Table 5: Resolution Scheme Comparison

OT may be required and we propose to include a limit on the OT in our model to control the trade-off between the delay and the grade of service.

The original formulation we proposed in [CHJ09a] suffered of a poor scalability and long resolution time. In [CJH09b], we proposed an iterative heuristic and a column generation formulation to struggle with the shortcoming of the initial formulation. The heuristic can quickly provide acceptable solutions, but the column generation should be preferred since it leads to near-optimal solutions at the expense, in worst cases, of a reasonable increase of the resolution times.

The asynchronous transfer is well suited to bursty traffic. Nevertheless, the solution proposed here is not fully reactive because it is based on off-line routing computation. In addition, the impact on the delay and fairness is significant. To avoid the rigidity of off-line solutions, we propose to explore a reactive solution supported by a translucent architecture. Recourse to electrical buffering can be used to modify a burst in time, space and spectral domain and is thus highly effective to solve contention. However, losses can still occur due to buffer overflows. Therefore, the main issue addressed by translucence is the buffer occupancy. In Chapter 6, a careful observation of the traffic in OBS networks suggests the relevance of traffic grooming to balance the impact of signal conversion. This is confirmed in Chapter 7, where we rpesent the traffic grooming in translucent OBS networks. In Chapter 8, we propose the CAROBS transmission scheme that intensively exploits traffic grooming, and we demonstrate that CAROBS, used in a translucent architecture, can provided loss-free transmission, without mitigating the assets of OBS regarding reactivity.

CHAPTER 6

A Further Observation of OBS Networks

6.1 Observation of the Loss Process in OBS Core Networks

OBS networks are quite apart of previous architectures. First, its bufferless packet-oriented nature is unprecedented. Second, the in-advance one way signaling protocol has been specifically designed and impacts the loss process. In this section, we propose to observe the traffic in a core network to identify favorable patterns and drive the design of a reliable OBS network.

6.1.1 OBS Transparency Property

The first important feature can be expressed by "flows of different connections never mix". In a network with buffers, mixing different flows modifies their characteristics. Indeed, multiplexing different flows, which usually "smooths" their characteristics, has absolutely no effect in an OBS core network.

In an OBS network, the switches are pre-configured so that the data plane is switched all optically. The bursts thus cut-through without being delayed. It entails that: *(i)* The gap between two successive bursts is not affected by a node traversal and each sequence of bursts remains unchanged when going through a node – except that some bursts can be discarded due to unresolved collisions

; *(ii)* Only the distance between two nodes determines the delay a given packet encounters and the number of traversed switches does not matter.

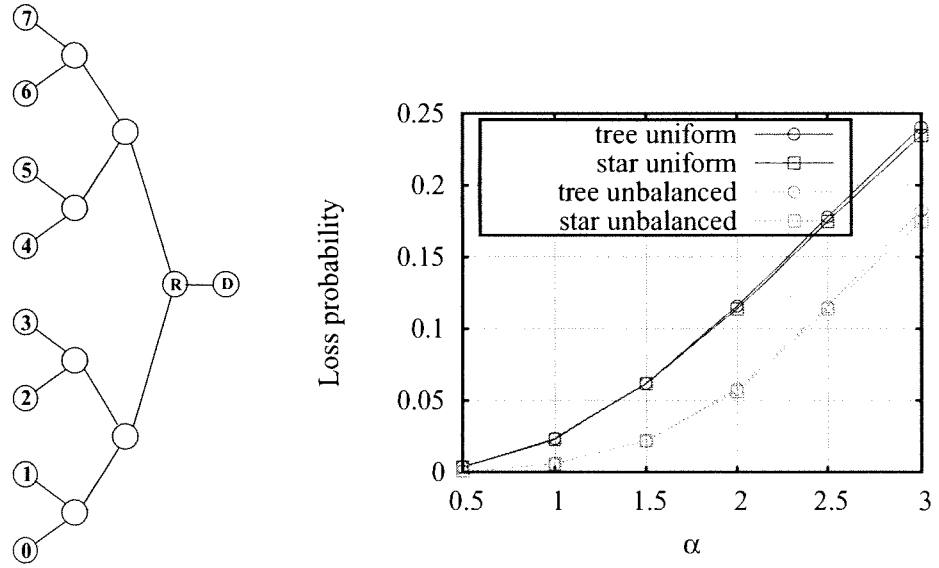
The implication is significant. Consider, for instance, the tree topology represented in Figure 1(a). Under the assumption that any outgoing link (links are directed downwards) has enough wavelengths to solve every contention (i.e., the number of overlapping bursts is always lower than the number of wavelengths), the traffic observed on link $r \rightarrow d$ is identical to the traffic observed on a derived star topology where each leave is directly connected to r with the same path length. Indeed, if two bursts overlap somewhere in the tree, they will overlap until the last link $r \rightarrow d$. As the travel time is not modified in the star topology, they also overlap in the star topology. Conversely, if two bursts overlap in the star topology, then they also overlap somewhere in the tree, up to the last link.

Now if, at some given time, the overlapping bursts outnumber the number of wavelengths, one (or more) burst must be dropped. In the star topology, loss occurs exclusively at r whereas it can occur at any node in the tree. The bursts are thus not dropped in the same order and the loss probability is not necessarily the same.

Let us experiment the accuracy of the star approximation derived from the tree represented in Figure 1(a). Each leaf sends identical traffic to node d (Poisson arrival of bursts of constant size arriving with exponentially distributed τ_i). In the "uniform scenario", the overall load is uniformly distributed among all sources whereas in the "unbalanced scenario", the load is distributed so that $a_i = 2 \times a_{i-1}$. The bursts are processed in the order of their arrival and each link has $W = 3$ wavelengths. Figure 1(b) reports the overall loss probability of both scenarios on the tree, and the derived star topologies. The initial overall load of 8 Gbps is tuned by a multiplicative coefficient (α).

The star topology is shown to provide an accurate estimation of the loss probability obtained with the tree topology in both balanced and unbalanced scenarios. A very small gap appears under very high load (i.e., when the loss probability is unacceptably high). Except for those scenarios of limited interest, the star topology can be used as a faithful approximation. It can considerably

simplify the studies and reduce the simulation speeds, as it reduces the number of iterations of the fixed point method [RVZW03].



(a) A 3-stage binary tree

(b) Loss Probability on Tree and Star Topologies

Figure 6.1: Illustration of the Transparency in OBS Networks

Those results corroborate the statement that the transparency of the data plane preserves the statistical characteristics of the traffic in a core network. It results that the traffic therein observed is completely defined at the emission. This observation is very attractive and drove our study. In the remainder of this section, we propose to identify impacting parameters and favorable core traffic profile in order to formulate recommendations regarding the emission process.

6.1.2 Loss Independent Arrival Property (LIA)

The bursts aggregated by a source of traffic are emitted by a network access point without knowledge about the fate of the previously sent bursts. The only information comes through acknowledgments, arriving at arbitrary intervals, which are too large to be of any help for real time traffic management. In particular, a source that sends bursts, keeps on sending bursts even if some of them are rejected in the core network due to collisions.

Because of the transparency property, such a behavior remains unchanged in the core network:

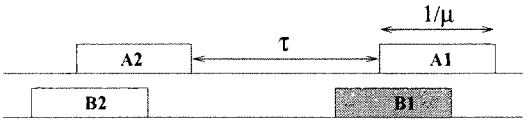


Figure 6.2: Core Network Traffic Arrival

The arrival of a burst submitted by an input port to a set of output ports is independent of the state of the output ports and the fate of the previous bursts. In addition, the bursts submitted by a given input port cannot overlap since they have been served by the same output port in the previous node. As a consequence, the traffic arrival in an OBS core switch can be described by the diagram on Figure 6.2. The input port either submits a burst of average duration $1/\mu$ or waits τ time units before submitting the next burst. Due to statistical multiplexing, several connections can be superimposed and an input port can submit bursts from a large number of connections. Hence, the burst submission process of an input port is a mix of several arrival processes.

We denote by $\lambda = 1/\tau$ the birth rate of an idle input port and $\bar{\lambda} = 1/(\tau + 1/\mu)$ the corresponding arrival rate. The burst submission will remain the same, whether the burst is dropped or served. In the example, burst B_1 is dropped, but this event has no impact on the next arrival. In the sequel, an input port is referred to as *active* for the whole duration of the burst, whether the burst is successfully transmitted or not. This behavior differs from the one described by Engset models (see 4.2.2), where a source switches to state "idle" as soon as a burst is rejected.

When we observe a group of N input ports, a direct consequence of the above remark is that, under the simplified model we assume for traffics, the number of active ports is distributed according to a binomial law:

$$P(n) = \binom{N}{n} \alpha^n (1-\alpha)^{N-n}, \tag{32}$$

where α stands for the load offered by a source:

$$\alpha = \frac{1}{\mu(1/\mu + \tau)} = \frac{1}{1 + \tau\mu}.$$

Note here that, in accordance to the LIA property, the behavior of the source is not related to dropping process of the system, as opposed to Engset models (see Section 4.2.2).

6.1.3 OBS Core Network Traffic Model (LCH$^+$)

Traditional models assume that requests (packets, requests for connection, etc.) arrive according to a Poisson process where Erlang models applies. In an OBS core network, however, any node multiplexes a finite (usually small) number of incoming links and Erlang model must be replaced by models that take advantage of the finite number of sources such as the streamline effect [PCG$^+$05] or the Engset model (see, e.g., [FH04, Sys86]). Those models, described in Section 4.2.2 do not reflect the "loss independent arrival" property (LIA). The model proposed in [ZYL04b] can be easily adapted to reflect the independent arrival property, but describes Poisson processes. As far as we know, only the "Lost Call Held" model (LCH) [Sys86, Vit95] combines the finite number of sources and the LIA property, but, unfortunately, it describes a system where segmentation is performed [VZJC02]. We describe below a variation (LCH$^+$) that discards the burst segmentation. We assume exponentially distributed burst sizes and define the load of source i by $\alpha_i = 1/(1 + \tau_i\mu_i)$ and its birth rate while in idle state by $\lambda_i = 1/\tau_i$ according to Section 6.1.2.

Because of the independent arrival property (LIA), the state of an input port is independent of the state of the system. In addition, the transparency assumption implies that if a burst is served, the input port and the output port remain both active as long as the input port keeps on receiving data (i.e., for the whole duration of the burst). Those two properties imply that an input port is active for the whole duration of its burst, whether it is served or dropped.

Figure 6.3 describes the model of an outgoing link with W output ports (servers) and S identical input ports (sources). In state (i, j), i servers are busy (and associated with i input ports currently receiving data) and j bursts are being dropped (i.e., j input ports are currently receiving data to be dropped). Consequently, $i + j$ sources are active and $S - (i + j)$ sources (qualified as *idle*) can submit a new burst (the birth rate is $(S - i - j)\lambda$). If a wavelength is available ($i < W$), the next

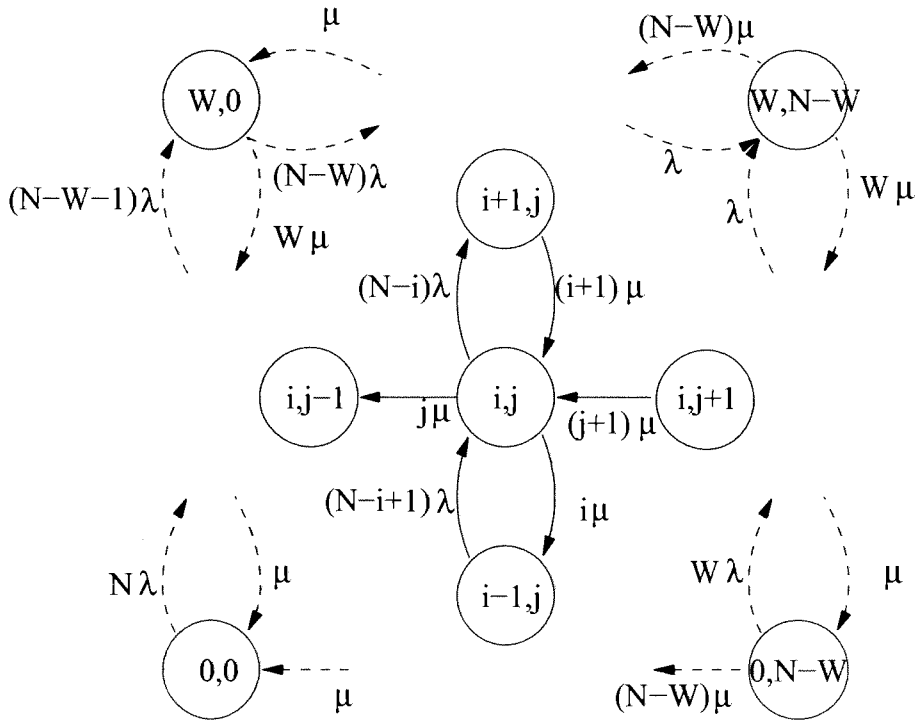


Figure 6.3: LCH$^+$ Model

burst is served and the system switches to state $(i+1, j)$ (the input port is connected to a server and remains active for the duration of the burst). Otherwise ($i = W$), the burst is dropped and the next state is $(W, j+1)$. Here again, the input port remains active as long as it keeps on receiving bursts, although they are dropped (as opposed to an Engset system where an input port would stay *idle*).

When an input port stops receiving data, it becomes *idle* ($i+j$ is decreased by one). If its burst has been dropped (rate $j\mu$), a "dropped input port" becomes *idle* and the next state is $(i, j-1)$. Otherwise (rate $i\mu$), a "served input port" becomes *idle* and a server is released, leading to state $(i-1, j)$. Such a behavior differs from the LCH model where a server that becomes available can serve the remaining of a burst being dropped (the system switches to state $(i, j-1)$ if $j > 0$ and to state $i-1, 0$ otherwise).

The solution of the LCH$^+$ model provides the probability $P_{i,j}$ for the system to be in state (i, j). A client is rejected if it is submitted while the system is in any state (W, j). The loss probability is

the ratio of the rejection rate over the total submission rate:

$$\mathrm{LCH}^{+}(\alpha, N, W) = \frac{\sum_{i=0}^{N-W-1} (N - (W+i))P_{W,i}}{\sum_{w=0}^{W} \sum_{i=0}^{N-W-1} (N - (w+i))P_{w,i}}. \quad (33)$$

6.1.4 Impact of a Finite Number of Incoming Flows

Consider an optical link ℓ with two wavelengths at 10Gbps. Let an overall incoming load of 4 Gbps be equally distributed among S input ports requesting ℓ and following the behavior described in Section 6.1.2. The loss probability predictions reported on Figure 6.4 validate the relevance of the LCH^{+} and Engset models whereas the streamline formula [PCG+05] quickly converges toward the Erlang-B loss formula that overestimates the loss.

This simple experiment illustrates that, at equal load, it is beneficial to reduce the number of sources. In the context of our study, i.e., a core node, a source is an active input port. Assuming full wavelength conversion, permutation of the wavelengths in the wavelength assignment leads to an equivalent configuration and, if we systematically select the available wavelength with the lowest index, the number of sources seen by an optical link ℓ equates the maximum number of overlapping bursts requesting ℓ. This value is referred to as the overlapping degree of link ℓ and is denoted by Ω_ℓ.

Observation 1: At constant load, as in classical queuing models, a smaller number of streams improves the performance.

6.1.5 Impact of the Burst Size

We consider two streams S_1 and S_2 of load a_1 and a_2 in a system with one wavelength ($W = 1$). The overall load A stays constant, equal to 2 Gbps, but the contribution of each connection varies (x-axis represents $\frac{a_2}{a_1 + a_2}$).

Several sets of experiments are plotted on Figures 5(a) and 5(b). For each experiment, $b_1 = 1$ and b_2 is increased. The values of b_2 are reported on the corresponding curve on Figure 5(a) where

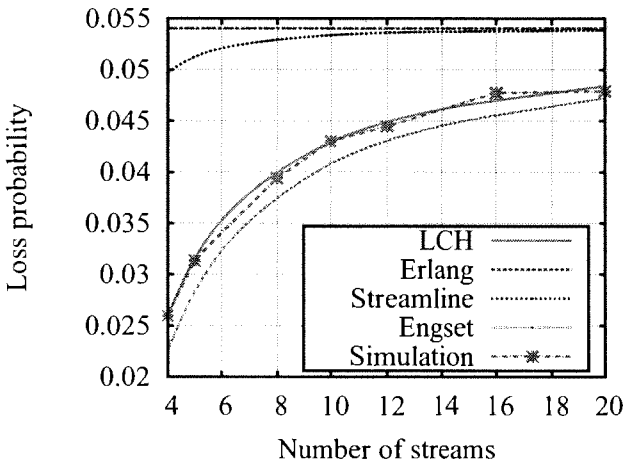


Figure 6.4: Influence of the number of streams

the overall loss probability is depicted. The line styles of Figure 5(a) are preserved in Figure 5(b) where the loss probability of each connection is represented.

With equal burst sizes, the worst case is obtained for the balanced configuration where $a_1 = a_2$. Unbalancing the traffic reduces the loss probability of the dominant stream and increases the one of the soft stream, resulting in the reduction of the overall loss probability.

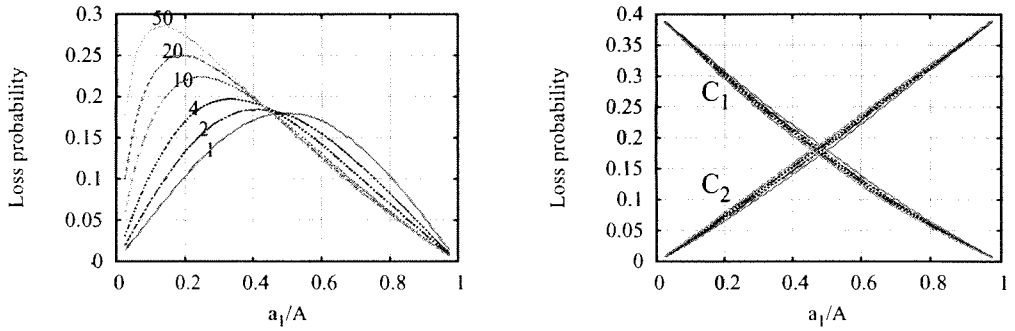
The size ratio has no significant impact on the individual source loss probability, but clearly influences the overall performances. At equal load, the arrival rate decreases as the burst size ratio increases. Thus, the overall dropping probability is more impacted by the stream with the shortest bursts. A consequence of this variation is that the worst case moves away from the configuration with equal loads.

Observation 2: The worst case loss probability increases with the gap between b_2 and b_1.

Observation 3: For any fixed value of b_1/b_2, the actual size of the bursts has no influence on the loss probability.

6.2 Emission Process in OBS

The observation of the traffic in the core network presented in the previous section promoted the study of the emission process. Indeed, the transparency property gives the exclusive responsibility of



(a) Influence of b_2 on overall loss ($b_1 = 1$)

(b) Influence of b_2 on connection loss ($b_1 = 1$)

Figure 6.5: Influence of Aggregation Parameters on Loss Probability

the traffic profile to the emission process. The emission process consists in the aggregation process and the access to the medium (including scheduling, OT decision and wavelength selection). In this section, we describe the burst emission and identify configurations that can reduce the overlapping degree in the core network.

6.2.1 Aggregation

Aggregation takes place in aggregation queues (AQs) managed by ingress nodes. Each AQ is assigned a particular combination of traffic characteristics – usually an optical egress node and possibly a class of service – and stores the incoming IP packets that match with those characteristics. Once the aggregation is completed, the packets in the AQ are grouped into a burst sent to the MAC layer so that they are all signaled by a single header.

The burst is triggered either once the AQ reaches a given size or after a timer expires. The former criterion supplies bursts of equal sizes but the aggregation duration increases for low loaded AQs. The latter criterion bounds the aggregation duration, but low loaded AQs generate small bursts and increase the signaling overhead. A combination of both criteria is commonly accepted as a reasonable compromise [CLCQ02]. Several variations have been proposed to improve the aggregation process (e.g., [HWMA03]) but all basically conform with the described mechanism.

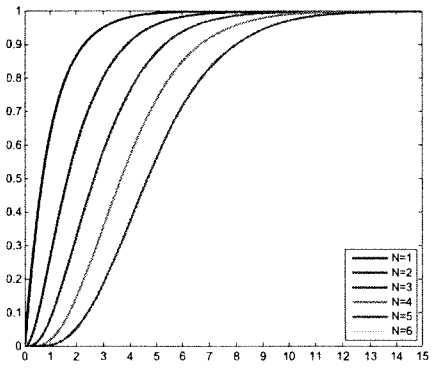
Delay consideration apart, several contributions recommend to set the same burst size for all

the sources (e.g., [YCQ02, LQXX04]), as confirmed by Observation 2 in Section 6.1.5. We assume here that all bursts have the same size Np, where p is the size of the packets arriving according to a Poisson process and N is the number of packets. The service rate is denoted by $\mu = C/(Np)$ where C is the transport capacity of the wavelengths. The mean aggregation duration is equal to N/λ and the burst are submitted to the MAC layer at a rate $\Lambda = \lambda/N$ following an Erlang-N distribution, which Cumulative Distribution Function (CDF) is written as follows:

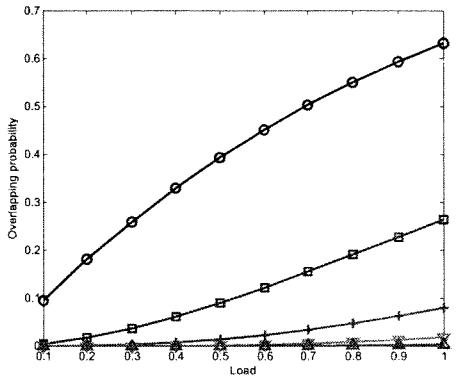
$$f(x, N, \lambda) = 1 - \sum_{n=0}^{N-1} e^{-\lambda x}(\lambda x)^n/n!. \tag{34}$$

Figure 6(a) plots the CDF of an Erlang-N distribution with packets arrival rate $\Lambda = 1$ and various bursts size N. Two bursts overlap if their inter-arrival is shorter than their duration ($1/\mu = Np/C$), i.e., with probability $f(1/\mu, N, \lambda)$. Figure 6(b) plots such a probability regarding N and the load of the AQ. It reports that the overlapping probability decreases with larger N. In other words, building longer bursts smooths the traffic. This statement is confirmed by Figure 6(c) that plots the overlapping degree (Ω_ℓ) measured on a simulation of the aggregation process. Figure 6(c) also reveals a very interesting property: For a given load, Ω_ℓ is reduced if less AQs are involved. For example, let us consider a load of 1.6 Erlang that fills bursts of 10 packets ($N = 10$). If the load is handled by a single AQ, up to three bursts will overlap at the exit of the aggregation process. Now, if the load is spread among several AQs, the resulting overlapping degree increases to 4, 4 and 16 with respectively 2, 4 and 8 equally loaded AQs.

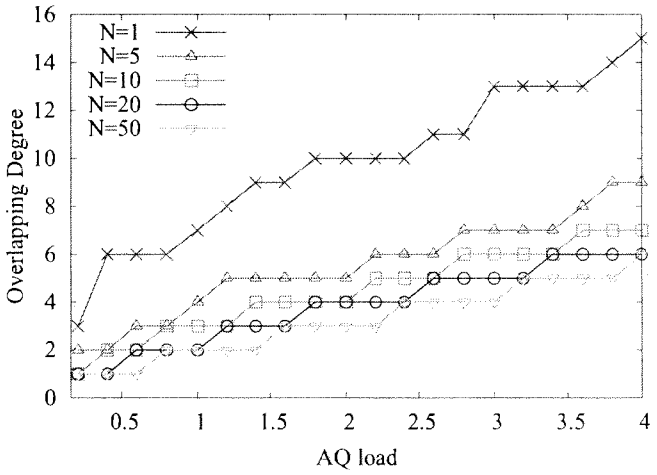
We now propose an experiment to confirm the Observation 1 – i.e., the relation between the overlapping degree and the loss probability. The simulation involves two ingress nodes connected to a core node v. Each ingress node generates a load of 1.6 Erlang competing for four wavelengths in v. Incoming packets are uniformly spread among 1, 2, 4, 8 or 16 AQs that generate bursts of size N packets. The burst, once aggregated, is scheduled on the first available period. Figure 7(a) plots the loss probability and confirms its correlation with the overlapping degree. Concentrating AQs systematically reduces the loss. Accordingly to the behavior of the Erlang-N distribution, increasing



(a) Erlang-N CDF (Cumulative distribution function)



(b) Overlapping Probability

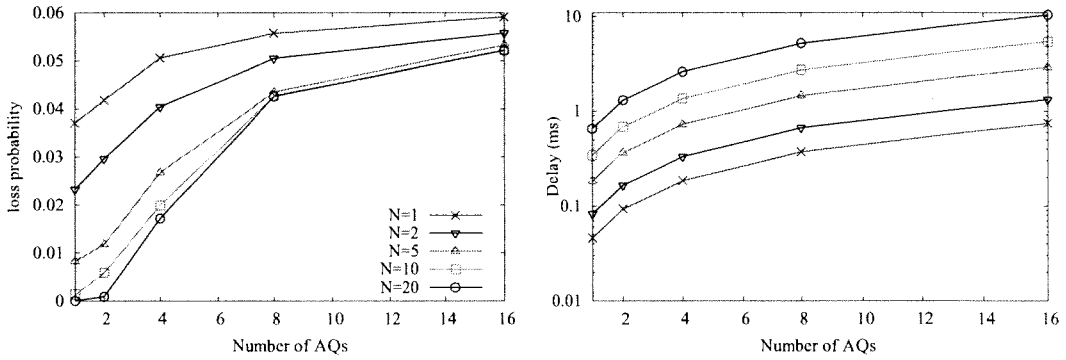


(c) Overlapping Degree

Figure 6.6: Observation of the Traffic Generated by the Aggregation Process

the size of the bursts also helps reducing the loss rate, but the effect decreases when increasing N.

The aggregation delay reported on Figure 7(b) illustrates the impact of the aggregation parameters on the aggregation process. Increasing N naturally increases the aggregation delay linearly. On the opposite, reducing the number of AQs speeds-up the aggregation process and reduces its delay. This is due to the fact that the resulting AQ will be more loaded. Thus, merging AQs is highly beneficial because it directly reduces the overlapping degree, but also it allows the increase the length of the bursts while keeping similar delays in order to accentuate the benefits.



(a) Loss (b) Delay

Figure 6.7: Impact of the Aggregation Process on the Performances in a OBS Core Network

6.2.2 Medium Access

The MAC layer is in charge of the wavelength assignment and the scheduling of newly aggregated bursts. Denoting by t_{READY} the date at which the burst is aggregated and by OT its offset time, the medium access controller looks for an available wavelength, from time $t_{\text{READY}} + OT$, for the whole duration of the burst.

In the mean time, the burst is stored in electrical buffers and the emission can be postponed if no wavelengths are available, or in order to control the profile of the emitted traffic. Especially, the MAC layer can serialize the emission of the bursts to limit the overlapping degree to a maximum value $\Omega^{\max}$. In that case, a burst is not scheduled as long as $\Omega^{\max}$ or more bursts are being emitted. The overlapping degree in the MAC layer is indeed defined by the aggregation process and the MAC can further smooth the traffic. In [LQXX04], emission serialization has been shown effective to reduce the contention probability.

This method, however, impacts the time spent in the MAC buffers and, consequently, the buffering capacity requirements. Figure 6.8 plots the average medium access delay regarding the load of the AQ and the overlapping restriction $\Omega^{\max}$ for $N = 10$. It suggests the relevance of using aggregation to smooth the traffic. For example, with $\Omega^{\max} = 2$, bursts generated by a single AQ loaded at 1 Erlangs will spend about 0.1ms in the MAC whereas they will spend 0.7ms if the load is spread among two equally loaded AQs restricted to one wavelength each.

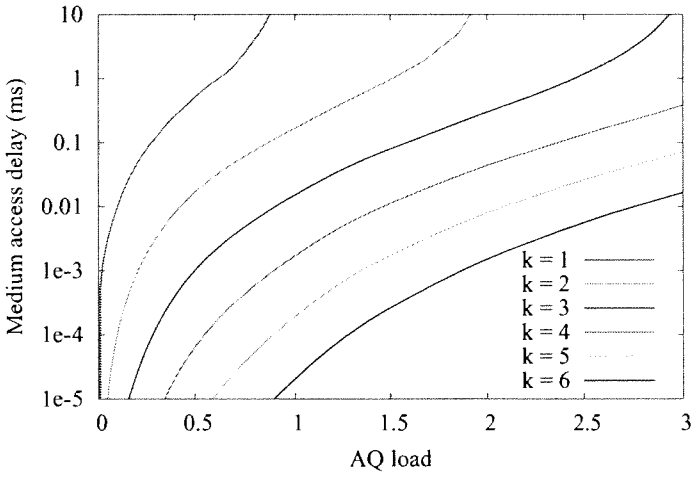


Figure 6.8: Medium Access Delay with Overlap Restriction

6.3 Recommendations Regarding the Aggregation Process

The transparency property entrusts the traffic profile in the core network to the emission process: The traffic emitted by the edge nodes preserves its statistical properties all along its journey in the core network. A careful observation of the performances of an OBS core node highlighted the importance to reduce the overlapping degrees in the core network.

The study of the emission process led to the identification of two parameters related to the overlapping degree: (1) the length of the bursts, and (2) the number of aggregation processes. Thus, traffic smoothing can be done by merging several AQs together. The so-called "AQ-grooming" also improves the aggregation process and reduces its duration. The main difficulty relies to the fact that AQs are defined according to the topology, since, at each node, at least one queue is required for each destination. As a sequel, the AQ grooming may lead to mix payload supplied by different AQs imposes demultiplexing operations in the core network.

In OCS networks, multiplexing the traffic of several flows within a light-path is referred to as "traffic grooming". In the next chapter, we describe how multiplexing and demultiplexing can be operated in electrical domain in the context of OCS networks. Then, we transpose this architecture into an OBS network.

CHAPTER 7

Traffic Grooming through Electrical Domain

In Chapter 6, we demonstrated the importance of the emission process in OBS networks: Because of the data plane transparency, the traffic profile in the core network is completely defined at its emission. The observation of the traffic in the core network has highlighted the benefits that can be expected by merging several AQs: The improvement of the aggregation process and traffic smoothing.

Mixing heterogeneous traffic of different AQs (i.e., with different egress nodes) within a burst can be assimilated to "traffic grooming". "Traffic grooming" has been initially proposed in the context of OCS networks. Therein, it consists in multiplexing heterogeneous traffic (i.e., with different egress nodes) within a single wavelength. The groomed traffic is only discriminated in some nodes where demultiplexing occurs to separate the data with different egress nodes. Besides those nodes, the groomed traffic is served without any reconfiguration of the optical layer. The "de-grooming" is complicated in optical domains, hence, it is commonly done in electrical domain.

Recourse to electrical processing of the data plane however addresses shortcomings, reflected by the increase of the end-to-end delay. Therefore, with translucent architectures, the benefits of traffic grooming must be put into perspective with the impact of electrical processing. In this chapter, we first present the OCS architectures that perform traffic grooming. Then, we transpose those architectures into an OBS network. Finally, we discuss the tradeoff between electrical processing

and traffic grooming.

7.1 Traffic Grooming in OCS Networks

In an OCS network, the finest granularity managed by the optical devices is the wavelength. The data are transmitted along a so-called light-path, which connects two nodes on a given wavelength. "Traffic grooming" in OCS networks refers to multiplexing several flows, i.e., data with different ingress node and/or egress node, within a light-path.

7.1.1 Synchronous Traffic Grooming with SONET/SDH

The SONET protocol is a widely deployed standard. A light-path is split into several segments connected by advanced equipment that can process the signal in the electrical domain. On a segment, the signal is transported all-optically, i.e., switched at the wavelength level. The end-nodes of the segments operate in the electrical domain and can switch at a sub-wavelength granularity to carry out multiplexing and demultiplexing.

The performance of the network highly depends on the efficiency of the end-nodes to operate multiplexing and demultiplexing, i.e., the efficiency to discriminate the sub-flows transported on an incoming signal. With the SONET/SDH protocols, the traffic is identified via a "clock". The transmission is synchronous, and each timeslot is dedicated to a single connection. The end-nodes only have to agree on the arrangement of the frames prior to the transmission in order to be able to discriminate them.

As the multiplexing is operated in electrical domain, the involved traffic is indeed re-emitted at that point, and scheduled so as to avoid the contentions. Back to the example described in Figure 4.3, a solution is depicted on Figure 7.1. Grooming equipment are installed in nodes C and E to perform multiplexing and demultiplexing operations, respectively. With this configuration, all the requests can be served on the same wavelength.

This architecture hardly handles dynamic traffic, because the logical connectivity is ruled by the configuration of the switches and consequently, any change of the logical connectivity is subject to

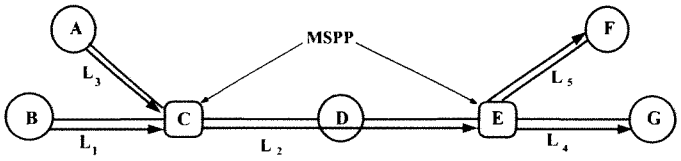


Figure 7.1: Provisioning Scenario of Figure 4.3 with SONET (OCS)

significant latency. Let us assume the arrival of a new connection between D and E. As D is not equipped with SONET/SDH interfaces, it cannot access the wavelength dedicated to the light-path L_2 and the new connection initiates the establishment of a new light-path on a second wavelength from D to E. As a result, the transmission is delayed by the latency of light-path establishment. Moreover, the new light-path will be restricted to traffic starting at node D.

The shortcoming of SONET/SDH is caused by the fact that only the end-nodes of a segment are logically connected. Intermediate nodes, as they are unable to synchronize their emission with the transiting flow, cannot access the resources. To sidestep this issue, the resource utilization must be known by the intermediate nodes. This mandates "in-advance" signaling protocols. This solution, referred to as light-trails, is discussed in the next section.

7.1.2 Asynchronous Traffic Grooming with Light-trails

A light-trail transforms a light-path into a directed bus (see Section 3.3): Once a light-path is established, it can be accessed asynchronously by any of its nodes to transmit to one or more downstream node(s) on the light-path. A network provisioning transmission over a light-trail can be viewed as an "OBS over OCS" network. The transmission is done through OCS equipment, but it is not circuit oriented. A Light-trail Access Unit (LAU, see 2.1.4) is added to the nodes to enable asynchronous access to the light-trail in a OBS fashion. From the OCS perspective, light-trails are likened to traffic grooming since they can carry several flows without optical switch reconfiguration.

The establishment of a connection over a light-trail consists in setting signal blockers appropriately to let the signal reach the destination. Conflicting accesses to the medium can be avoided with signaling mechanisms similar to the JET protocol.

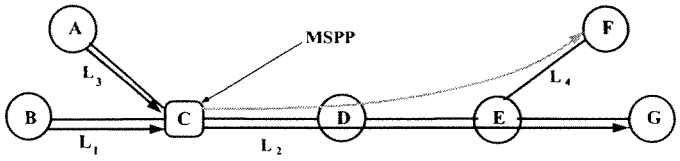


Figure 7.2: Provisioning Scenario of Figure 4.3 with Light-trails and MSPP (OCS)

An important advantage of light-trails over SONET/SDH is the "drop and continue" feature: At any node, the signal can be retrieved and either blocked or forwarded. This enables multiplexing flows that travel along the same path. A particular attention must be paid on the access to the medium in order to prevent conflicting transmissions (see Section 3.3.3). The problematic part is the multiplexing of converging flows, that can be provided through recourse to electrical processing. This is illustrated in Figure 7.2. In node C, the incoming flows contend for link $C \rightarrow D$, but the contention can be avoided assuming the incoming bursts are rescheduled from the electrical domain. Light-trails L_2 and L_4 are established from C to G and F respectively. Node C sends the traffic of C_2 and C_4 on L_4, whereas the traffic of C_3 is carried by L_2. Finally, the traffic of C_1 can use either L_2 or L_4. The set of light-trails provides a full connectivity between any two nodes (in one direction). Thus, any new connection could be multiplexed on the existing light-trails.

7.1.3 Conclusion

Light-trails open the possibility to perform all-optical traffic grooming in OCS networks, by transmitting bursts over light-paths. Thus, traffic grooming is achieved at the level of the wavelength. An OBS network manages the transmission of bursts. After each burst, the switch is reconfigured to serve the next burst. Thus, traffic grooming must be achieved at the level of the burst, especially if it is intended to exploit the benefits enlightened in Chapter 6.

7.2 Aggregation of Heterogeneous Bursts

In this section, we describe how the aggregation process can be improved by mixing heterogeneous traffic in a burst. The described mechanisms are illustrated on the example depicted in Figure 7.3.

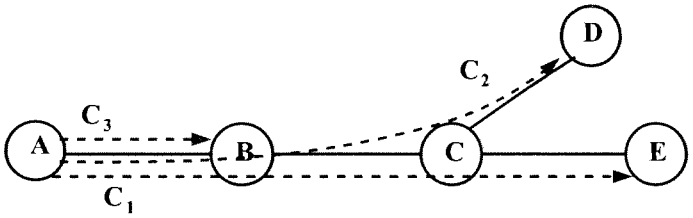


Figure 7.3: A Sample Topology to Illustrate Traffic Grooming

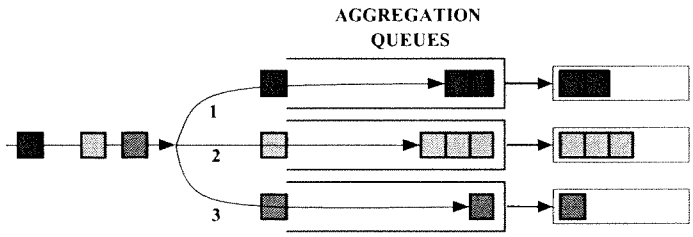


Figure 7.4: Traffic Grooming at the Application Level

7.2.1 Aggregation Process

From the application point of view, the aggregation process can be seen as traffic grooming since, at a given hierarchical level, the aggregation process mixes several packets that share similar characteristics, but possibly generated by distinct processes. The aggregation process typically groups data with the same destination. Figure 7.4 illustrates the aggregation in Node A of the example depicted in Figure 7.3. Node A manages three AQs (one for each flow), which trigger their bursts independently.

With a pure size-based aggregation, the average time spent in AQ_i (associated with connection C_i) is equal to B/α_i, where α_i denotes the load of connection C_i and B the size of the bursts. In Chapter 6, we showed that increasing B improves the traffic smoothing, but increases the aggregation delay. The aggregation delay can be bounded by triggering bursts before they reach the expected size B, according to a timer. However, in that case, softly loaded AQs generate small bursts.

Traffic grooming in the aggregation consists in mixing traffic from different AQs within a burst. In the remainder of this section, we describe how traffic grooming can be used to avoid the emission of small bursts and speed-up the aggregation process.

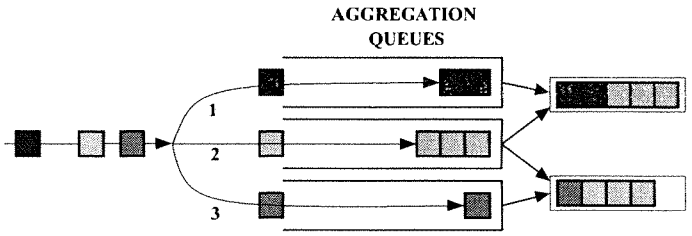


Figure 7.5: Traffic Grooming at the Burst Level

7.2.2 Collaborative Aggregation Processes

In [FZJ05], the small bursts triggered by lightly loaded AQs are complemented with the payload of some other AQs. The resulting burst is composed of several sub-bursts supplied by different AQs, thus, possibly with different egress nodes. The burst is sent to the destination of the first sub-burst, in accordance to the classical OBS scheme. Therein, the first sub-burst is discarded from the burst and the remainder of the burst is forwarded toward the next destination node. The demultiplexing operation will be described in Section 7.3

The process is illustrated in Figure 7.5, that illustrates the aggregation in Node A of the scenario presented by Figure 7.3: Once the timer of AQ_2 expires, its content is complemented with the content of AQ_1 or AQ_3 to lengthen the resulting burst, sent to Node B (the destination of C_2). Thus, the aggregation process of AQ_1 and AQ_3 is preempted and the aggregation delay is reduced. In addition, as the emission of the traffic emanating from different AQs is serialized, the overlapping degree of the emitted traffic is reduced.

However, whereas this solution is significantly effective in case of softly loaded AQs, the benefits are mitigated when the load increases because traffic grooming only occurs in case of time-based triggered bursts, i.e., in case an AQ is unable to fill-up a burst before the timer has expired.

7.2.3 Aggregation by Pools

In order to increase traffic grooming, and to extend it to highly loaded scenarios, the size-based triggering should take into account collaborative burst filling. Aggregation by pool proposes to associate the triggering events to a set of AQs instead of a single AQ (Figure 8.1): AQs are grouped

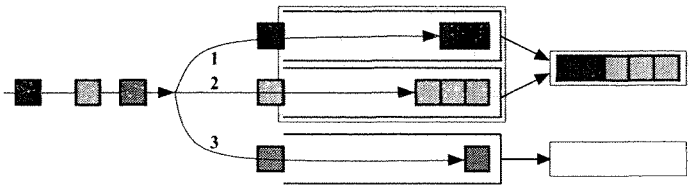


Figure 7.6: Aggregation by Pool

into aggregation pools (APs). An AP triggers a burst once the combined load of its AQs reaches the maximum burst size. The classical aggregation process is improved here because several AQs contribute to the filling of the burst. The burst is filled up more quickly and, if the burst triggering is time-driven, then the collaboration of several AQs increases its size (as in [FZJ05]).

In the example depicted in Figure 7.6, node A manages two APs: AP_E contains AQ_1 and AQ_2, and AP_D contains AQ_3. Although neither AQ_1 nor AQ_2 fulfill the size-based criterion, they can trigger their payload without waiting for the time-based criterion since, together, they can fill-up the burst.

The efficiency of traffic grooming depends on the AP definition. A larger number of AQ included in each AP increases the load of the AP and consequently improves the aggregation process. Nevertheless, the resulting burst will include more sub-bursts and, as a result of this increased heterogeneity, the number of demultiplexing operation will increase. Those operations can mitigate the benefits of traffic grooming. This issue is discussed in the next section.

7.3 Demultiplexing Operations

The payload emanating from different AQs must be discriminated somewhere in the core network. Thus, the generation of heterogeneous bursts imposes bursts demultiplexing. The node where demultiplexing takes place is decided at the emission and depends on the content of the burst.

7.3.1 Simple Discard

In [FZJ05], the demultiplexing is done in the destination nodes of each sub-burst. The burst is converted toward electrical domain and the sub-burst that reached its destination is transmitted to the IP sink while the remaining of the burst is re-emitted, possibly with handling local sub-bursts of ingress traffic. This policy imposes that all sub-bursts travel together up to their respective destination. As a sequel, in the example depicted in Figure 7.3, if a burst contains a sub-burst of each connection, then the traffic destined to node E will use a longer path since it must accompaign the traffic of connection C_3 to node D before reaching node E. In addition, each time a sub-burst is discarded, the remaining sub-bursts are delayed by signal conversion and re-emission. Nevertheless, the numerical results reported in [FZJ05] demonstrate the benefits of traffic grooming on the end-to-end delay, which is reduced when the number of sub-bursts included in the bursts increases. The most significant impact, however, relies on the loss probability, which is strongly reduced because a burst that discards a sub-burst accesses electrical buffering and, consequently, is protected from contention.

With such a translucent architecture, recourse to electrical buffering reduces the loss probability and its impact on the end-to-end delay is counterbalanced by traffic grooming. Consequently, it is relevant to enhance traffic grooming in order to increase the acceptable recourse to electrical buffering (i.e., the amount of payload that can be converted without impacting the end-to-end delay). With the solution described in this section, traffic grooming possibly increases the distance traveled by the payload. This drawback is solved in the next section.

7.3.2 Burst Re-aggregation

To avoid the extension of the path, in [CHJ09c, CJH09a], we considered the intermediate re-aggregation in the core network. Figure 7.7 represents an OBS core node capable to operate burst re-aggregation. It operates as an edge node and as a core node.

Ingress data are placed in the aggregation module, until the queue contains enough data to fill

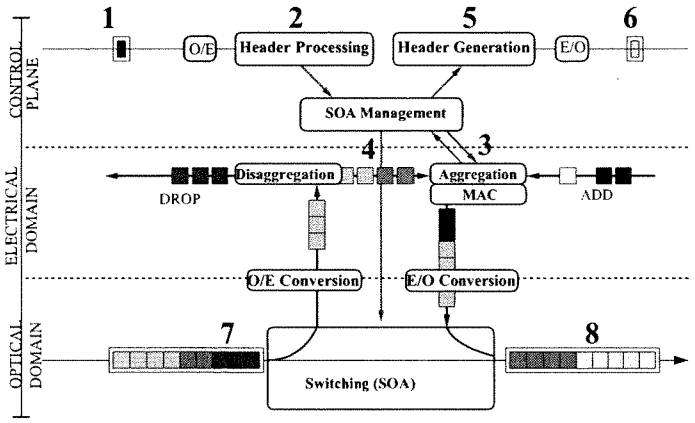


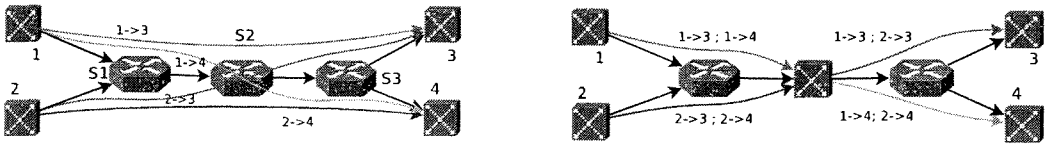
Figure 7.7: Re-aggregation Capable Node

up a burst. Bursts arriving from optical ports either cut-through the switch toward the next node or they are converted toward electrical domain and submitted to the disaggregation module. The payload that reached its destination is sent to the IP sink and the remainder is submitted to the local aggregation process, just as ingress traffic. That traffic is thus mixed with ingress traffic and, by contributing to the local aggregation process, increases traffic grooming.

The re-aggregation allows to separate sub-bursts that must be forwarded to different output ports. In our example, a burst composed of sub-bursts of C_1 and C_2 does not have to reach both nodes D and E. It is rather de-aggregated in node C (or B) and the sub-bursts are re-emitted independently.

The described node is very similar to a classical OBS node that acts both as an edge and a core node. The only hardware modification is the connection between the disaggregation and the aggregation module. The impact on the equipment cost only relies on additional electrical buffering capacity requirements. On the one hand, the buffers handle more traffic, but on the other hand, the traffic spends less time in the aggregation queue. This tradeoff must be discussed simultaneously with the delay, which faces the same compromise.

On the scenario depicted in Figure 7.8, we assume traffic from nodes v_1 and v_2 to nodes v_3 and v_4. $C_{i,j}$ denotes the connection from v_i to v_j. With end-to-end aggregation (Figure 8(a)), each ingress node manages one AQ per destination and the bursts compete in switch S_1 to reach S_2. If



(a) End-to-End Aggregation

(b) Intermediate Re-aggregation

Figure 7.8: Aggregation Grooming with Re-aggregation

re-aggregation is featured in S_2 (Figure 8(b)), ingress node v_i only maintains one AP that mixes packets of $C_{i,3}$ and $C_{i,4}$ in bursts sent to S_2. Note that contention can still occur in node S_1. The surviving bursts are disassembled in S_2, that maintains one AP per destination, regardless of the origin.

The benefits of traffic grooming are expected in the reduction of the overlapping degree on links connecting S_1. Let us set $N = 10$ and the connection load of 0.4. According to Figure 6(c), the maximum number of overlapping bursts arriving in S_1 (denoted by Ω_{S_1}) is eight (two for each connection). With re-aggregation featured in S_2, aggregation grooming reduces the number of APs to 2, and Ω_{S_1} to six (three per incoming flow of S_1). As loss occurs when the overlapping degree outnumbers the wavelengths, traffic grooming is expected to reduce the loss probability (as previously illustrated in Section 6.2.1).

7.4 Evaluation of the Trade-off between Translucence and Traffic Grooming

Our primary objective here is to confront the antagonist effects of re-aggregation and traffic grooming and to discuss the resulting trade-off. In this section, we describe a simple sequential greedy strategy to plan the grooming operations. This simple strategy aims to intensify the traffic grooming, but does not ensure loss-less transmission.

7.4.1 Grooming Configuration Used

We call $\mathcal{R}$ the set of nodes where re-aggregation takes place. We assume this set is built off-line. AQs $A_1^{v'}$ and $A_2^{v'}$ managed by node v' are included in the pool AP^{v',v_r} if $v_r \in \mathcal{R}$ and $A_1^{v'}$ and $A_2^{v'}$ are associated with the same path between v' and v_r. In other words, if node $v_r \in (R)$, all the AQs of any node v' whose path includes v_r and share the same sub-path from v' to v_r are merged in the same AP, denoted AP^{v',v_r}.

The bursts of AP^{v',v_r} are sent toward v_r, where re-aggregation occurs. This definition of the pools aims at maximizing the number of AQ included in each AP so as to minimize the aggregation delay. With this pool definition, any node performing intermediate re-aggregation will systematically re-aggregate all incoming bursts.

Locating the re-aggregation processes is a key issue. Loss-less configuration can be obtained with the tools developed for the GRWA problem in the context of SONET/SDH networks (see Section 3.2.2). However, whereas the objective of "classical" GRWA tools is to maximize the number of granted connections with a limited access to the grooming operations, we showed that in OBS networks, the traffic grooming must be accentuated in order to increase its benefits. In the scope of this chapter, we propose to perform re-aggregation in the nodes that are used by the largest number of AQs. This approach tends to minimize the resulting number of APs, and conforms to the conclusions of Chapter 6. This simple greedy strategy is used in the next section to discuss the trade-off between traffic grooming and electrical processing.

7.4.2 Numerical Results

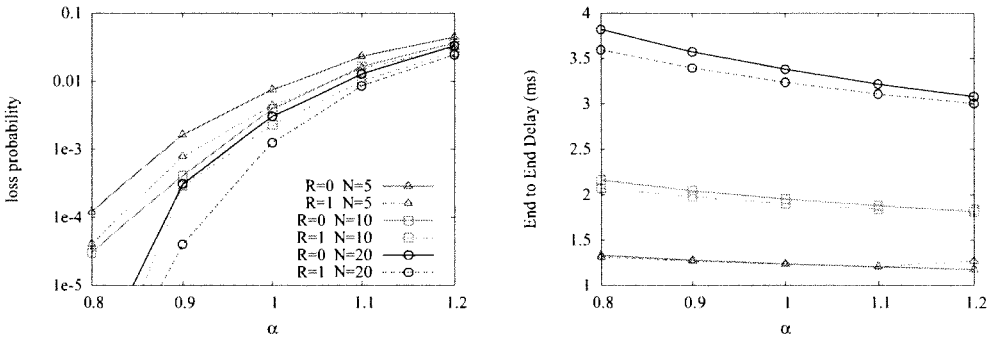
We use the EONET topology and the traffic matrix found in [BGH+04] with $C = 10$ Gbps. The overall load is tuned with a multiplicative coefficient α. Full wavelength conversion is featured in every node using equipment described in [LTL+06a] so that the conversion delay can be neglected. We assume that opto-electrical conversion is performed at the same rate as the transport capacity 0.1 ms.Mb^{-1}. The header processing time is set to 50μs ([SHM+05]). Incoming packets have equal

sizes (1 Mb) and arrive according to a Poisson distribution. The routes are computed with the model proposed in [Cou05] that aims to minimize the bottleneck load. Finally, we set the number of wavelengths per link W to 30.

Figure 9(a) reports the loss probability with $R \in \{0, 1\}$. First note that, as observed in Section 6.2.1, increasing the burst length N reduces the loss probability: With $R = 0$, increasing N from five packets to 10 packets and then to 20 packets successively reduces the loss probability from 16.4×10^{-4} to 4.1×10^{-4} and 3.1×10^{-4} with $\alpha = 0.9$. The benefit is mitigated with higher loads, but remains significant (from 0.023 to 0.017 and 0.013 with $\alpha = 1.1$). The re-aggregation ($R = 1$) further reduces the loss probability. With $N = 10$, the loss probability is reduced from 25% under high load up to 80% under low load. This reduction is largely attributable to the electrical buffering availability in the re-aggregation capable node, where no burst is dropped.

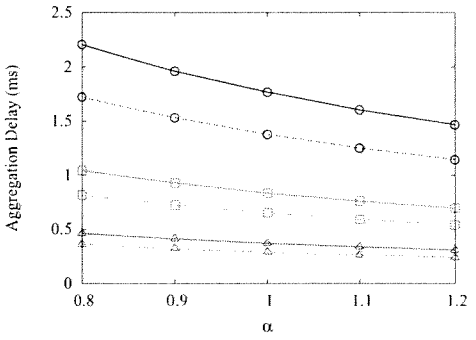
The impact of the re-aggregation on the delay results from the dis-aggregation and the medium access procedures, that are performed more often. Thus, the corresponding delays increase (Figure 9(e) and Figure 9(d), respectively). Note that the contribution of the MAC delay becomes significant under high load ($\alpha \geq 1.2$), but is less impacted with longer bursts, because the reduction of the overlapping degree increases the wavelength availability. With re-aggregation, additional aggregation procedures are requested as well. However, as a result of aggregation by pool, the overall aggregation delay is reduced, though it is performed more often (see Figure 9(c)). The aggregation delay is shown to be the most significant factor on the end-to-end delay and it drives the fluctuations of the end-to-end delay (Figure 9(b)). Therefore, the use of re-aggregation in one node significantly reduces the loss probability, and the effects of electrical processing on the delay are completely compensated by the reduction of the aggregation delay, attributable to traffic grooming.

Traffic grooming enabled with one re-aggregation node compensates the extra delay imposed in this node for electrical processing. Let us now discuss the evolution of this trade-off with more grooming capacity. Figure 7.10 reports the loss probability and the end-to-end delay for an increasing number of R translucent nodes. Increasing the re-aggregation capacity improves the access to

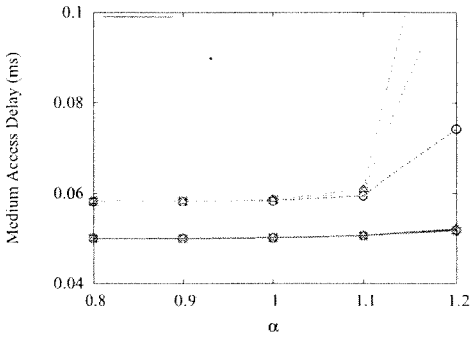


(a) Loss

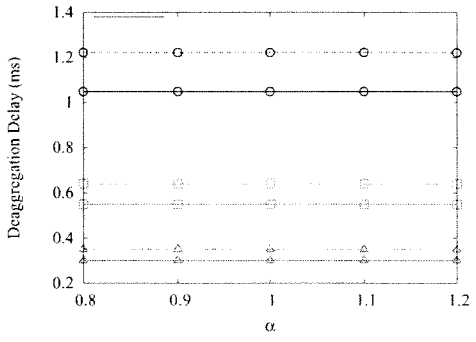
(b) End-to-End Delay



(c) Aggregation Delay

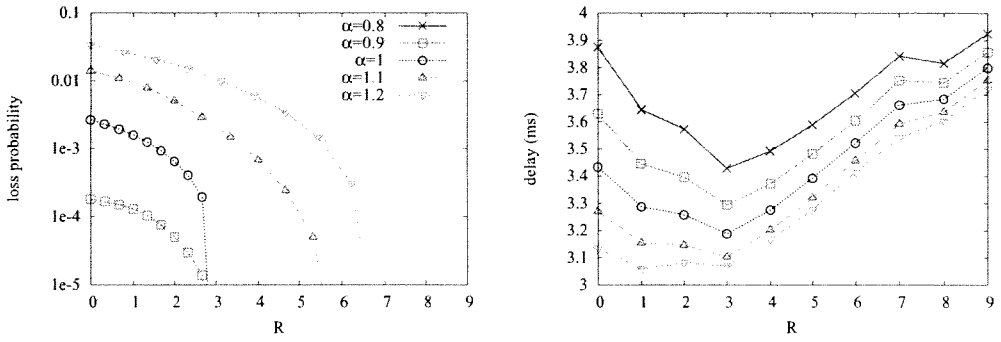


(d) Medium Access



(e) Disaggregation Delay

Figure 7.9: Impact of the AQ Grooming with Various Burst Sizes



(a) Impact on the Loss (N=20)

(b) Impact on the Delay (N=20)

Figure 7.10: Impact of the Number R of Translucent Nodes

electrical buffering and consequently systematically reduces the loss probability.

While increasing R, a larger number of AQs can be merged, but, above a given capacity ($R = 3$ in our experiments), the benefits of traffic grooming are overridden by the effects of electrical processing, because of the multiplication of the re-aggregation procedures.

7.5 Conclusion

In this chapter, we described an OBS architecture that exploits traffic grooming. This architecture operates multiplexing and demultiplexing in the electrical domain, like SONET/SDH systems. However, whereas traffic grooming in OCS networks is mainly helpful to reduce the transport granularity (and consequently increase the bandwidth usage), in OBS networks, it highly improves the aggregation delay. Although the impact of electrical processing on the end-to-end delay is a serious threat, our experiments showed that the reduction of the aggregation delay counter-balances the effect of signal conversion, which, however, significantly reduces the contention probability.

Loss-less transmission can be provided, similarly as in SONET/SDH systems, by multiplexing merging flows from the electrical domain. The re-aggregation can be performed with nearly-basic OBS equipment and the configuration of the grooming operations, although it is based on the global routing table, involves only the source. This may be a true advantage over SONET/SDH systems, which imposes end-nodes synchronization prior to transmission. In addition, the asynchronous

nature of OBS is preserved and ensures a better reactivity to traffic variations.

However, loss-less transmission is based on a global routing configuration of the network. It compromises the dynamism and the reactivity of OBS in cases where re-routing is required. In the next chapter, we present the CAROBS transmission scheme, which exploits traffic grooming dynamically, and without recourse to electrical buffering. With CAROBS, grooming is not performed in electrical domain and the end-to-end delay is significantly reduced. The benefits of CAROBS can be spent in dynamic and on-demand access to electrical buffering to solve contention.

CHAPTER 8

CAROBS: Reliable Burst Transmission with Dynamic Traffic Grooming

In the previous chapter, we proposed a static architecture that enables traffic grooming with intermediate re-aggregation. Though the electrical processing involved in the intermediate re-aggregation affects the end-to-end delay, we showed that this shortcoming can be compensated by traffic grooming. As a result, the architecture is shown efficient to reduce the loss probability – thanks to the electrical processing –, without impacting the end-to-end delay.

In other words, traffic grooming indirectly provides a delay budget that can be spent on re-aggregation. In this chapter, we propose a framework for dynamic and all-optical traffic grooming. We demonstrate the significant benefits of the so-called CAROBS architecture. Then, we propose to use re-aggregation as a reactive solution to contention.

Experiments show that CAROBS can compensate the effects of re-aggregation to solve 100% of the contentions without impacting the end-to-end delay, nor the equipment cost. The priceless asset of this architecture is that it remains completely reactive and can be deployed with standard OBS equipment. Thus, it can be combined with any other pro-active or reactive mechanisms to further improve the performances.

8.1 Transparent Traffic Grooming in OBS Networks

Traffic grooming in optical domain consists in modifying the content of a burst – either inserting or discarding a fragment of the burst – without recourse to electrical processing. Those operations will be referred to as optical MUX/DEMUX operations respectively. Optical DEMUX denotes the segmentation of an optical signal into a number of sub-signals to be switched differently. On the opposite, an optical MUX combines signals from different input ports toward a given output port.

Those operations can be achieved naturally within the OBS framework: As the bursts are signaled in advance, switch reconfigurations can be pre-programmed. Nevertheless, the signal arriving during the configuration update is lost and successive units to be de-multiplexed must be separated by the reconfiguration time.

With SOAs, the reconfiguration time δ^s is in the order of the nanosecond [EBS02] and the resource wastage is absorbed for units as small as few micro-seconds. Thus, the use of optical MUX/DEMUX is reasonable in OBS networks.

In the literature, optical traffic grooming proposals are restricted to the data plane. They aim at arranging the bursts so as to minimize the number of switch reconfiguration. In [SQ04a], core nodes are equipped with FDLs, that are used to reduce the gap between two successive bursts that travel along a common set of links. Therein, traffic grooming operates in the core network.

In [LQ04], the scheme is extended in the edge nodes where several bursts traveling on a common sub-path are sent in a row. As in native OBS, the content of an AQ is assembled in a burst, either once it reaches a given size or once a timer has expired. In the framework found in [LQ04], such events can entail the triggering of additional bursts, all emitted in a blast. The bursts are switched the same way in the first several nodes, but are all signaled independently. From the control plane point of view, those bursts are seen as distinct units, but they request the same switch configuration in the first few switches. The benefits of traffic grooming are actual: On the one hand, the burst serialization reduces the overlapping degree in the core node, on the other hand, the preemption of the aggregation process reduces the aggregation delay.

The CAROBS framework proposed in this chapter intends to accentuate the benefits of traffic grooming. Firstly, traffic grooming is explicitly included in the aggregation process, secondly, the control plane is informed of the grooming decision, which enables optical multiplexing.

8.2 The Emission Process in CAROBS

In CAROBS, the traffic is not mixed within a burst. Traffic grooming rather operates in the control plane: Several bursts, called "car-bursts" in this context are sent in a row, as distinct units, in a virtual unit called "burst-train". The cars are signaled by the same header, associated with the train.

This scheme is an hybrid between the solutions found in [FZJ05] and [LQ04]: As in [LQ04] (and unlike in [FZJ05]), the multiplexed cars (i.e., the content of each AQ) remain separated in the data plane, but, as in [FZJ05] (and unlike in [LQ04]), they are signaled by a single header. The separation in the data plane enables optical demultiplexing and the joint signaling enables optical multiplexing traffic grooming in the core network.

8.2.1 Aggregation in CAROBS

The aggregation process is based on aggregation pools (see Section 7.2.3). Figure 8.1 describes the process in the CAROBS framework. AQs are grouped into APs and each AQ generates a so-called car-burst (denoted by $c[s_c, e_c]$ where s_c and e_c are the starting and ending times of the car), submitted to the train assembler of the AP. The train assembler organizes the cars within a train, which is submitted to the medium access controller, which computes the OT of the train and schedules the departure date. Note that the cars are separated by a guard period equal to the switch reconfiguration time. This gap is mandatory to enable optical demultiplexing, as detailed later in Section 8.3.1.

A key aspect of the aggregation process is the design of the APs. It dictates the degree of traffic grooming, but also the complexity of the de-grooming operations. Increasing the population of an

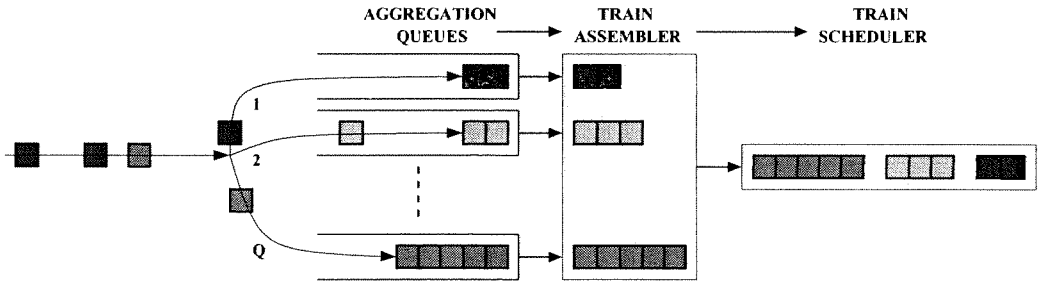


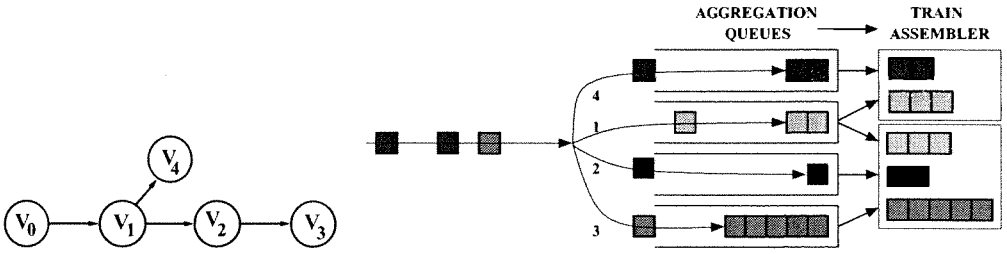
Figure 8.1: Aggregation Pool

AP intensifies traffic grooming, but it may complicate the management of the train in the core network.

Several APs definitions can be envisioned. Assuming that a route is assigned to each AQ, we characterize different types of pools by the topology induced by the set of paths of the involved AQs. For instance, a path-pool contains AQs such that all the destination nodes are along the same path, whereas a tree-pool contains AQs, whose routes diverge but never converge.

Considering the topology depicted on Figure 2(a), Figure 2(b) illustrates the path-pool aggregation of node v_0. The node manages two pools: pool P_3 with destination v_3 that contains AQs with destination v_1,v_2 and v_3 ; and pool P_4 with destination v_4, filled with traffic destined to v_1 and v_4. Note that AQ_1 with destination v_1 belongs to both pools. The train assembly is either initiated by a pool, once it can be supplied enough payload by its AQs, or by an AQ, once the waiting time has exceeded a pre-defined limit. In our example, any pool that initiates train assembly (size-based criterion) receives the content of AQ_1. In the case where AQ_1 initiates the train triggering (time-based criterion), then the train assembly involves the most loaded pool AQ_1 belongs to.

With tree-pools, all AQs belong to the same pool. Such a configuration accentuates traffic grooming, however it entails a splitting of the train at node v_1 and an additional header must be created. With path-pools, DEMUX operations only deal with cars that reached their destinations. Thus, a train is never split. The impact of pool definition is worth further investigation, but in this thesis, we will only consider the path-pools to keep a simple control plane.



(a) Example of topology (b) Aggregation by Pool

Figure 8.2: Pool Definition

8.2.2 Signaling in CAROBS

Optical demultiplexing of a train is enabled by the fast switching devices of OBS. Indeed, from the data plane point of view, each car is seen as an independent burst and can be handled by a custom configuration, independently of the other cars in the train. However, all the cars in a train are signaled by the same header. The train header contains the information related to each car and the node deduces the appropriate configuration to be set for each car.

Figure 8.3 depicts the train header. It is composed of two sections. The "train section" includes the information related to the train, and common for every car: The wavelength that carries the train (WL), the gap between the header and the head of the train (i.e., the offset time – OT), and the destination of the train (Dest.).

The "car section" contains the information specific to each of the "N" cars of the train. For each car, the header stores its destination, the distance to the head (Δ_c), the size, and, optionally, additional fields. Thus, denoting by C the data rate of the wavelength used by a train, the arrival date s_c and the termination date e_c of a car c can be computed as follows:

$$s_c = t^H + OT + \Delta_c/C$$

$$e_c = s_c + size_c/C$$

Note that if the head car is discarded from the train, or if a car is inserted at the head of the train, the OT of the train and Δ_c of each car must be updated. In addition, if a car is inserted at the tail,

its destination must be stored in the train section.

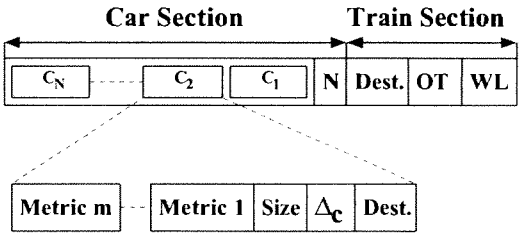


Figure 8.3: CAROBS Header

8.2.3 Offset Time Computation

The offset time OT_v^c of a car c at node v is the gap between the header arrival date, denoted by t_v^{H}, and c arrival date, denoted by t_v^c ($OT_v^c = t_v^c - t_v^{\mathrm{H}}$). It is set at the emission and consumed at each node by the header processing time δ^{H}: With δ^{P} denoting the propagation time of the next link, $t_{v+1}^{\mathrm{H}} = t_v^{\mathrm{H}} + \delta^{\mathrm{H}} + \delta^{\mathrm{P}}$ whereas $t_{v+1}^c = t_v^c + \delta^{\mathrm{P}}$.

In order to ensure that a car c never catches up its header, the source node must assign it an Offset Time OT^c larger than the overall processing time ot^c required up to the destination of c ($ot^c = \ell^c \times \delta^{\mathrm{H}}$, where ℓ^c denotes the number of hops to reach the destination of c).

In CAROBS, the OT is computed for a train, but it must conform to the OT constraint of all its cars. The Offset Time OT^T of train T is equal to the OT of the head car c_0, and, denoting by $\Delta_c = s_c - s_{c_0}$ the gap between the head of the train and the head of car c, we have $t_v^c = t_v^{c_0} + \Delta_c$. As a result, if the OT of train T is OT^T, then each of its car c is assigned, implicitely an offset time OT^c equals to $OT^T + \Delta_c$. Finally, the OT constraints in a train can be expressed as follows:

$$OT^T \geq \max_c (ot^c - \Delta_c) \tag{35}$$

The cars with whom the value $(ot^c - \Delta_c)$ is maximum are called the dominant cars. They are not assigned extra offset time (EOT), i.e., $ot^c = OT^c$.

On the example depicted in Figure 8.4, considering that the time unit is the length of an IP

packet and the header processing time is equal to four time units, $OT^T = \max_{c \in c_1, c_2, c_3} (ot^c - \Delta_c) = \max\{8, 9, 10\}$. The dominant car is c_3 and c_1 and c_2 are assigned an extra offset time (EOT).

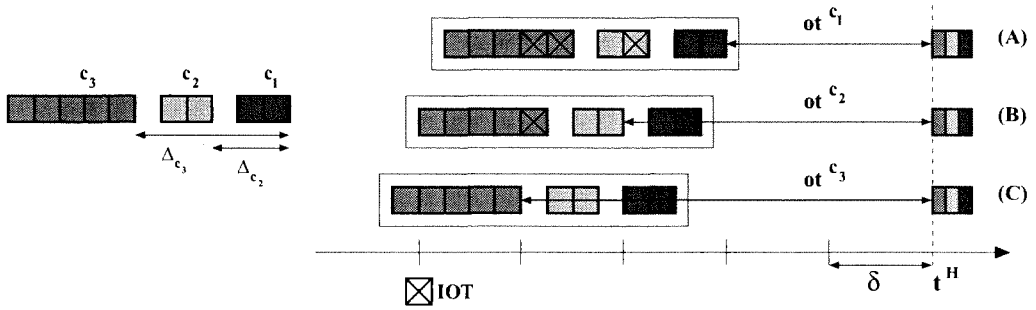


Figure 8.4: CAROBS OT Computation (Sorted Cars)

Figure 8.5 illustrates the same process, but the cars are not sorted the same way. In that case, c_2 imposes the train departure time and increases the value of the OT. Indeed, the minimum OTs are obtained by sorting the cars according to their OT requirements (as on Figure 8.4). This way, cars from short connections are transmitted while cars from longer connections wait for their OT. As a result, the EOT of each car is reduced and so is OT^T.

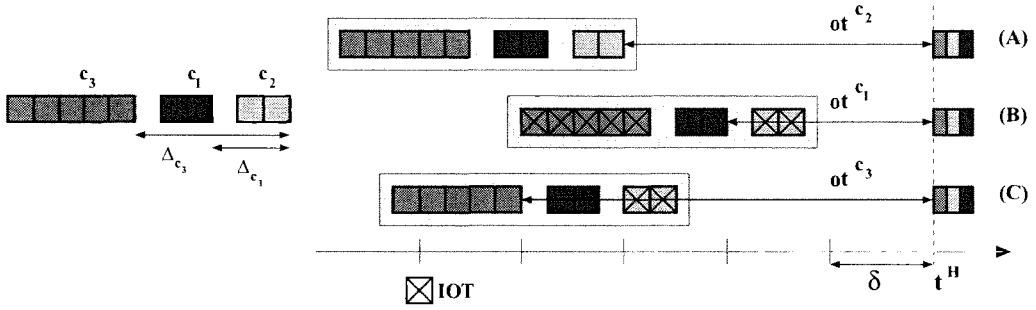


Figure 8.5: CAROBS OT Computation (Unsorted Cars)

8.2.4 Curbed Train Assembly (CTA)

In Section 8.2.3, we described the OT computation policy in CAROBS. We showed that the OT of each car is ruled by its position in the train and that its EOT is minimum if the cars are sorted by increasing OT requirements. We propose here to constrain the size of the cars to reduce further the OT.

On the example of Figure 8.6(a), all the cars are dominant. This configuration is attained if $OT^c = ot^c = OT^T + \Delta_c$, i.e., if the duration of each car amounts the difference of its OT with the OT of the following car (minus δ^S). If the size of a car is reduced, then the preceding cars are assigned a larger OT (Figure 8.6(B)). In opposition, if a car is enlarged, then the OT of the succeeding cars is increased (Figure 8.6(C)).

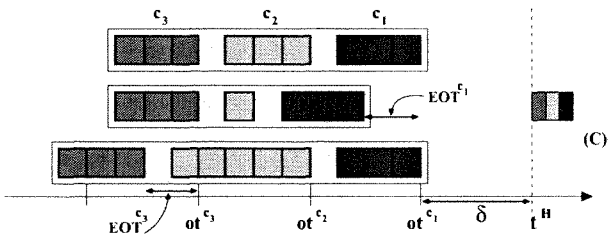


Figure 8.6: Impact of Car Size on the OT

We propose here to assemble the trains so as to avoid EOT. The proposed Curbed Train Algorithm (Algorithm 4) is authorized to chop the cars to limits Δ_c to the difference between the OT requirements of car c and the head car. On Figure 8.7, we assume that δ^H and δ^S respectively amount to the duration of four packets and one packet. The cars are triggered once the pool contains nine packets. We consider a configuration with eight packets in the pool. The next packet arrival provokes the triggering of the cars. If the next packet arrives in AQ_2 (Figure 8.7(B)), then each AQ submits a car of length three and the searched configuration is attained. If the next car is stored in AQ_1 (Figure 8.7(A)), then AQ_2 cannot fill the gap with the following car, imposed by the OT requirement. In that case, the length of c_1, launched by AQ_1, is increased so as to complement c_2. This solution is not possible if the packet is buffered in AQ_3 (Figure 8.7(C)). In that case, the content of AQ_3 is not included in the train.

The last illustration suggests a possible shortcoming of the assembly process since the emission of a car is conditional to the load of the previous AQs. It may entail short trains and seems to disadvantage the traffic of longer connections. This problem can be solved by postponing the triggering, either by triggering according to a larger occupancy, or with finer criteria related to the actual occupancy of each AQ. For example, the train may be triggered only once each AQ can submit

long enough cars.

In the next section, we present the core of a CAROBS network. The in-advance signaling of a sequence of cars enables car insertion in a transit train. The resulting "core grooming" – precisely described in Section 8.3.2 – offers an effective alternative to drain the traffic of the AQs that hardly access the local trains.

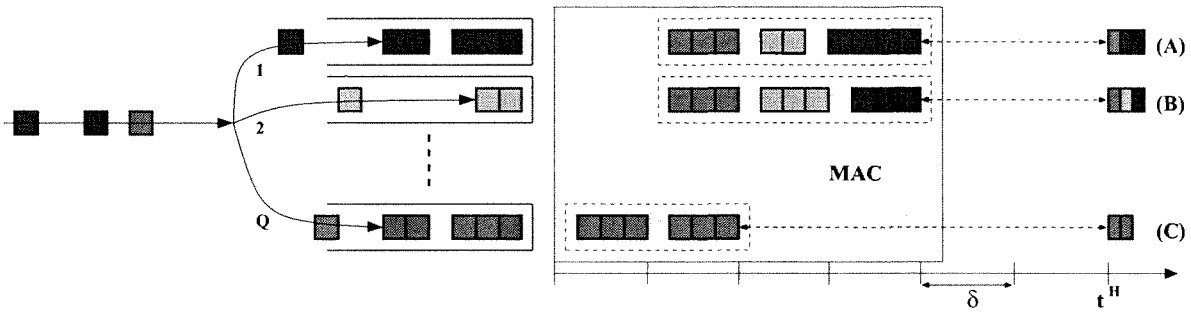


Figure 8.7: Curbed Train Assembly

Algorithm 4 CTA algorithm

AQs are sorted by increasing path length ℓ
for $i \in [0, \#AQ]$ **do**
  $down \leftarrow \ell_i \times \delta^H$
  $up \leftarrow (\ell_{i+1}) \times \delta^H - \delta^S$
  $idle\_period \leftarrow (up - down) \times C$
  if $\{AQ_i.size() \leq idle\_period\}$ **then**
    $s \leftarrow idle\_period$
    $s' \leftarrow 0$
  else
    $s' \leftarrow idle\_period - AQ_i.size()$
    append a car of size s from AQ_{i-1} to the last inserted car
    $s \leftarrow AQ_i.size()$
  end if
  add a car of size s from AQ_i in the train
  $down \leftarrow down + (s + s')/C + \delta^H$
  if $down = up$ **then**
    $up \leftarrow \ell_{i+1} \times \delta^H$
  else
    return the train
  end if
end for

8.3 CAROBS in the Core Network

CAROBS takes advantage of the in-advance signaling of OBS to perform optical multiplexing and demultiplexing. As described in Section 8.2.2, the header signals a train, i.e., a sequence of cars, and the node can handle each car independently. In Section 8.3.1, we describe the switching in CAROBS. Then, the core grooming is described in Section 8.3.2.

8.3.1 Core Switching

The core node behavior is illustrated in Figure 8.8. The train header contains the arrival date, the duration and the destination of each car of the train (1). The information is retrieved after signal conversion and supplied to the SOA manager (SOAm) (2), responsible for configuring the switch. For each car, the SOAm stores a switch configuration and the arrival date of the car. Once the current car is fully transmitted, the SOAm sets the switch for the next car. With path-pools, the configuration either forwards the car toward the next node (the car stays in the train) or toward signal converters if the car burst has reached its destination (it is then disaggregated and exits the optical core network) or if it faces adversity (contention, insufficient signal quality ...).

Thanks to the "in-advance" signaling, the SOAm can preempt the local aggregation process and plan car insertion in the train (3), thus achieving core grooming. The updated information of the train (taking into account car removals and insertions) is transmitted to the header factory (4). The header is updated with the information of the cars composing the train (5) and sent toward the next node (6). At the train arrival (7), the SOAm configures the switch on time to discard and insert cars in the train in transit toward the next node (8).

8.3.2 Core Grooming

In a core node, some cars may be discarded from the train, either if they have reached their destination or if they cannot be transmitted on the output port. On Figure 8.9, the head car of the train is discarded in node v_1. The resulting train exhibits idle periods at the head and in the tail.

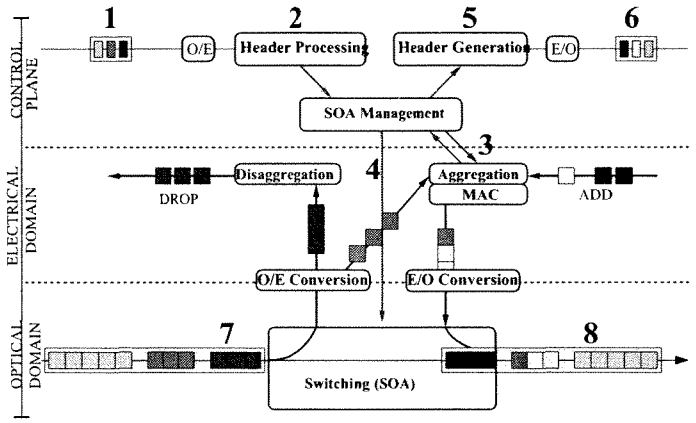


Figure 8.8: CAROBS Switching

Let us assume that we want to keep the cars sorted by increasing OT requirement. The head of the train is always the closest to its destination. Eligible AQs for head insertion are thus only those whose destination is the same as the head car or closer. For each AQ, the longest car that can be inserted is $c[t^{\mathrm{H}} + \ell^{c}, s_{c_0} - \delta^{\mathrm{S}}]$. Thus, if the head car is one hop away from its destination, the inserted car must have the same destination and its duration is equal to the EOT of c_0 minus δ^{S}. If so, the reconfiguration time δ^{S} can be neglected because no reconfiguration is required between the cars.

Denote by c_n the last car of the train. A car c_m can be appended to the tail of the train if $\ell^{c_m} \geq \ell^{c_n}$ and if $EOT_m = e_{c_n} - t^{\mathrm{H}} - \ell^{c_m} \times \delta^{\mathrm{H}} \geq 0$.

Once the eligible AQs for head or tail insertion are identified, the corresponding cars are created and sent to the MAC buffer. The car conversion is then scheduled in order to be synchronized with the train transit. Note that two successive cars must be separated by the switch reconfiguration time δ^{S} (few nanoseconds with SOA switches).

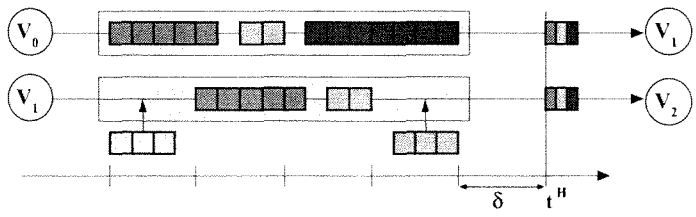


Figure 8.9: Core Grooming

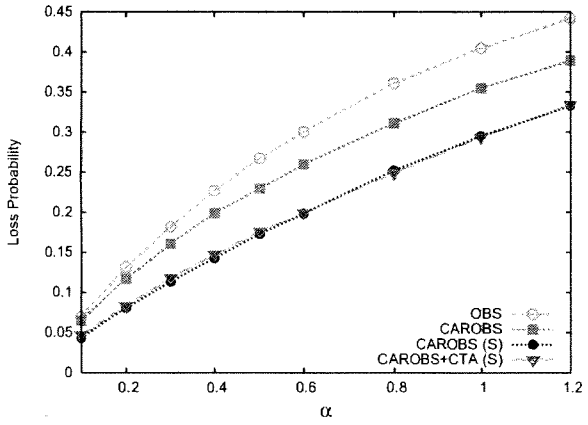


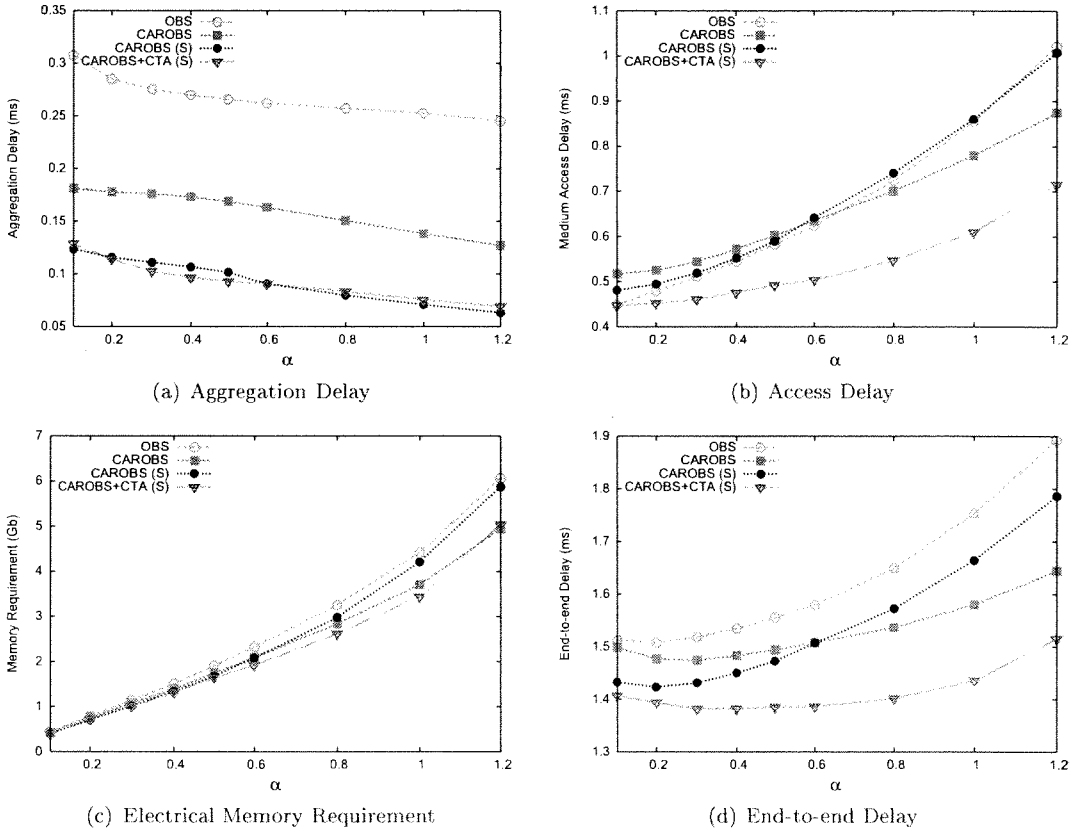
Figure 8.10: Impact of CAROBS on the Loss Probability

8.4 Performance Evaluation of CAROBS

We performed experiments on the EONET network. Fibers offer 50 wavelengths at 10 Gbps. The original traffic matrix found in [BGH+04] is adjusted by a factor α. Packets of constant size (100 kb) are supplied to AQs according to a Poisson distribution. The aggregation timer is set to $0.5ms$ and the maximum train size is $10Mb$.

Our reference scenario uses the native OBS-JET (referred to as "OBS" in figures). We evaluate the performances of CAROBS without segmentation (i.e., a contention entails the drop of the whole train, referred to as "CAROBS") and with segmentation (only the contending cars are dropped, referred to as "CAROBS(S)"). The slight reduction of the loss probability earned by CAROBS without segmentation (15%, see Figure 8.10) confirms that traffic grooming smoothes the traffic. With segmentation, the reduction reaches up to 40%. This improvement is mainly attributable to the segmentation of the train, but also to the intensification of the core grooming: A discarded car opens an idle slot in the train and offers more opportunities of car insertion to further nodes.

The core grooming intensification due to segmentation is validated by Figure 11(a) and Figure 11(b). Figure 11(a) shows that the aggregation delay obtained with OBS is reduced by 40% by using CAROBS without segmentation. Further reduction of an additional 20% is observed when segmentation is performed. Indeed, the core grooming pre-empts the aggregation process. Figure 11(b) shows that the medium access delay (i.e., the time between train triggering and train launch)



(a) Aggregation Delay

(b) Access Delay

(c) Electrical Memory Requirement

(d) End-to-end Delay

Figure 8.11: Impact of CAROBS on Delay Related Metrics

increases with CAROBS (up to 10%), mainly because of the EOT assigned to some cars. As the core grooming does not impact the offset time of the train, the effect of extra offset-time is less obvious with segmentation (intermediate nodes have more opportunity to insert cars in trains in transit). When the load increases, the use of segmentation entails a larger delay, however always shorter than with classical OBS. Indeed, as the loss probability decreases, the resource utilization increases and consequently, the access to the medium is more congested. With CTA, the trains are assembled in a way to avoid EOT, as reflected by the reduction of the medium access delay.

Combining the reduction of the aggregation and the medium access delay leads to a significant reduction of the end-to-end delay when CAROBS is used with CTA (Figure 11(d)). In addition, as the payload occupies memory during the aggregation process and the wait for medium access, the CTA reduces the memory occupancy. The sum of the maximum memory usage of each node measured during the simulations is reported on Figure 11(c). CAROBS reduces the memory requirement as

compared with the classical OBS. This reduction is mitigated by the use of segmentation because in that case, the loss probability is reduced and consequently increases the effective load in the network (especially under heavy load). The impact of CTA is clearly reflected in that situation: Although the loss probability is not impacted by CTA, the memory requirement decreases by approximately 15%. This is the direct consequence of the reduction of the emission delay.

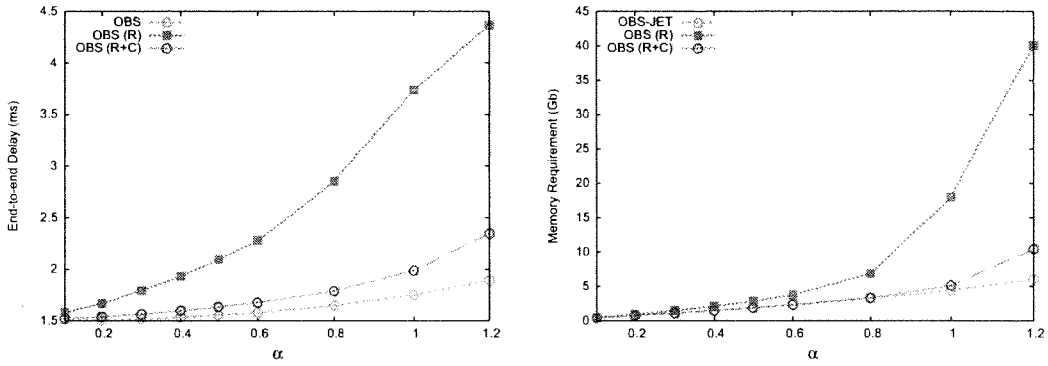
Those observations demonstrate the potential of CAROBS: Without any specific equipment, CAROBS reduces the loss probability, the end-to-end delay and the memory requirement. As CTA reduces the OT, it further improves the end-to-end delay and the memory requirements, without impacting the loss probability.

8.5 CAROBS in Translucent Mode: A Reliable and Dynamic Solution

8.5.1 Limitations of Translucent Architectures

Electrical buffering is often conceived as a set of queues where the converted bursts are stored, waiting for their re-emission. Such an architecture raises two main shortcomings. Firstly, it imposes the installation of dedicated memory. Secondly the stored bursts compete with ingress bursts triggered by the aggregation module and with transit bursts in the optical domain.

The memory occupancy is a crucial aspect since it determines the contention resolution capacity of the system. Indeed, a contending burst will be dropped if the buffers are full (buffer overflow). In an OBS network, electrical buffers are required for two purposes. Firstly, edge nodes manage the aggregation process and must be equipped with enough memory. Then, once a burst is triggered, its departure date is computed and the burst is stored by the MAC layer. The time spent in the MAC layer amounts at least the offset time of the burst, but it also depends on the transport resource availability. Thus, the memory requirement is directly connected to the time spent in the edge nodes, i.e., to the aggregation and medium access delay. Thus, reducing the memory requirements



(a) End-to-End Delay (b) Memory Requirements

Figure 8.12: Impact of Re-aggregation

consists in reducing the emission delay and the drawbacks of signal conversion can be faced jointly.

The architecture depicted on Figure 7.7 is thus well suited to provide electrical buffering. Instead of accessing dedicated buffers, the contending payload is submitted to the local aggregation process. Combining the electrical buffers assigned to contention resolution and those used for aggregation avoids the installation of dedicated memory. In addition, the converted payload contributes to the local aggregation process and consequently reduces its duration, with a beneficial impact on the end-to-end delay and on the memory requirement.

However, the memory requirements are directly impacted by the amount of converted payload, i.e., by the contention rate. Figure 12(a) and Figure 12(b) respectively report the end-to-end delay and the memory requirement of an all-optical OBS network in the scenario used in Section 8.4. The flag "R" denotes the use of re-aggregation to solve contention and "C" the availability of full wavelength converters at each node. When the load increases, the access to the medium is more congested and the bursts spend more time in the MAC buffers. Consequently, the memory requirement increases. With re-aggregation ("OBS(R)"), the amount of data intended to access the medium increases and consequently, the delay and the memory requirements highly increase. The cost of re-aggregation remains reasonable for low load, i.e., when the contention rate is fairly small. Nevertheless, increasing the load drastically increases that cost to an unacceptable level. Thus, though it allows contention resolutions, straightforward and systematic recourse to electrical

buffers is of disputable relevance. The joint use of alternate contention resolution mechanisms prior to re-aggregation reduces the amount of data involved in the re-aggregation process and decreases the cost of re-aggregation. The results reported in Figure 8.12 show that wavelength converters ("OBS(R+C)") can thwart the drawbacks of re-aggregation for $\alpha \leq 1$.

Featuring wavelength converters at each node drastically increases the cost of the network. In [SR09], a limited wavelength conversion capacity is installed at each node to mitigate the impact on the network cost. The performance of this solution is ruled by the tradeoff between the equipment cost and the recourse to electrical buffering.

8.5.2 CAROBS in Translucent Mode

The numerical results reported in previous section demonstrate the correlation between the emission delay and the memory occupancy. They also demonstrate that the reduction of the contention rate helps reducing the emission delay and consequently improves the memory availability.

The CAROBS transmission scheme appears as an adapted complement to the re-aggregation since it reduces the time spent in AQs and smoothes the traffic, resulting in a lower contention probability. In addition, the core node treats each car sequentially and the dropping process is managed at the level of the cars, naturally leading to a transparent segmentation process, which reduces the volume of data involved in the re-aggregation process.

Indeed, operating in translucent mode is very similar as operating in native transparent mode. The only difference lies in the treatment of the contending cars. Whereas they are dropped in the transparent mode, we propose here to convert them to electrical domain and to submit them to the local aggregation process, just as the ingress traffic.

With this architecture, no memory is dedicated to contention resolution. The contention resolution rather uses the basic facilities. However, the contending payload increases the payload emitted at each link and the memory requirement in the MAC will increase. On the other hand, traffic

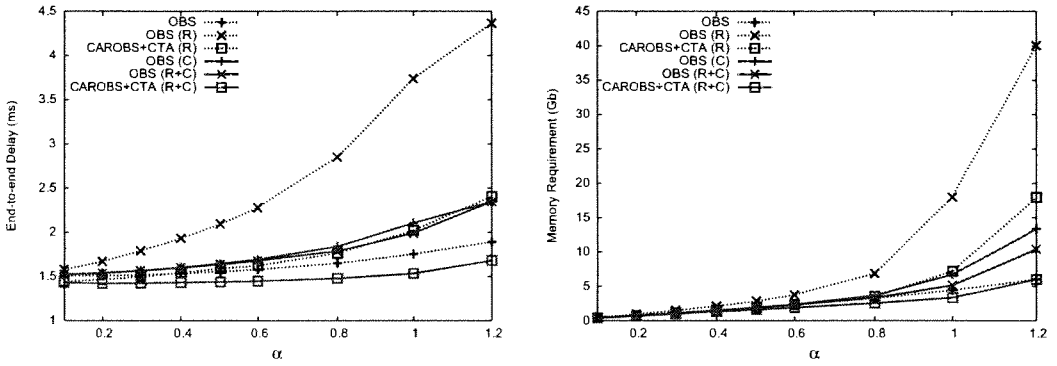
grooming reduces the time spent in the AQs, and consequently the memory occupancy. In addition, car insertions in transit trains preempt the aggregation process and further balance the cost of contention resolution.

8.5.3 Performances of CAROBS in Translucent Mode

Our experiments complement those reported in Section 8.4. Therein, CAROBS operates in transparent mode and results in a high dropping probability, though it is reduced as compared with a basic OBS network. In Section 8.5.1, we showed that re-aggregation has a dramatic impact on the memory requirements and the delay, and that wavelength converters can considerably mitigate those drawbacks. Our concern here is to position CAROBS as an alternate complement to re-aggregation to replace the use of expensive wavelength converters. In this section, CAROBS is implemented with the CTA algorithm. On the figures, the flag "R" denotes the use of re-aggregation to solve the contentions and the flag "C" denotes the use of wavelength conversion.

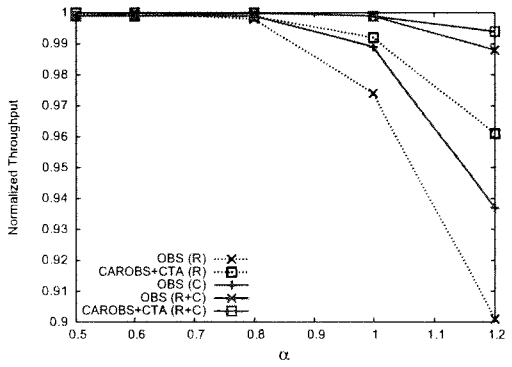
We assume that the memory budget is unlimited and measure the memory requirements (Figure 13(b)) and the end-to-end delay (Figure 13(a)) entailed by 100% contention resolution. Results obtained in Section 8.5.1 – without traffic grooming – are reported on those figures and are used as reference scenario.

CAROBS reduces the cost of re-aggregation (i.e., its impact on the end-to-end delay and on the memory requirement) by 50%. With the CTA policy, the OT of the cars is reduced and consequently the time spent in electrical buffers and the memory requirements decrease. As a result, CAROBS with the CTA train assembly completely compensates the cost of re-aggregation for $\alpha \leq 0.8$, though the contention rate can reach up to 36%. In other words, the loss-less transfer is achieved with the same memory requirements and end-to-end delay. For higher load, CAROBS cannot fully counterbalance the effects of re-aggregation and the cost increases. When OBS is equipped with wavelength converters, the loss probability is highly reduced and the memory requirement and end-to-end delay are increased. This is because the bursts that reach the destination spend more time in the network



(a) End-to-end Delay with CAROBS and Re-aggregation

(b) Memory Requirements with CAROBS and Re-aggregation



(c) Normalized Throughput

Figure 8.13: Impact of CAROBS on Re-aggregation Cost

as compared with those that are dropped. The memory requirement and end-to-end delay are indeed similar as those measured with CAROBS in translucent mode, though the latter configuration is loss-less.

Indeed, CAROBS can be considered as a cheap alternative to wavelength conversion to face the effects of translucence. In translucent mode, with $\alpha \leq 0.8$, CAROBS entails the same re-aggregation costs as wavelength converters, though it does not impact the CAPEX. Anyway, if wavelength converters were installed, then CAROBS remains helpful and further reduces the cost of re-aggregation, at a lower level than any native OBS configuration.

8.6 Conclusions

CAROBS enhances the solution found in [FZJ05] because traffic grooming is operated in the optical domain and, thanks to aggregation by pools, it is also efficient in highly loaded scenarios.

Aggregation by pool is also an improvement as compared with the transparent traffic grooming proposed in [LQ04]. Therein, a burst can preempt the aggregation process so as to follow the previously sent burst, but the aggregation premption is less systematic and rather opportunistic. Moreover, CAROBS performs grooming of the control plane as well, and enables the insertion of cars in transit trains to accentuate the benefits of traffic grooming.

CAROBS thus highly exploits traffic grooming to reduce the end-to-end delay and memory requirements. We have proposed to spend those benefits to counter-balance on-demand re-aggregation procedures, enabled for contending payload. The re-aggregation enables electrical buffering and its capacity to solve the contentions depends on the capacity and availability of the buffers. As CAROBS strongly reduces the time spent in aggregation queues and improves the availability of the electrical buffers, it counter-balances the cost of re-aggregation so that 100% of the contentions are solved without impacting the delay nor the memory requirement, even in quite congested states.

The proposed architecture is expected to fulfill the objectives of this thesis since it reactively solves the problem of contention in a genuine OBS framework, thus preserving the assets of OBS regarding the reactivity.

CHAPTER 9

Conclusions and Possible Extensions

9.1 Summary of the Thesis

The deployment of Passive Optical Networks between the users and the Metropolitan Area Networks (MANs) aims to provide the transport capacity required by the next generation traffic, and is expected to drastically modify the traffic profile in the MANs. Progressively, the common services will quickly migrate to the Internet, and new applications will emerge. The high variability of the traffic, resulting from the variety of the applications, will probably be preserved by the passive equipment deployed in the access. Thus, following the extension of the optical network, MANs will receive a heavy load of highly variable traffic.

This evolution threatens the integrity of the MANs because of the static provisioning of the current circuit oriented network (OCS). Indeed, OCS is restricted to static optical configuration due to the low speed of its switches. The latency to change the configuration addresses several issues. In the scope of this thesis, our concern is related to the efficiency of TDM, especially in dynamic contexts. Multiplexing several flows within a wavelength is referred to as traffic grooming. Currently, two architectures provide traffic grooming. The first one, SONET/SDH, is widely deployed and relies on electrical processing to synchronize the emission of several flows without contention. The second one offers an asynchronous access to a wavelength, provided that an adapted signaling protocol is

available.

When the optical fiber has been extended to the MAN, the lack of flexibility of the network, has been addressed and led to a fairly intensive study of packet oriented transmission over optical fibers (OPS), with the objective to design a dynamic and fully reactive optical layer. The lack of buffering in the core network - required during the header processing in charge of the routing decisions - has finally been accommodated a decade ago, with the proposal of Optical Burst Switching. Similarly as OCS, OBS sidesteps core buffering by setting the optical equipment prior to the data arrival. However, whereas in OCS networks, the latency induced by the circuit establishment jeopardizes the reactivity, OBS has been designed to minimize the latency in order to increase the reactivity. Consequently, the resource reservation is not acknowledged, what raises the contention problem, which is a critical obstacle to the maturity of OBS. Thus, the contention problem has been much written about, but remains a wide open issue, particularly challenging while focussing on the reactivity of the optical layer.

OBS is highly flexible. Due to its very fast switches, it has no lower bound on the size of the bursts, whose, however, can be few microseconds as well days, weeks or even years long. Also, the one-way signaling protocol intends to improve the reactivity of OBS, but the signaling can easily be modified to secure the transmission, while mitigating the reactivity. Ultimately, OBS can operate in circuit mode, with an enhanced multiplexing potential and reactivity resulting from the fast switching devices. For instance, in OCS, two flows can be merged via electrical processing. Similarly, contention can be avoided via electrical buffering. Recourse to electrical buffering however increases the end-to-end delay and we proposed to perform intermediate re-aggregation in order to achieve traffic grooming at the level of the burst. The traffic grooming reduces the aggregation delay and can compensate the effect of translucence on the end-to-end delay, while taking advantage of electrical buffering to prevent contentions. Nevertheless, this architecture relies on a pre-planned access to electrical buffers and consequently does not fulfill our objective.

A priceless property of OBS is its ability to read the future. The offset time is quite exclusive to

OBS and it has been quickly exploited, at first, for service differentiation. Thanks to the offset time, the nodes accumulate many information that can be helpful to drive some decisions. Indeed, instead of consulting an historical of metrics, the node can refer to a window pointed to the future. This opens the possibility to plan operation prior to a burst arrival. This is the essence of the CAROBS transmission scheme. Several bursts, called cars in CAROBS, are sent together within a so-called train, The cars, although they do not have the same destination, are signaled by a single header that supplies the required information to the intermediate nodes so that they can demultiplex the cars of the train. In addition, a node can synchronize the emission of a car so that it is inserted within a train expected to transit in the near future. Thus, CAROBS all-optically and dynamically exploits traffic grooming. The process significantly reduces the delay and memory requirement so as to compensate for the expense of contention resolution via electrical buffering.

9.2 Further Research: Impact of CAROBS on the End-to-end Performances

As observed on a given hierarchical level, the reactivity of CAROBS is an attractive property. Nevertheless, it is worthwhile to evaluate how the deployment of CAROBS in the MAN would impact the end-to-end performances.

9.2.1 Integration in the Protocol Stack

In OBS networks, contentions can occur even under low load and may entail data loss in the cases where the available contention resolution mechanisms fail. In that case, the congestion controller of the Transport Control Protocol (TCP) assumes the network is congested and reduces the load of the sources concerned. This assumption is not necessarily relevant in OBS networks. In Section 4.3.2, we presented few mechanisms devised to improve the performances of throughput of TCP over OBS.

With CAROBS in translucent mode, OBS is endowed with a reliable reactive response to the

contentions and the losses are caused by buffer overflows. This is much closer to the TCP assumption, but the system is not equivalent to IP networks. It would be meaningful to investigate the "TCP-friendship" of CAROBS.

9.2.2 CAROBS in the WAN?

Nevertheless, it is fairly reasonable to assume that, following MAN upgrades, the bottleneck will keep on climbing the hierarchy. Currently, there is no obvious argument to urge the deployment of a dynamic architecture in the WANs. Therein, the traffic stability is protected by the large number of connections multiplexed within each flow, despite the reactivity of CAROBS may preserve the traffic variability up to the WAN. The impact of transparency in the access motivated this thesis, but in long-haul networks, the variability of the traffic is absorbed.

Nevertheless, though the traffic may remain stable, the reactivity of CAROBS may be helpful to deal with failure and signal impairment. Let us discuss those two issues.

CAROBS for a Reactive Fault Management

In OCS networks, extra resources must be allocated to the services that do not tolerate service perturbation. This waste of bandwidth can be saved with restoration mechanisms, but in that case, the service is perturbed, or even interrupted, during the connection re-routing, provided the alternative resources are indeed available.

The restoration mechanisms are much more adapted to OBS networks, since the traffic is served burst by burst. The restoration mechanism is basically the same as in OCS networks, except that the traffic emitted before the re-configuration completion, lost in OCS networks, can be deflected to turn around the cut link. Moreover, the use of an alternate route does not impose any additional latency as in OCS networks. The traffic perturbation thus only relies to the ability to handle the deflected traffic.

With CAROBS, the trains are explicitly requested to transit via a given set of nodes. This restricts the alternative routes and may mitigate the restoration process. An efficient alternative

may be to take advantage of the link failure to enhance the traffic grooming by submitting the trains to re-aggregation instead of deflection routing. This way, all the cars affected by the failures are assigned to new trains emitted on an alternate path.

To maintain the network throughput after a failure, spare capacity might be required on each link. This solution can be evaluated by comparing the extra capacity required by OBS and OCS to maintain a given throughput in case of failure.

Dynamic Signal Management with CAROBS

All along its transmission, the signal accumulates impairments, which can compromise the transmission in the case the receiver is unable to retrieve the information from the signal. This is an important concern in long-haul networks. In OCS networks, MEMS-based equipment have a very limited impact on the signal, but this issue however remains of interest and is usually handled by the use of evenly spaced signal regenerators along the path.

In OBS networks, this appears critical because the signal is split several times at each switch, each split resulting in a strength reduction by half. In Section 8.2.2, we described the train header and illustrate the possibility to include optional information for each car. In particular, an estimation of the Q-factor of the car would enable signal management: If the signal is estimated too weak to reach the next node, then the car is dropped from the train, converted to the electrical domain and submitted to the local aggregation process, just as a contending car. After the aggregation process, the content of the car will be re-emitted and carried by a brand new signal.

Several approximations of the Q-factor can be found in the literature, but a clever alternative may be to measure the Q-factor of the converted cars and map the values with a particular historic. This way, the Q-factor of a car may be estimated by consulting the data-base with a look-up on the traveled path and the used wavelength.

Bibliography

[Agg94] A. Aggarwal. Efficient routing and scheduling algorithms for optical networks. *fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 412–423, 1994.

[Bat02] D. Battu. *Télécommunications: principes, infrastructures et services.* InterEditions, 2002.

[BGH+04] A. Betker, C. Gerlach, R. Hulsermann, M. Jager, M. Barry, S. Bodamer, J. Spath, C. Gauger, and M. Kohn. Reference transport network scenarios. Technical report, MultiTeraNet Project, 2004.

[BH94] R.A. Barry and P.A. Humblet. On the number of wavelengths and switches in all-optical networks. *IEEE Transactions on Communications*, 42(234):583–591, 1994.

[BH07] A. Belbekkouche and A. Hafid. An Adaptive Reinforcement Learning-based Approach to Reduce Blocking Probability in Bufferless OBS Networks. *Proceedings of the IEEE ICC*, pages 2377–2382, 2007.

[BM00a] D. Banerjee and B. Mukherjee. Wavelength-routed optical networks: linear formulation, resource budgeting tradeoffs, and a reconfiguration study. *IEEE/ACM Transactions on Networking*, 8(5):598–607, 2000.

[BM09] A. Barradas and M. Medeiros. Pre-planned optical burst switched routing strategies considering the streamline effect. *Photonic Network Communications*, September 2009.

[Bou07] N. Bouabdallah. Sub-Wavelength Solutions for Next-Generation Optical Networks [Topics in Optical Communications]. *IEEE Communications Magazine*, 45(8):36–43, 2007.

[BS05] N. Barakat and EH Sargent. Dual-header optical burst switching: a new architecture for WDM burst-switched networks. *Proceedings of the IEEE INFOCOM*, 1:685–693, 2005.

[CCL+02] J. Cao, W. S. Cleveland, D. Lin, , and D. X. Sun. The effect of statistical multiplexing on the long-range dependence of internet packet traffic, bell labs, tech. rep., 2002 (online), available at: http://cm.bell-labs.com/cm/ms/departments/sia/doc/multiplex.pdf. 2002.

[CEJ05a] T. Coutelen, H. Elbiaze, and B. Jaumard. An Efficient Adaptive Offset Mechanism to Reduce Burst Losses in OBS Networks. In *Proceedings of the IEEE GLOBECOM*, 2005.

[CEJM05] T. Coutelen, H. Elbiaze, B. Jaumard, and A. Metnani. Measurment-Based Alternative Routing Strategies in Optical Burst-Switched Networks. In *Proceedings of the IEEE ICTON*, 2005.

[CHJ09a] T. Coutelen, G. Hébuterne, and B. Jaumard. An OBS RWA Formulation for Asynchronous Loss-less Transfer in OBS Networks. In *Proceedings of HPSR*, 2009.

[CHJ09b] T. Coutelen, G. Hébuterne, and B. Jaumard. Core OBS Traffic Properties and Behavior. *Les cahiers du Gerad*, Mai 2009.

[CHJ09c] T. Coutelen, G. Hébuterne, and B. Jaumard. Is It Worth to Keep an All Optical OBS Data Plane? In *Proceedings of CCECE*, 2009.

[CJH09a] T. Coutelen, B. Jaumard, and G. Hébuterne. A Translucent OBS Node Architecture to Improve Traffic Emission and Loss Probability. In *Proceedings of the IEEE WOBS*, 2009.

[CJH09b] T. Coutelen, B. Jaumard, and G. Hébuterne. Improving RWA-OBS Formulation and Solution. In *Proceedings of the IEEE WOBS*, 2009.

[CJH10a] T. Coutelen, B. Jaumard, and G. Hébuterne. A Viable Translucent Architecture for Lossless OBS Networks. In *to appear in Proceedings of the IEEE ICC*, 2010.

[CJH10b] T. Coutelen, B. Jaumard, and G. Hébuterne. An Enhanced Train Assembly Policy for Loss-less OBS with CAROBS. In *To appear in Proceedings of the IEEE CNSR*, 2010.

[CLCQ02] X. Cao, J. Li, Y. Chen, and C. Qiao. TCP/IP packets assembly over optical burst switching network. *Proceedings of the IEEE GLOBECOM*, 3:2808–2812, 2002.

[Cou05] T. Coutelen. Accès et routage optique en mode de commutation de rafales. Master's thesis, Université de Montréal, 2005.

[CZB04] B. CHEN, W. ZHONG, and S. K. BOSE. A path inflation control strategy for dynamic traffic grooming in IP/MPLS over WDM network. *IEEE Communications Letters*, 8(11):680–682, 2004.

[DH04] Z. Dutton and L.V. Hau. Storing and processing optical information with ultraslow light in Bose-Einstein condensates. *A Physical Review*, 70(5):53831.1–53831.19, 2004.

[DHGY03] J. Doucette, D. He, W.D. Grover, and O. Yang. Algorithmic approaches for efficient enumeration of candidate p-cycles and capacitated p-cycle network design. In *Proc. DRCN*, volume 3, pages 212–220, 2003.

[DHL+03] HJS Dorren, MT Hill, Y. Liu, N. Calabretta, A. Srivatsa, FM Huijskens, H. de Waardt, and GD Khoe. Optical packet switching and buffering by using all-optical signal processing methods. *IEEE Journal of Lightwave Technology*, 21(1):2–12, 2003.

[EBS02] TS. El-Bawab and J.D. Shin. Optical Packet Switching in Core Network : Between Vision and Reality. *IEEE Communications Magazine*, 40(9):60–65, 2002.

[EGH+09] J. Erman, A. Gerber, M.T. Hajiaghayi, D. Pei, and O. Spatscheck. Network-aware forward caching. In *Proceedings of the 18th international conference on World wide web*, pages 291–300. ACM New York, NY, USA, 2009.

[FH04] G. Fiche and G. Hebuterne. *Communicating Systems & Networks: Traffic & Performance (Innovative Technology Series. Information Systems and Networks)*. KOGAN PAGE, 2004.

[FHS04] J. Fang, W. He, and A.K. Somani. Optimal light trail design in WDM optical networks. In *International Conference on Communications (ICC2004)*, pages 1699–1703, 2004.

[FVS04] M.T. Frederick, N.A. VanderHorn, and A.K. Somani. Light trails: a sub-wavelength solution for optical networking. In *High Performance Switching and Routing (HPSR) Workshop*, pages 175–179, 2004.

[FZJ05] F. Farahmand, Q. Zhang, and J.P. Jue. Dynamic traffic grooming in optical burst-switched networks. *Journal of Lightwave Technology*, 23(10):3167, 2005.

[Gau03] C. M. Gauger. Dimensioning of FDL Buffers fr Optical Burst Switching Nodes. *Next Generation Optical Network Design and Modelling*, pages 117–132, 2003.

[GC03a] A. Gumaste and I. Chlamtac. Light-trails: a novel conceptual framework for conducting optical communications. In *Proceedings of HPSR*, 2003.

[GC03b] A. Gumaste and I. Chlamtac. Mesh implementation of light-trails: A solution to IP centric communication. In *12th Proceedings of IEEE International Conference on Computers, Communication and Networks (ICCCN), Dallas, TX*, 2003.

[GK09] G. Gurel and E. Karasan. Using multiple per egress burstifiers for enhanced TCP performance in OBS networks. *Photonic Network Communications*, pages 105–117, 2009.

[GKC03] A. Gumaste, G. Kuper, and I. Chlamtac. Optimizing light-trail assignment to WDM networks for dynamic IP centric traffic. In *12th IEEE Workshop on Local and Metropolitan Area Networks*, pages 25–28, 2003.

[GKS04] CM Gauger, M. Kohn, and J. Scharf. Comparison of contention resolution strategies in OBS network scenarios. In *Proceedings of the IEEE ICTON*, volume 1, pages 18–21, 2004.

[GLW+05] H. Guo, Z. Lan, J. Wu, Z. Gao, X. Li, J. Lin, and Y. Ji. A testbed for optical burst switching network. In *Optical Fiber Communication Conference, 2005. Technical Digest. OFC/NFOEC*, volume 5, 2005.

[GM03] A. Gençata and B. Mukherjee. Virtual-topology adaptation for WDM mesh networks under dynamic traffic. *IEEE/ACM Transactions on Networking*, 11(2):236–247, 2003.

[GS98] W.D. Grover and D. Stamatelakis. Cycle-oriented distributed preconfiguration: ring-like speed with mesh-like capacity for self-planning network restoration. In *IEEE International Conference on Communications*, volume 1, pages 537–543. Citeseer, 1998.

[GS08] C. Grimm and G. Schlüchtermann. *IP-Traffic Theory and Performance*. Springer, 2008.

[GWM02] K. Grobe, M. Wiegand, and J. McCall. Optical metropolitan DWDM networks an overview. *BT Technology Journal*, 20(4):27–44, 2002.

[HD07] S. Huang and R. Dutta. Dynamic Traffic Grooming: The Changing Role of Traffic Grooming. *Communications Surveys & Tutorials, IEEE*, 9(1):32–50, 2007.

[HHK06] D.W. Hong, C.S. Hong, and W.S. Kim. A Segment-based Protection Scheme for MPLS Network Survivability. In *Network Operations and Management Symposium, 2006. NOMS 2006. 10th IEEE/IFIP*, pages 1–4, 2006.

[HHM05] Y. Huang, JP Heritage, and B. Mukherjee. Dynamic routing with preplanned congestion avoidance for survivable optical burst-switched (OBS) networks. In *Proceedings of the IEEE OFC*, volume 3, pages 1708–1712, 2005.

[HJJ05] A. Houle, A. Jarray, and B. Jaumard. Minimum Cost Dimensioning of Ring Optical Networks. *Les Cahiers du GERAD*, Apr. 2005.

[HJS05] A.C. Houle, B. Jaumard, and Y. Solari. Addressing the GRWA problem in WDM networks with a tabu search algorithm. In *Canadian Conference on Electrical and Computer Engineering*, pages 1630–1633, May 2005.

[HL04] JQ Hu and B. Leida. Traffic grooming, routing, and wavelength assignment in optical WDM mesh networks. In *Proceedings of the IEEE INFOCOM*, volume 1, pages 495–501, 2004.

[HL05] Q.-D. Ho and M.-S. Lee. Practical Dynamic Traffic Grooming in Large WDM Mesh Networks. *Proc. 2nd Int'l. IEEE Create-Net Wksp. Traffic Grooming*, 2:1194–1196, 2005.

[HWMA03] T. Hashigushi, X. Wang, H. Morikawa, and T. Aoyama. Burst Assembly Mechanism with Delay Reduction for OBS Networks. In *Proc. COIN/ACOFT*, pages 664–666, 2003.

[INT10] http://www.intunenetworks.com [last visited: april 10, 2010].

[JBCB07] A. Jaekel, A. Bari, Y. Chen, and S. Bandyopadhyay. New Techniques for Efficient Traffic Grooming in WDM Mesh Networks. *Proceedings of the IEEE ICCCN*, pages 303–308, 2007.

[JJ05] A. Jarray and B. Jaumard. Exact ILP solution for the grooming problem in WDM ring networks. In *Proceedings of the IEEE ICC*, volume 3, 2005.

[JMT06] B. Jaumard, C. Meyer, and B. Thiongane. ILP formulations for the RWA problem for symmetrical systems. In P. Pardalos and M. Resende, editors, *Handbook for Optimization in Telecommunications*, chapter 23, pages 637–678. Kluwer, 2006.

[JMT07b] B. Jaumard, C. Meyer, and B. Thiongane. Comparison of ILP Formulations for the RWA Problem. *Optical Switching and Networking*, 2007.

[JMT07c] B. Jaumard, C. Meyer, and B. Thiongane. Decomposition Methods for the RWA Problem. *Discrete Applied Mathematics*, page to appear, 2007.

[JMY06b] B. Jaumard, C. Meyer, and X. Yu. How much wavelength conversion allows a reduction in the blocking rate ? *IEEE Journal of Optical Networking*, 5(12):881–900, 2006.

[JV05] Jason P. Jue and Vinod M. Vokkarane. *Optical Burst Switched Networks*. Springer, 2005.

[Kar00] S. Kartalopoulos. *Introduction to DWDM technology: Data in a rainbow*. Socitey of Photo-Optical, 2000.

[KB03] E. Kozlovski and P. Bayvel. Link failure restoration in wavelength-routed optical burst switched (WR-OBS) networks. *Proceedings of the IEEE OFC*, pages 222–223, 2003.

[KKK02] S. Kim, N. Kim, and M. Kang. Contention resolution for optical burst switching networks using alternative routing. In *Proceedings of the IEEE ICC*, volume 5, pages 2678–2681, 2002.

[KKY+06] Y.M. Kim, T.H. Kim, J.J. Yoo, B.W. Kim, and H.S. Park. Performance comparisons of restoration techniques for fault management in OBS networks. *Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference*, 2, 2006.

[KVJ03] R. Karanam, V. Vokkarane, and J. Jue. Intermediate node initiated (INI) signalling: a hybrid resevation technique for optical burst-switched networks. In *Proceedings of the IEEE OFC*, volume 1, pages 213–215, March 2003.

[LCA06] ITU-T G.7042/Y.1305 Link Capacity Adjustment Scheme (LCAS) for Virtual Concatenated Signals. 2006.

[LGC03a] J. Li, M. Gurusamy, and K. C. Chua. Load Balancing Using Adaptive Alternate Routing in IP-over-WDM Optical Burst Switching Networks. In *Proceedings of SPIE*, page 336, 2003.

[LGW01] A. Leon-Garcia and I. Widjaja. *Communication Networks - Fundamental Concepts and Key Architectures* -. MC Graw Hill, 2001.

[LNJ+05] V.T. Le, S.H. Ngo, X. Jiang, S. Horiguchi, and Y. Inoguchi. A Hybrid Algorithm for Dynamic Lightpath Protection in Survivable WDM Optical Networks. *Proceedings of the 8th International Symposium on Parallel Architectures, Algorithms and Networks*, pages 484–489, 2005.

[LQ04] J. Li and C. Qiao. Schedule burst proactively for optical burst switched networks. *Computer Networks*, 44(5):617–629, 2004.

[LQXX04] J. Li, C. Qiao, J. Xu, and D. Xu. Maximizing throughput for optical burst switching networks. *Proceedings of the IEEE INFOCOM*, 2004.

[LTL+06a] Y. Liu, E. Tangdiongga, Z. Li, H. de Waardt, A.M.J. Koonen, G.D. Khoe, H.J.S. Dorren, X. Shu, and I. Bennion. Error-free 320 Gb/s SOA-based Wavelength Conversion using Optical Filtering. *Proceedings of the IEEE OFC*, 2006.

[Mai08] M. Maier. *Optical Switching Networks*. Cambridge University Press, 2008.

[MC00] S. McCreary and K. Claffy. Trends in wide area IP traffic patterns. In *ITC Specialist Seminar*, 2000.

[Met05] A. Metnani. Routage par déflexion dans les réseaux tout optique à commutation de bursts. Master's thesis, Université de Montréal, 2005.

[MFPA09] G. Maier, A. Feldmann, V. Paxson, and M. Allman. On dominant characteristics of residential broadband internet traffic. In *Proc. ACM IMC*, 2009.

[Mim02] Mimi Dannhardt. Ethernet Over SONET. *White Paper*, February 2002.

[MLC+07] JM Martinez, Y. Liu, R. Clavero, AMJ Koonen, J. Herrera, F. Ramos, HJS Dorren, and J. Marti. All-Optical Processing Based on a Logic xor Gate and a Flip-Flop Memory for Packet-Switched Networks. *Photonics Technology Letters, IEEE*, 19(17):1316–1318, 2007.

[MZV+02] F. Masetti, D. Zriny, D. Verchère, J. Blanton, T. Kim, J. Talley, D. Chiaroni, A. Jourdan, J.-C. Jacquinot, C. Coeurjolly, P. Poignant, M. Renaud, G. Eilenberger, S. Bunse, W. Latenshleager, J. Wolde, and U. Bilgak. Design and Implementation of a Multi-Terabit Optical Burst/Packet Router prototype. In *Proceedings of the IEEE OFC*, pages FD1–1,FD1–3, March 2002.

[Nor06] Nortel. Delivering dynamic and efficient wavelength networking with eROADM technology. *White Paper*, 2006.

[OMKB09] I. Ouarda, M. Menif, M. Koubaa, and M. Bakri. Quality of transmission impact on routing and wavelength assignement rules in hybrid optical networks. In *Broadband Communications, Networks, and Systems, 2009. BROADNETS 2009. Sixth International Conference on*, pages 1 –7, sept. 2009.

[OMPR03] S. Ovadia, C. Maciocco, M. Paniccia, and R. Rajaduray. Photonic burst switching (PBS) architecture for hop and span-constrained optical networks. *IEEE Communications Magazine*, 41(11):S24–S32, 2003.

[PAM+05] H. Pan, T. Abe, Y. Mori, Y.B. Choi, and H. Okada. Feedback-based Load Balancing Routing for Optical Burst Switching Networks. In *Proceedings of the IEEE APCC*, pages 1033–1037, 2005.

[PCG+05] M.H. Phùng, K.C. Chua, M. Gurusamy, M. Motani, and T.C. Wong. The streamline effect in OBS networks and its application in load balancing. In *Proceedings of the IEEE Broadnets*, pages 304–311, Oct. 2005.

[PHNL06] Q.V. Phung, D. Habibi, H.N. Nguyen, and K. Lo. A Segmentation Method for Shared Protection in WDM Mesh Networks. *Proceedings of the IEEE ICON*, 2:1–6, 2006.

[PSCG06] M.H. Phúng, D. Shan, K.C. Chua, and M. Gurusamy. Performance analysis of a bufferless OBS node considering the streamline effect. *IEEE Communications Letters*, 10:293–295, April 2006.

[Qia00] Chunming Qiao. Labeled optical burst switching for IP-over-WDM integration. *IEEE Communications Magazine*, 38:104–114, September 2000.

[QWL06] Chunming Qiao, Wei Wei, and Xin Liu. Extending generalized multiprotocol label switching (GMPLS) for polymorphous, agile, and transparent optical networks (PATON). *IEEE Communications Magazine*, 44:104–114, Dec. 2006.

[QY99] C. Qiao and M. Yoo. Optical Burst Switching (OBS) - A New Paradigm for an Optical Internet. *Journal of High Speed Networks*, 8(1):69–84, January 1999.

[RCR05] L.C. Resendo, Ld.C. Calmon, and M.R.N. Ribeiro. Simple ILP approaches to grooming, routing, and wavelength assignment in WDM mesh networks. In *SBMO/IEEE MTT-S International Conference on Microwave and Optoelectronics*, pages 616–619, July 2005.

[RS95] R. Ramaswami and K.N. Sivarajan. Routing and wavelength assignment in all-optical networks. *IEEE/ACM Transactions on Networking*, 5(3):489–501, October 1995.

[RS07] S. Ramasubramanian and A. K. Somani. Dynamic Survivable Network Design for Path Level Traffic Grooming in WDM Optical Networks. In *Proceedings of the IEEE GLOBECOM*, pages 2359–2363, 2007.

[RVZW03] Z. Rosberg, Hai Le Vu, M. Zukerman, and J. White. Performance analyses of optical burst-switching networks . *IEEE Journal of Selected Areas in Communications*, 21:1187–1197, September 2003.

[RZG06] K. Ratnam, L. Zhou, and M. Gurusamy. Efficient Multi-Layer Operational Strategies for Survivable IP-over-WDM Networks. *IEEE Journal of Selected Areas in Communications*, 24(8):16–31, 2006.

[SCD+07] F. Solano, LF Caro, JC De Oliveira, R. Fabregat, and JL Marzo. G+: Enhanced Traffic Grooming in WDM Mesh Networks using Lighttours. *IEEE Journal of Selected Areas in Communications*, 25(5):1034–1047, 2007.

[SHM+05] Y. Sun, T. Hashiguchi, V.Q. Minh, X. Wang, H. Morikawa, and T. Aoyama. Design and Implementation of an Optical Burst-switched network testbed. *Communications Magazine, IEEE*, 43(11):848–855, 2005.

[SHZ07] B. Shihada, P.H. Ho, and Q. Zhang. TCP-ENG: Dynamic Explicit Congestion Notification for TCP over OBS Networks. In *Proceedings of the IEEE ICCCN*, pages 516–521, 2007.

[SJ09] S. Sebbah and B. Jaumard. A resilient transparent optical network design with a preconfigured extended-tree protection. In *In IEEE International Conference on Communications 2009*, pages 1–6, 2009.

[SME06] S. Said, H. Mouftah, and H. Elbiaze. A QoS-Based Restoration Mechanism for OBS Networks. In *Proceedings of the IEEE ICTON*, volume 3, pages 27–30, June 2006.

[SQ04a] S. Sheeshia and C. Qiao. Burst grooming in optical-burst-switched networks. *Matrix*, 1:3, 2004.

[SR04] K. Sathyamurthy and S. Ramasubramanian. Benefits of link protection at connection granularity. In *Proceedings of the IEEE Broadnets*, pages 300–309, 2004.

[SR09] Michele Savi and Carla Raffaelli. Hybrid contention resolution in optical switching fabric with qos traffic. In *Proceedings of the IEEE WOBS*, 2009.

[SS06a] A.A.M. Saleh and J.M. Simmons. Evolution toward the next-generation core optical network. *IEEE Journal of Lightwave Technology*, 24:3303–3321, Sept. 2006.

[SSS01] M. Sridharan, A. K. Somani, and M. Salapaka. Approaches for capacity and revenue optimization in survivable wdm networks. *Journal of High Speed Networks*, 10:109–125, 2001.

[SSS06] M. Sivakumar, K.M. Sivalingam, and S. Subramaniam. On Factors Affecting the Performance of Dynamically Groomed Optical WDM Mesh Networks. *Photonic Network Communications*, 12(1):15–28, 2006.

[Sys86] R. Syski. *Introduction to Congestion Theory in Telephone Systems*. Elsevier Science Ltd, 1986.

[TBZ+07] M. Tornatore, A. Baruffaldi, H. Zhu, B. Mukherjee, and A. Pattavina. Dynamic traffic grooming of subwavelength connections with known duration. *Proceedings of the IEEE OFC*, pages 1–3, 2007.

[TMP02] M. Tornatore, G. Maier, and A. Pattavina. WDM network optimization by ILP based on source formulation. In *Proceedings of the IEEE INFOCOM*, volume 3, pages 1813–1821, 2002.

[TOZ+05] M. Tornatore, C. Ou, J. Zhang, A. Pattavina, and B. Mukherjee. PHOTO: An Efficient Shared-Path-Protection Strategy Based on Connection-Holding-Time Awareness. *IEEE Journal of Lightwave Technology*, 23(10):3138–3146, 2005.

[TR05] J. Teng and G.N. Rouskas. Traffic engineering approach to path selection in optical burst switching networks. *IEEE Journal of Optical Networking*, pages 759–777, 2005.

[TVJ03] G.P.V. Thodime, V.M. Vokkarane, and J.P. Jue. Dynamic congestion-based load balanced routing in optical burst-switched networks. In *Proceedings of the IEEE GLOBECOM*, volume 5, pages 2628–2632, 2003.

[VBMS05] N.A. VanderHorn, S. Balasubramanian, M. Mina, and A.K. Somani. Light-trail test bed for IP-centric applications. *IEEE Commun. Mag*, 43(9), 2005.

[VCA02] ITU-T G.707 Network Node Interface for the Synchronous Digital Hierarchy. 2002.

[Vit95] A.J. Viterbi. *CDMA principles of spread spectrum communication Addison-Wesley wireless communications series.* Addison-Wesley Pub. Co., 1995.

[VJ03] V. Vokkarane and J. Jue. Burst Segmentation: an Approach for Reducing Packet Loss in Optical Burst Switched Networks. *Optical Networks Magazine*, 4(6):81–89, Nov./Dec. 2003.

[VJV09] B. Vignac, B. Jaumard, and F. Vanderbeck. Hierarchical optimization procedure for traffic grooming in WDM optical networks. In *Proceedings of the 13th international conference on Optical Network Design and Modeling*, pages 171–176. Institute of Electrical and Electronics Engineers Inc., The, 2009.

[VZJC02] V.M Vokkarane, Q. Zhang, JP. Jue, and B. Chen. Generalized Burst Assembly and Scheduling techniques for QoS Support in Optical Burst-Switched Networks. In *Proceedings of the IEEE GLOBECOM*, volume 3, 2002.

[WH97] Qin Wei and G. Hébuterne. Optical Deflection Networks based on "Manhattan Street" : torus or grid? In *Proceedings of the IEEE ICC*, 1997.

[WHLW03] H. Wen, R. He, L. Li, and S. Wang. Dynamic traffic-grooming algorithms in wavelength-division-multiplexing mesh networks. *IEEE Journal of Optical Networking*, 2(4):100–110, 2003.

[WM00] J.Y. Wei and R.I. McFarland. Just-in-time Signaling for WDM Optical Burst Switching Networks. *IEEE Journal of Lightwave Technology*, 18(12):2019–2037, December 2000.

[WSB02] Jian Wang, L. Sahasrabuddhe, and Mukherjee B. Path vs. subpath vs. link restoration for fault management in IP-over-WDM networks: performance comparisons using GMPLS control signaling. *IEEE Communications Magazine*, 40:80–87, Nov. 2002.

[WSLW05] H. Wen, H. Song, L. Li, and S. Wang. Load balancing contention resolution in obs networks based on gmpls. *International Journal of High Performance Computing and Networking*, 3(1):25–32, 2005.

[XQLX03] J. Xu, C. Qiao, J. Li, and G. Xu. Efficient channel scheduling algorithms in optical burst switched networks. In *Proceedings of the IEEE INFOCOM*, pages 2268–2278, 2003.

[XTKE+04] Yufeng Xin, Jing Teng, G. Karmous-Edwards, G.N. Rouskas, and D. Stevenson. Fault management with fast restoration for optical burst switched networks. In *Proceedings of the IEEE Broadnets*, pages 34–42, 2004.

[XWCL05a] Chunsheng Xin, Bin Wang, Xiaojun Cao, and Jikai Li. A heuristic logical topology design algorithm for multi-hop dynamic traffic grooming in WDM optical networks. In *Proceedings of the IEEE GLOBECOM*, volume 4, pages 2102–2106, Nov. 2005.

[XWCL05b] Chunsheng Xin, Bin Wang, Xiaojun Cao, and Jikai Li;. Formulation of multi-hop dynamic traffic grooming in WDM optical networks. In *Proceedings of the IEEE Broadnets*, volume 2, pages 1203–1208, Oct. 2005.

[YCQ02] X. Yu, Y. Chen, and C. Qiao. A Study of Traffic Statistics of Assembled Burst Traffic in Optical Burst Switched Networks. In *Proc. Opticomm*, volume 4874, pages 149–159, 2002.

[YJQ98] M. Yoo, M. Jeong, and C. Qiao. A new optical burst switching (OBS) protocol for supporting quality of service. *Proceedings of the SPIE*, 3531:396–405, November 1998.

[YLC05] O. Yu, M. Liao, and Y. Cao. Synchronous stream optical burst switching. In *Proceedings of the IEEE Broadnets*, pages 524–529, 2005.

[YLC06] O. Yu, M. Liao, and Y. Cao. Multi-granular Stream Optical Burst Switching. In *Proceedings of the IEEE Broadnets*, pages 1–10, 2006.

[YQ99] M. Yoo and C. Qiao. Supporting multiple classes of services in IP over WDM networks. *Global Telecommunications Conference, 1999. GLOBECOM'99*, 1, 1999.

[YQL04] X. Yu, C. Qiao, and Y. Liu. TCP Implementations and False Time Out Detection in OBS Networks. In *Proceedings of the IEEE INFOCOM*, volume 2, pages 774–784, march 2004.

[YR04b] W. Yao and B. Ramatnurthy. Constrained Dynamic Traffic Grooming in WDM Mesh Networks with Link Bundled Auxiliary Graph Model. *Wksp. High Performance Switching and Routing*, pages 287–291, 2004.

[YR04c] W. Yao and B. Ramatnurthy. Rerouting Schemes for Dynamic Traffic Grooming in Optical WDM Mesh Networks. *Proceedings of the IEEE GLOBECOM*, 3:1793–1797, 2004.

[YR06] L. Yang and G.N. Rouskas. Adaptive path selection in optical burst switched networks. In *IEEE Journal of Lightwave Technology*, volume 24, pages 3002–3011, 2006.

[ZM02b] Keyao Zhu and B. Mukherjee. Traffic grooming in an optical WDM mesh network. *IEEE Journal of Selected Areas in Communications*, 20:122–133, Jan. 2002.

[ZM04a] J. Zhang and B. Mukheriee. A review of fault management in WDM mesh networks: basic concepts and research challenges. *Network, IEEE*, 18(2):41–48, 2004.

[ZXTT05] W. Zhang, G. Xue, J. Tang, and K. Thulasiraman. Dynamic light trail routing and protection issues in WDM optical networks. In *Proc. IEEE Globecom 2005*, 2005.

[ZYL04b] Y. Zhang, O. Yang, and H. Liu. A Lagrangean relaxation and subgradient framework for the routing and wavelength assignment problem in WDM networks. *IEEE Journal on Selected Areas in Communications*, 22:1752 – 1765, November 2004.