PART III

OPS-based wide area networks

Chapter 7

Introduction to the OPS-based wide area network

7.1 State-of-the-art

In this thesis, we consider the general switch architecture with full connectivity and wavelength conversion shown in Figure 7.1 and capable to switch asynchronous, variable-length packets [35]. This switch acts as an output queueing switch; it uses a feed-forward configuration [60] and the optical buffer is made by B FDLs. The electronic Switch Control Logic (SCL) takes all the decisions regarding the configuration of the hardware to realize the proper switching actions. When a packet arrives, the SCL examines the header and lookups the forwarding table to determine the output fiber, determining also the network path. Successively, the SCL performs the following functions:

- choose which wavelength of the output fiber will be used to transmit the packet, in order to properly control the output interface;
- decide whether the packet has to be delayed by using the FDLs or it has to be dropped, since the required queuing resource is congested.

These decisions are routing independent and all the wavelengths of a given output fiber are equivalent for routing purposes but are not from the contention resolution point of view. The choices of wavelength and delay are actually correlated, being the need to delay a packet related to the availability of the wavelength selected. This is what we call the *Wavelength and Delay Selection* (WDS) problem. We therefore consider contention resolution policies able to exploit only the *time and wavelength* domains and not the space domain.

The technology limitation of the optical queuing motivates significant research efforts in recent years dealing with the design of simple WDS contention resolution policies (see for instance [48] [99] [15]). Almost all of them solve the contentions on a per-packet basis, i.e. the WDS algorithm is executed at each incoming packet (we call this approach *connectionless*). This means that once the forwarding component has

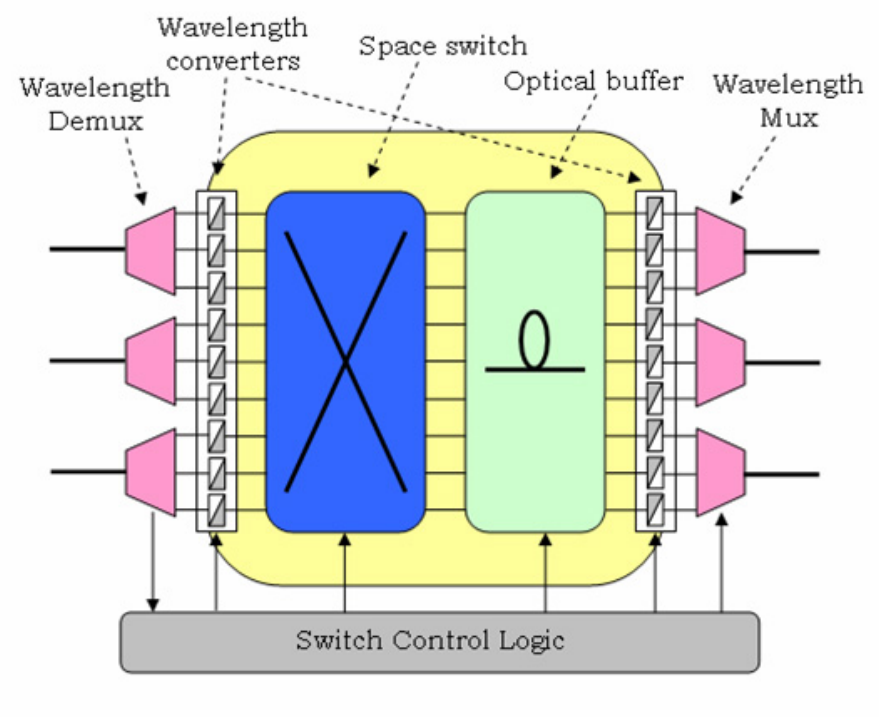

Figure 7.1: The considered switch architecture

decided to which output port p the packet should be sent, the functions performed by the SCL are:

1. Search for the set of wavelength $W \in p$ not busy;
   (a) If $W = \emptyset$ (i.e., all queues are full), the packet is lost;
   (b) If $W \neq \emptyset$, select the wavelength $w \in W$ to transmit the packet on;
2. Determine the delay D_j and select the FDL j to send the packet to.

The choice of the wavelength in step 1.b is the key point and can be implemented by following different policies, producing different processing loads at the SCL and different resource utilizations. In [15] several heuristic WDS algorithms were presented and studied, showing that the choice of the algorithm may significantly change the performance. In fact, when a packet has to be buffered, the choice of the delay is not free, since the number of delays available within a FDL buffer is discrete. As explained in [14], this creates gaps between queued packets that can be considered equivalent to an increase of the packet service time, meaning an artificial increase in the traffic load (*excess load*). It has been demonstrated in [99] that a WDS algorithm (called VOID algorithm) that aims at minimizing those gaps gives best performance with respect to other policies. Nonetheless, the computational complexity of the VOID algorithm is very high since it requires to know the length and the duration of every gap in the queues. A simplification of this algorithm called MINGAP is proposed in

[15]; it selects the FDL with the minimum gap only between the last queued packet and the new one.

Despite the fact that the WDS algorithm can be relatively simple to implement, taking per-packet decisions requires too much computations considering that each switch has several ports, each port several wavelengths and each wavelength transports packets at 10 Gbit/s or above. To overtake this problem, OPS concepts are recently extended to a connection-oriented network scenario [16], for instance based on MPLS. In this scenario, a suitable design of WDS algorithms permits to obtain fairly good performance, by exploiting queuing behaviors related to the connection-oriented nature of the traffic, but with a significant saving in term of processing effort for the switch control with respect to the connectionless case.

7.2 The connection-oriented OPS network

The connection-oriented OPS network comprises several nodes connected in a mesh topology. Based on destination address and quality of service requirements, packets coming from the client networks are classified at the edge nodes into a finite number of subsets such as the *Forwarding Equivalent Classes* (FECs) concept defined in MPLS environment. Each FEC is identified by an additional *label* added to the packets. Edge nodes are in charge of setting up and maintain the unidirectional Optical Virtual Circuits (OVCs) throughout the network. Packets belonging to the same FEC are identical from a forwarding point of view and are transferred from source to destination along the OVC which corresponds to their label. On each core node, a simple label matching operation is performed on a pre-computed OVC forwarding table, thus simplifying and speeding up the forwarding function.

Due to the high number of traffic flows being typically transported by a WDM network, the adoption of a pure MPLS-like labeling scheme may result in an excess of per-flow information to be handled by the optical nodes. In order to avoid scalability problems, here we assume that each OVC represents a top-level explicitly routed path formed by an aggregation of lower-level connections including several traffic flows such as what proposed in [68]. Following this approach, the number of OVCs managed by a single optical core router is not supposed to be too high and to affect the correct label processing.

Figure 7.2 shows an example of the connection-oriented OPS network. An OVC forwarding table is setup in the node 1 which indicates that packets with label 25 coming from port 0 with a pink wavelength should be forwarded to the output port 2 with a blue wavelength and a new label 12.

While the information on the output fibre is given by the routing protocols (no subject of study here), the choice of the wavelength may be taken locally by each node taking into account the availability of time resources. This problem can be solved by following different strategies:

- **Static**. The OVC is assigned to a given wavelength at OVC setup and this assignment is hold over the OVC life. Therefore packets belonging to the same

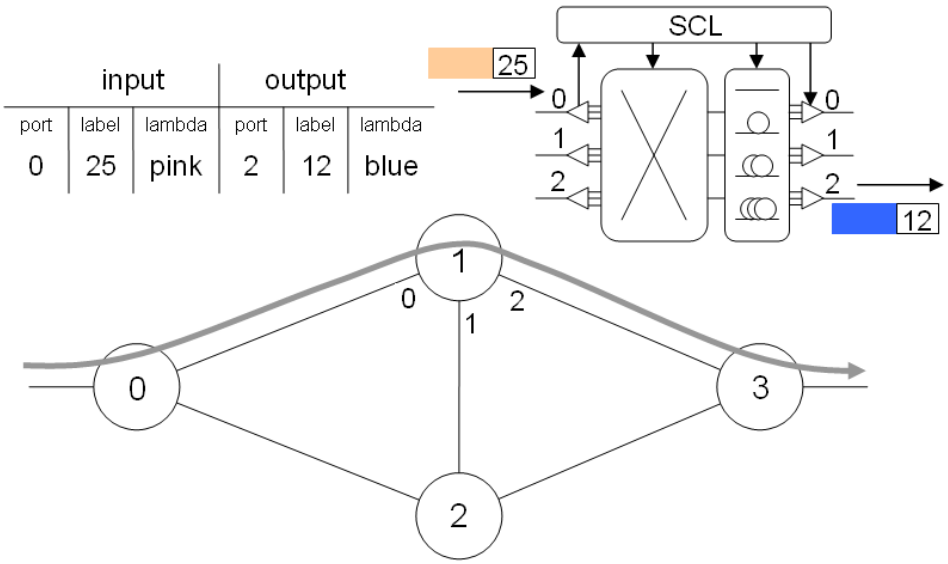

Figure 7.2: Connection-oriented OPS network

OVC are always switched to the same wavelength and the contentions can be only solved in time;

- **Dynamic**. The OVC is assigned to a wavelength at OVC setup but it can be changed during OVC life. When heavy congestion arises on the assigned wavelength (i.e., when the time domain cannot solve a contention), the OVC is temporary switched to another wavelength. When congestion disappears, the OVC is switched back to the original wavelength.

The static wavelength selection requires minimum control complexity since processing is performed only at OVC setup. At the same time, it preserves the correct order of packets belonging to the same OVC since new arrivals cannot overtake older packets. However it does not optimize the resources obtaining high PLR figures. On the other hand, when a dynamic algorithm is executed, the OVC is switched to an alternative wavelength that is not (or is less) congested and new incoming packets on that OVC will experience in general less queuing time than older packets and will very likely overtake them along the network path. At the same time, the amount of execution of the algorithm affects the processing load on the SCL ranging from no efforts if static approach is used to fairly demanding efforts if a new wavelength search is executed per each incoming packet (e.g., [99] [15]).

Delivering out-of-order packets is a serious problem since causes expensive reordering operations to be performed at the edges of the optical network and makes mandatory the use of very large memories due to the high speed of optical links. It is demonstrated in [71] [65] that even a small percentage of out-of-ordered packets seriously affects TCP behavior, causing a considerable throughput degradation at the application level. We focus on this problem in Chapter 9.

Static

When a packet belonging to OVC l arrives:

1. Look-up the forwarding table to determine $n(l)^{out}$ and $\lambda^{out}(l)$ assigned to l;
2. Determine the minimum delay D_j available for $\lambda^{out}(l)$;
   (a) If $D_j > D_M$, the packet is lost;
   (b) If $D_j \leq D_M$, select the FDL j to send the packet to.

Dynamic

When a packet belonging to OVC l arrives:

1. Look-up the forwarding table to determine $n^{out}(l)$ and $\lambda^{out}(l)$ assigned to l;
2. Determine the minimum delay D_j available for $\lambda^{out}(c)$;
3. If $D_j > D_M$ (i.e., $\lambda^{out}(l)$ is congested), search for the set of wavelength $\Lambda \in n^{out}(l)$ not busy;
   (a) If $\Lambda = \emptyset$, the packet is lost;
   (b) If $\Lambda \neq \emptyset$, select the new wavelength $\lambda_{new} \in \Lambda$ to transmit the packet on and update the forwarding table $\lambda^{out}(l) = \lambda_{new}$;
4. Determine the delay D_j and select the FDL j to send the packet to.

Figure 7.3: Contention resolution techniques in connection-oriented OPS networks.

7.3 Problems addressed in this thesis

In this thesis we focus on a connection-oriented OPS network scenario taking into account both static and dynamic approach. We address two problems, namely the problem of setting up of the OVC, properly configuring the forwarding table at the nodes, and the problem of providing QoS.

Concerning the former problem, at the OVC setup, each node must assign both the output port and the output wavelength to the OVC in such a way that the packets belonging to that OVC are always switched to the same output. This double setup problem is different with respect to the *classical* RWA problem in circuit-switched network because here the wavelengths are shared among several OVCs (in a packet-switched basis). In this study we do not deal with the problem of selecting the output port which depends on the routing protocol but we are interested in the election of the wavelength which may be set locally by each node using a *OVC-to-wavelength setup assignment* (OWSA) algorithm. In particular we show that intelligent OWSA procedures can considerably improve the performance of the switches. The intelligence relies on grouping the flows coming from the same input wavelength which allows to obtain the conflict-free situations and hence reduce the contention probability.

Concerning the latter problem, existing solutions to provide QoS in OPS networks are based on the following strategy: 1) design a contention resolution algorithm which minimizes the Packet Loss Rate (PLR), thus 2) apply a QoS mechanism (some form of resources reservation on top of the contention resolution algorithm) able to differentiate the PLR among two or more classes. Given that we are dealing with a connection-oriented model, here we suggest a new method based on the well known ATM scheme of defining different service categories which consists of defining different OPS service categories, each one based on a different contention resolution algorithm specifically designed to cope with the requirements of that category. With this technique, besides the PLR, also the preserving the correct packet sequence and the computational complexity can be considered as important metrics for the QoS provisioning problem.

Let us spent some words on the out-of-order packet delivery problem. As already outlined above it is well known that packet loss as well as out-of-order packet delivery and delay variations affects end-to-end protocols behavior and may cause throughput impairments [65] [71].

When considering TCP-based traffic it is well known that these phenomena influence the typical congestion control mechanisms adopted by the protocol and may result in a reduction of the transmission window size and consequently in bandwidth under-utilization. In particular the TCP congestion control is very affected by the loss or the out-of-order delivery of bursts of segments. This is exactly what may happen in the OPS network where traffic is typically groomed and several IP datagrams (and therefore TCP segments) are multiplexed in an optical packet, because optical packets must satisfy a minimum length requirement to guarantee a reasonable switching efficiency. Therefore out-of-order or delayed delivery of just one optical packet may result in out-of-order or delayed delivery of several TCP segments triggering (multiple duplicate ACKS and/or timeouts that expire) congestion control mechanisms and

causing unnecessary reduction of the window size.

Another example of how out-of-sequence packets may affect application performance is the case of delay-sensitive UDP-based traffic, such as real-time traffic. In fact unordered packets may arrive too late and/or the delay required to reorder several out-of-sequence packets may be too high with respect to the timing requirements of the application.

These brief and simple examples make evident the need to limit the number of unordered packets. In general out-of-order delivery is caused by the fact that packets belonging to the same flow of information can take different paths through the network and then can experience different delays [5]. In traditional connection-oriented networks, packet reordering is not an issues since packets belonging to the same connection are supposed to follow the same virtual network path and therefore are delivered in the correct sequence, unless packet loss occurs.

In an OPS network, using the wavelength domain for contention resolution (i.e. using dynamic policies), this may not be the case. Packets traveling along the same network path may use different wavelengths according to the choices of the algorithm. Therefore it may happen that packets of the same flow are delivered out of sequence, even though still following the same network path. Intuitively the reason is that the OVC is switched to an alternative wavelength that is less congested and new incoming packets on that OVC will experience less queuing time than older packets.

A possible solution could be to assume that this problem is solved at the egress edge-nodes that should take care of re-sequencing the various packet flows. This assumption in our view is not very realistic. It can be feasible for some flow of high value traffic, but is unlikely that will happen for all the flows of best effort traffic, because of the amount of memory and processing effort that would be necessary. Therefore we argue that it is important and necessary to control out-of-order delivery of packets directly in the OPS network nodes.

As explained above, this may cause an undesirable reduction in throughput as well as costly reordering operations with consequent unacceptable delays.

7.4 Simulation scenario

The performances of the proposed mechanisms are evaluated in order to assess their merits. The simulation results presented in the following section have been obtained by means of an ad-hoc event-driven simulator of the optical packet switch. The parameters of the switch are:

- N indicates the number of input and output fibers;
- W indicates the number of wavelengths per fiber;
- $\mathbf{Q}_B$ indicates the set of possible delays of B FDLs;
- D indicates the delay granularity;
- L indicates the average number of OVCs per input wavelength.

- ρ indicates the offered load.
- $\mathbf{M}$ indicates the fibre-to-fibre traffic matrix, whose generic element $\mathbf{M}_{i,j}$ is a real number ranging between 0 and 1 representing the percentage of traffic coming from input fibre i and going to output fibre j with respect to ρ. Three different traffic matrix are defined named: *uniform* $\mathbf{M}^U$, *power-of-two* $\mathbf{M}^P$, and *unbalanced* $\mathbf{M}^B$. For the case of $N = 4$, the matrices are as follows:

$$M^U = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad M^P = \frac{1}{15} \begin{bmatrix} 1 & 2 & 4 & 8 \\ 2 & 4 & 8 & 1 \\ 4 & 8 & 1 & 2 \\ 8 & 1 & 2 & 4 \end{bmatrix}$$

$$M^B = \frac{1}{30} \begin{bmatrix} 15 & 0 & 0 & 0 \\ 15 & 3 & 10 & 2 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 10 & 0 \end{bmatrix}$$

  These matrices represent a good sample of all possible traffic patterns: the classical uniform matric to evaluate a fair situation, the power-of-two matrix which demonstrates performance degradations when applied to the switches, and the unbalanced matrix to consider not balanced situations.

The distribution of the OVC requests follow an exponential model: both the interarrival time and connection duration of the OVCs are exponential distributed. The mean value of the interarrival times, connection duration, and required bandwidth are selected accordingly to generate the required offered load ρ.

The interarrival time of the packets is exponential distributed with a mean that depends on the OVC bandwidth. The packets are an exponential distributed size with average and minimum lengths of 500 and 40 bytes respectively.

We define the following measures to evaluate the performance of the switch:

- *Average Packet Loss Rate* (PLR). It is the usual performance measure for packet switches and also indicates the capability of an algorithm to reduce the congestion situation.
- *Out-of-sequence packets* (OS). The out-of-sequence packets delivery causes considerable throughput degradations and delay increases [71] due to the expensive reordering operations to be performed at the edges of the optical network. This measure indicates the percentage of out-of-sequence packets belonging to the same OVC.
- *Forwarding opacity* (FO). It is measured as the percentage of packets that are forwarded searching a new wavelength over the total number of simulated packets. The resulting value estimates the overload on the switch control function. The higher the percentage, the higher the overload.

In the following performance evaluation sections, we will show only the most significant measures according to the purposes of the study.