Optical switching and wire-speed packet processing in optical backbone networks

P. Pavón-Mariño, J. Veiga-Gontán, J. García-Haro

Department of Information Technologies and Communications, Polytechnic University of Cartagena, Plaza del Hospital 1, Edificio Antigones, E-30202, Spain.
pablo.pavon@upct.es, javg@alu.upct.es, joang.haro@upct.es

Abstract

In this paper, the wire-speed packet processing issue is analyzed for Optical Packet/Burst Switching (OPS/OBS) switching nodes, and wavelength-routing (WR) switching nodes. Some assumptions are made which allow to obtain simple estimates of these values, which are then compared with the packet processing performance of commercial electronic devices which implement this type of tasks in state-of-the-art high-performance packet switches. Results show that in the OPS/OBS alternatives, wire-speed packet processing cannot be considered a challenge. Furthermore, it is not dependent on the binary rate of packet payload. In wavelength routing networks, the reduction in electronic processing provided by transparent lightpath switching in wavelength routing networks is considered. Nevertheless, in these networks the number of packets to process is tied to the transmission binary rates. Then, this alternative reduces, but does not eliminate, the packet-processing bottleneck. This situation can promote in the future the OPS/OBS technologies.

1. Introduction

Optical WDM backbone networks interconnect major traffic centres, covering large distances. The switching nodes play a crucial role in this type of networks. Transmission bandwidth given by Wavelength Division Multiplexing (WDM) fiber links is in the order of tens of terabits per second. The interest is now in the design of switching mechanisms able to provide the required bandwidth distribution among traffic sources.

There are several proposals to implement these switching nodes. These proposals can be classified according to the degree of optical processing suffered by the traffic (supposed to be packet-based). In this classification, the deepest optical processing corresponds to those proposals where the switching and packet processing is made at the optical domain. The LASAGNE project [ram05] describes one of these alternatives. The lowest degree of optical processing occurs in the Optical Electronic Optical (O-E-O) switching technology [yoo96], where only transmission is performed at the optics. In between of both of the proposals there are a range of electronic/optical switching

combinations, like Wavelength Routing (WR), Optical Burst Switching (OBS) and Optical Packet Switching (OPS) [baw06].

In this paper, we address a feasibility study of the different approaches, from the point of view of the electronic control unit of the switch, and the number of switching decisions that should be taken per second. We define a switching decision event as, a traffic arrival to the node which requires scheduling and table lookup. Note that, regardless of the switching technique employed, these decisions in the switching nodes are still performed electronically. Therefore, this study is relevant and yields to a comparison among the different alternatives, which are inherently different:

- The decision event in OPS and OBS nodes occurs for each optical packet or burst arriving to the node. This is independent of the binary rate of the packet/burst payload. In this paper, we consider OPS nodes in which packet duration is in the order of 1 μs, while in OBS nodes burst duration is a parameter, ranging from tens of microseconds, to tens of milliseconds. A major difference between OPS and OBS nodes is given by the use of a dedicated control wavelength in OBS networks where the burst header information is transmitted, potentially in advance to the burst payload. This difference is not an issue in our discussion, as it does not affect the throughput of switching decisions to be taken.
- In WR nodes, the traffic from the lightpaths ending at the node is electronically switched. This corresponds to egress traffic to the node, and grooming traffic. Note that in this switching paradigm, the transmission rate and packet processing rates are coupled: the higher the transmission rate, the higher the number of packets per second to process.
- In O-E-O nodes, all the arriving traffic to the node has to be processed. Again, transmission rates and processing rate requirements are coupled.

The objective of this paper is to assess the packet processing throughput required by the electronic control units of the different switching proposals, comparing the processing capacities given by the state-of-the-art technology.

The rest of the paper is organized as follows. Section 2 summarizes the packet processing technologies suitable to build the control units in the optical nodes. Section 3 compares the state-of-the-art processing throughputs to the processing requirements estimated for the different optical switching alternatives. Finally, section 4 concludes.

2. Packet processing technologies

The packet traffic in IP networks is commonly divided into control plane traffic and data-plane traffic. The control plane traffic is composed by signalling information distributed across the network. It arrives sporadically, being a minimum percentage of the total traffic, and, in general, has no hard time restrictions. The processing of these packets is usually complex, and it may involve changes in the router state (i.e. the routing table). Examples of this type of traffic are the OSPF or BGP traffic in IP networks, or LDP traffic in MPLS networks. Electronic units of the switching nodes assign the processing of this type of traffic to general-purpose CPUs (Central

Processing Units) in the system. This type of packet processing is named the switch *slow-path*.

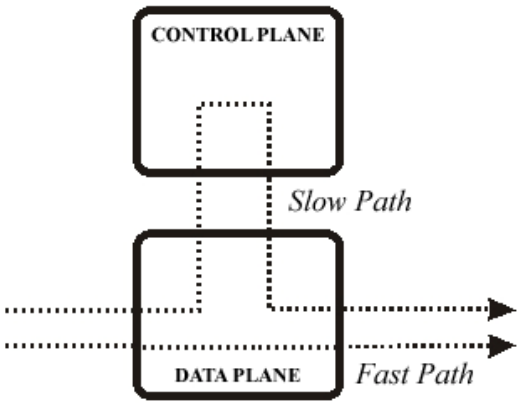

Fig. 1. Fast-path and slow-path

The "user" arriving traffic to the switching node, which does not carry control/signalling information and does not require payload processing, is named the data-plane traffic. It represents the major volume of network traffic. The switching node must perform a fast packet header processing, to forward the packet satisfying the appropriate Quality of Service (QoS). This type of processing is much simpler than the one required to control plane traffic packets. Therefore, in backbone switching nodes, this traffic is not usually processed by general purpose CPUs, but by specialized systems capable of a much faster packet processing time. This is called the switch *fast-path*. A typical sequence of fast-path processing involves:

(i) Store incoming packets in high capacity memories (usually DRAM memories).

(ii) Store packet header information in higher speed access memories (like SRAM memories).

(iii) Packet forwarding decision: This step consists on the necessary lookups to obtain the outgoing route of the packet. Fast lookups are performed by means of Ternary Context Addressable Memories (TCAM) or specific ASICs. The bottleneck of lookups is the number of memory accesses required. The benefit of using TCAM is to parallelize the lookup process. Solutions based on ASIC use embedded SRAM memories to achieve more random memory accesses per second than TCAM [dha03]. The forwarding tables depend on the protocols employed (IP routing tables, Next Hop Label Forwarding Entry (NHLFE), Incoming Label Map (ILM) and Stream-to-NHLFE (STN) tables in MPLS nodes.

(iv) Packet classification into different packet queues, according to QoS information. Implementation of possible early discard policies to prevent congestion. Scheduling decision on which queue is to be served, selecting the next packet to be transmitted

(attending to possible traffic shaping and traffic service contract issues). Core nodes usually implement a light version of these policies.

Several alternatives are employed in packet switching nodes to implement the mentioned tasks in the fast-path:

- *General purpose processor*. The option of using a general purpose processor in the fast-path implies both advantages and drawbacks. On one side, these systems can be programmed by using well-known high level languages, which speed up the time to market development cycle. On the other hand, the number of packets per second that can be processed is usually low. As an example, the Motorola ATCA-F300, based on MPC8245 PowerPC publishes a throughput of 1Gbps, (aprox. $1.5x10^6$ packets per second) [mot07].
- *Application-Specific Integrated Circuit (ASIC)*. It is an integrated circuit (IC) customized for a particular use. On one hand, the performance achieved with this solution is the highest possible with electronic technology. On the other hand, the cost to design and produce a specific ASIC is also very high, and the usual development time is between one and two years [hus00]. A long development cycle implies to forecast what features will need to be supported when the product is in the stores. Therefore, this solution is the less flexible. As an example, IBM research has designed a 4-Terabit packet switch using CMOS ASIC technology [abe03].
- *Field Programmable Gate Array (FPGA)*. It is a programmable device, with a shorter development cycle than ASICs. The disadvantages are high power consumption and higher cost. FPGA has a fixed architecture that implies the impossibility to achieve the process speed of an ASIC. Laboratory demonstrations have achieved up to 10Gbps with FTP connections using a FPGA [shr06].

  The latest proposals employ ASICs with FPGAs. The idea behind is to use ASIC for the parts of more time critical packet processing to achieve good performance and employ the FPGA for the configurable processing parts. The designer may predict which parts are likely to require a change on the processing type. This alternative maintains the disadvantage of ASICs of a very long development time.
- *Network Processors (NP)*. These are programmable processors specialized and optimized to be used in applications involving network routing and packet processing. An increasing set of architectures have arose in the last years. Typically, network processors include on-chip multiple small scale simple RISC processors running in parallel, devoted to the fast-path, and one general purpose processor for control plane packets (slow-path). This solution is situated in between of general purpose processor and ASIC based systems. Their advantages are a short time to market because of the flexibility given by the NP programming in specialized C-like languages. The NP vendors also offer libraries implementing the conventional protocol processing at different layers (SDH, PPP, IPv4, IPv6, MPLS, etc.), to speed-up the development process. The packets per second performance varies among different vendors and models. As an example, the Bay Microsystems Chesapeake NP model documentation [bay07] publishes a performance figure of $122x10^6$ packets per second, suitable for 40 Gbps interfaces. Table 1 briefly summarizes the published performance metrics of a set of NP products commercially available. Note that these values are strongly dependent on the particular packet processing performed, are in general brought from vendor catalogues, and should be compared with a great care.

Indeed they are expressed in different units and computed under slightly different conditions.

In addition, some preliminary projects exist to implement some of the tasks in the fast-path by means of optical processing. As an example, an all-optical contention resolution [ram05] scheme has been proposed in the LASAGNE project for OPS nodes. The tasks optically implemented are the decision on the optical delay to be assigned to the packet, processing the optical label attached to it. In [mur02] optical processing is also proposed to implement optically the delay assignment in OPS switches. Unfortunately these alternatives, although encouraging, are still in a preliminary stage.

Table 1. Performance values for commercial Network Processors. Information supplied by manufacturers.

MANUFACTURER	MODEL	PERFORMANCE
Bay Microsystems	Chesapeake	40 Gbps $122x10^6$ pps
Bay Microsystems	Biscayne	10 Gbps $41.5x10^6$ pps
Bay Microsystems	Montego	10 Gbps $31.25x10^6$ pps
Intel	IXP2855	10 Gbps $15x10^6$ pps *
Intel	IXP2400	4 Gbps $6x10^6$ pps *
IBM	PowerNP NP4GS3	3 Gbps $4.5x10^6$ pps
Vitesse	VSC2232	4 Gbps $6x10^6$ pps *
Vitesse	VSC2200	2,5 Gbps $3.75x10^6$ pps *
Agere	APP650	2,5 Gbps $3.75x10^6$ pps *
Agere	APP530	2,5 Gbps $3.75x10^6$ pps *
AMCC	nP7250	2,5 Gbps $3.75x10^6$ pps *

(*) Estimated values.

3. Performance figures comparison

In this section we try to address the following question for different optical switching technologies: are the packet processing issues a challenge when compared to packet processing throughputs achieved today?
To aid in the estimation of the decision throughput requirements, we focus on an example switching node, with 50 input WDM channels, and 50 output channels (figure 1). The number of input and output fiber links in which these channels are distributed is not relevant for our study. We suppose 60% utilization in each WDM channel. We

consider two separated examples, for nodes receiving traffic at 10 Gbps transmission rate, and 100 Gbps transmission rate for each WDM channel.

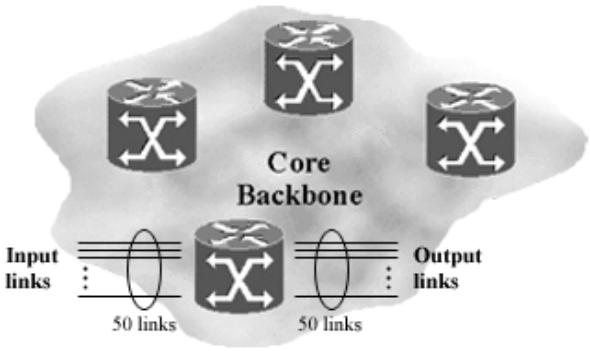

Fig. 2. Illustration of a core node with 50 input links and 50 output links

The number of decision events per second for each technology is calculated as follows:

- *OPS*. Considering an average packet duration of 1 μs (as proposed in [dit03]), and an average load of the 60%, in average 30 packet arrivals occur each microsecond. This yields to $3x10^7$ decisions per second. Note that these decision rates do NOT depend on the transmission rate employed in the packet payload.
- *OBS*. Several average burst sizes can be considered: {10, 100, 1000, 10000, 100000} μs (that is, from 10 μs to 10 ms). This yields to a decision rates ranging from $3x10^6$ to $3x10^3$ decisions per second. Again, these decision rates are independent from the transmission rate inside the burst payload.
- Wavelength-routing nodes/O-E-O nodes. We suppose an average packet size of 125 bytes. Te distinction between wavelength routing nodes and pure O-E-O nodes is given by the fact that in WR nodes, a fraction of the traffic traverses the node, without suffering electronic processing. The fraction employed depends on the network planning, and cannot be generalized. In our tests we consider three different degrees of O-O-O fractions of traffic: 70%, 50% and 20%. O-E-O switches can be thought as WR nodes with a 0% of O-O-O traversing traffic. The higher the degree, the lower the number of packets to be electronically processed.

Table 2 summarizes the average requirements in number of decision events per second in each of the above scenarios.

In the authors' opinion, some conclusions can be derived from table 2. First, the packet processing issue can be hardly graded as a challenge in OPS switching nodes, as the processing throughput is within the range of current commercial electronic systems. In OBS networks, the packet processing issue can be simply removed from the design constraints. In both switching technologies, the cost of the optical equipment is then the major design concern. The results obtained lead us to assert that switch designs should

not increase the cost of the optical side of the node, with the objective of reducing the electronic processing requirements.

Table 2. Average packet processing throughput required for different switching paradigms at 10 Gbps and 100 Gbps transmission rates.

LINKS / PACKET SIZE	10 GBPS	100 GBPS
OPS 1 µs	$30x10^6$ pps	$30x10^6$ pps
OBS 10 µs	$3x10^6$ pps	$3x10^6$ pps
OBS 100 µs	300,000 pps	300,000 pps
OBS 1 ms	30,000 pps	30,000 pps
OBS 10 ms	3,000 pps	3,000 pps
WR 70% traffic O-O-O	$90x10^6$ pps	$900x10^6$ pps
WR 50% traffic O-O-O	$150x10^6$ pps	$1{,}500x10^6$ pps
WR 20% traffic O-O-O	$240x10^6$ pps	$2{,}400x10^6$ pps
O-E-O	$300x10^6$ pps	$3{,}000x10^6$ pps

Regarding the wavelength-routing networks and O-E-O processing, a different analysis should be made. This is because, in both alternatives, the packet processing rate is tied to the binary rate of the packet transmission. As binary rates increase, harder constraints appear. The wavelength-routing approach provides a way to reduce the amount of traffic to be processed electronically, alleviating the packet processing problem. It is still unclear if the increase in the binary rates will force in the future to migrate to OPS/OBS switching paradigms where transmission rate and processing rate issues are decoupled.

5. Conclusions

In this paper, the packet processing issue in switching nodes is analyzed for different optical switching paradigms. The requirements in the number of decision events per second in different scenarios are compared to the performance of commercial electronic systems devoted to this type of processing in current electronic switches. Results show that packet processing speed is not a challenge in OPS and OBS nodes. On the other side, the coupling between transmission rate and processing rate that occur in WR and O-E-O switching, can be the source of a narrow bottleneck in the future. This could happen only if the improvements in the packet processing speed could not cope with the increase in the packet transmission rates. In authors' opinion, such a technological scenario, would strongly speedup the application of OPS/OBS switching techniques.

Acknowledgements

This research has been supported by the Spanish MCyT project grant TEC2004-05622-C04-02/TCM (ARPaq). Authors would like to thank also the e-Photon/One project and the COST 291 action.

References

[abe03] F. Abel, C. Minkenberg, R. Luijten, M. Gusat, I. Iliadis, "A Four-Terabit Packet Switch Supporting Long Round-Trip Times", IEEE Micro Magazine, vol. 23, no. 1, pp. 10–24, Jan./Feb. 2003.

[baw06] El-Bawab, Tarek S. "Optical Switching", Springer, 1 edition April 5, 2006.

[bay07] www.baymicrosystems.com, last consulted on Feb. 4, 2007.

[dha03] S. Dharmapurikar, P. Krishnamurthy, and D. E. Taylor, "Longest Prefix Matching using Bloom Filters," in ACM SIGCOMM'03, August 2003.

[dit03] Dittman L., et al.: The European IST Project DAVID: A Viable Approach Toward Optical Packet Switching, IEEE Journal on Selected Areas in Communications, vol. 21, no. pp. 1026-1040, 7, Sep. 2003.

[hus00] D. Husak, "Network Processors: A Definition and Comparison", White paper, C-PORT, http://www.cportcorp.com/solutions/docs/netprocessor_wp5-00.pdf, May. 3, 2000.

[mot07] www.motorola.com, last consulted on Feb. 4, 2007.

[mur02] M. Murata and K. Kitayama, "Ultrafast photonic label switch for asynchronous packets of variable length," Proc. IEEE INFOCOM 2002, New York, June 23–27, 2002.

[ram05] F. Ramos, E. Kehayas, J. M. Martinez, R. Clavero, J. Marti,L. Stampoulidis, D. Tsiokos, H. Avramopoulos, J. Zhang, P. V. Holm-Nielsen, N. Chi, P. Jeppesen, N. Yan, I. T. Monroy, A. M. J. Koonen,M. T. Hill, Y. Liu, H. J. S. Dorren, R. V. Caenegem, D. Colle, M. Pickavet,and B. Riposati, "IST-LASAGNE: Towards all-optical label swappingemploying optical logic gates and optical flip-flops," J. Lightw. Technol.,vol. 23, no. 10, pp. 2993–3011, Oct. 2005.

[shr06] H. Shrikumar, "40Gbps de-layered silicon protocol engine for TCP record," IEEE EDAA, pp. 188–193, 2006.

[yoo96] S. J. B. Yoo, "Wavelength conversion technologies for WDM network applications," IEEE J. Light. Tech., 14 (6), pp. 955–966, 1996.