Network Processors: A practical approach for achieving wire-speed packet processing in the emerging optical backbone networks

[1]J. Veiga-Gontán, [1]P. Pavón-Mariño, [1]J. García-Haro, [2]M. Rodelgo, [2]C. López-Bravo, [2]F. J. González-Castaño

[1]Department of Information Technologies and Communications, Polytechnic University of Cartagena, Plaza del Hospital 1, Edificio Antiguones, E-30202, Spain.

[2] Department of Telematics Engineering, University of Vigo, Spain

javg@alu.upct.es, {pablo.pavon, joang.haro}@upct.es, {mrodelgo, clbravo, javier}@det.uvigo.es

ABSTRACT

In this paper, the wire-speed packet processing issue is studied for Optical Packet/Burst Switching (OPS/OBS) switching nodes as well as for wavelength-routing (WR) switching nodes. Some assumptions are made to obtain simple estimates of the requirements in terms of forwarding events per second. This is compared with the packet processing performance of electronic devices which implement this type of tasks. Specially, the state-of-the-art in Network Processor (NP) technology is examined. NPs seem to offer the best flexibility/performance trade-off. Our results reveal 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 is considered. Nevertheless, this alternative reduces, but does not eliminate, the packet-processing bottleneck. In the authors' opinion this situation will promote OPS/OBS technologies.

Keywords: network processor, WDM, OPS, OBS, WR, PPL.

1. INTRODUCTION

Optical Wavelength Division Multiplexing (WDM) backbone networks have been deployed to interconnect major traffic centres. The WDM technology provides an enormous transmission bandwidth. The challenge is to find mechanisms capable of efficiently switching this amount of traffic. The alternatives considered are: Optical Packet/Burst Switching (OPS/OBS), Wavelength-Routing (WR), and Optical Electronic Optical (O-E-O) [1].

In this paper, we observe the electronic part of the switching nodes in OPS, OBS and WR networks. Our objective is to address a feasibility study, from the point of view of the electronic control unit of the switch, and the rate of forwarding decisions per second to be taken. We define a forwarding decision event as a traffic arrival to the node which requires scheduling and table lookup. Regardless of the switching technique employed, these decisions in the switching nodes are performed electronically, and will be like that in the short to medium term. In this paper, the requirements for wire-speed packet processing are estimated for each switching paradigm, and compared to the corresponding performance figure of commercial electronic control units.

The rest of the paper is organized as follows. Section 2 investigates the packet processing requirements of the nodes for the three switching paradigms under consideration. Section 3 provides some performance figures for a set of common control unit technologies, with special focus on Network Processors (NP). Finally, section 4 concludes.

2. ESTIMATION OF PACKET PROCESSING REQUIREMENTS

The network traffic is commonly divided into signaling information or control plane, and "user" information or data plane. Control plane traffic is a small fraction of the total traffic. In general, it has no hard time restrictions, and its processing in the node is diverse and potentially complex. Switching systems usually devote general purpose processors for the processing of the control plane traffic. This is called the switch slow-path. Data plane traffic is massive, and requires fast and simple processing based on packet header information. Backbone switching systems usually allocate specific systems for data plane traffic processing. This is called the switch fast-path. Our study is focused on the switch fast-path.

2.1 Packet processing estimation

In order to estimate the decision throughput requirements, we focus on an switching node example, with 50 input WDM channels, and 50 output channels. The number of input and output fibre links in which these channels are distributed is not relevant for our study. We suppose 60% utilization in each WDM channel. We

---

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.

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

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

- *OPS*. In OPS and OBS nodes, only the header is processed electronically. A forwarding decision is required for each arriving packet/burst, regardless of the packet/burst payload duration, and the payload data rate. Considering an average packet duration of 1 μs (as proposed in [2]), and an average load of the 60%, in average 30 packet arrivals occur each microsecond. This yields to $3x10^7$ decisions per second (packets per second, pps). Note that these decision rates do NOT depend on the transmission rate employed for 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 decision rates ranging from $3x10^6$ to $3x10^3$ decisions per second. As in the OPS case, these decision rates are independent from the transmission binary rate inside the burst payload.
- *WR/O-E-O*. The 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 (lightpath switching). The groomed traffic is defined as the traversing traffic which is not optically switched. Note that groomed traffic is processed electronically at a forwarding rate which depends on the binary rate in the input ports, and the average packet length. We suppose an average packet size of 125 bytes. The actual fraction of optical switching 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.

Table 1 summarizes the average requirements in number of forwarding decision events per second in each of the above scenarios. Regarding the WR 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 transmission binary rate. As binary rates increase, harder constraints appear. OPS/OBS switching paradigms offer the minor performance requirements and the most important: transmission rate and processing rate issues are decoupled.

Table 1. Average packet processing throughput required for different switching paradigms at 10 Gbps and 100 Gbps transmission rates.

PACKET SIZE	LINKS 10 Gbps	LINKS 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

Table 2. Performance values for commercial Network Processors. Information supplied by manufacturers. (*) Estimated values.

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
Xelerated	X11	40 Gbps / $60x10^6$ pps *
Intel	IXP2855	10 Gbps / $15x10^6$ pps *
Intel	IXP2400	4 Gbps / $6x10^6$ pps *
IBM	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 *

3. IMPLEMENTING THE SWITCHING SYSTEM

In this section we address the following question: is the packet processing throughput achieved today enough to be used in optical switching nodes?

3.1 Analysis of existing technologies

Data plane processing in the fast-path, starts by storing packet header information in high speed access memories. Then, a forwarding decision is to be taken, based on one or more table/s lookup/s, attending to QoS guarantees. The lookup process is speeded up using Ternary Context Addressable Memories (TCAM). Several technological options exist to implement the mentioned tasks:

– *General purpose processor*. 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 programming 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 rather low. As an example, the Motorola ATCA-F300, based on MPC8245 PowerPC publishes a throughput of 1Gbps, (aprox. $1.5x10^6$ packets per second) [3].
– *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 [4]. A long development cycle implies to forecast what features will need to be supported when the product is in the market. Therefore, this approach is the less flexible one. As an example, IBM research has designed a 4-Terabit packet switch using CMOS ASIC technology [5].
– *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 resulting in the impossibility to achieve the processing speed of an ASIC. As a performance example, laboratory experiments have achieved up to 10Gbps with FTP connections using a FPGA [6].

  The latest proposals employ ASICs with FPGAs. The idea behind is to use ASIC for the time critical packet processing, 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. In addition, 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. 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 using specialized C-like languages. The packets per second performance varies among different vendors and models. As an example, the Bay Microsystems Chesapeake NP model documentation [7] publishes a performance figure of $122x10^6$ packets per second, suitable for 40 Gbps interfaces.

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 [8] 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 [9] optical processing is also proposed to implement the delay assignment in OPS switches. Unfortunately these alternatives, although encouraging, are still in a very preliminary stage, and are out of the scope of this paper.

3.2 Comparison of Network Processors

Table 2 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.

3.3 Product development based on NPs

Network Processors usually integrate multiple processing units such as (i) a general purpose CPU core, (ii) a set of simpler micro-engines, and (iii) dedicated hardware for computer-intensive tasks i.e. header parsing, table look-up and encryption/decryption. The system manages a fast access to different types of memory units (SRAM, DRAM,CAM, TCAM,...), and I/O interfaces.

Following a system-on-a-chip (SoC) design method, the product is optimized to perform packet processing tasks at wire-speed. The general purpose CPU core, if any, is usually devoted to the slow path. The set of faster and simpler micro-engines are dedicated to the fast path. Commonly, some of the cores include hardware support for multi-threading, which essentially results in zero context-switch time between threads on the same core. Multi-core design allows processing several packets at a time and multi-thread avoids a micro-engine to be idle, while waiting for the completion of i.e. memory access instructions.

Handling such a multi-core and multi-thread specific system, can be a complex task for programmers. Assembler languages in the microengines are system specific. Usually, NP manufacturers distribute their products with a compiler for a specialized C-like higher level language, and a set of libraries implementing most common packet processing functionalities at different layers (SDH, PPP, IPv4, IPv6, MPLS, etc.). Other tools to speed up the development process are NP simulators and debuggers.

Nevertheless programmers still have to fight with the complex hardware architecture of NPs. Nowadays, innovative proposals to facilitate the development are revolutionizing the NP world. These proposals have the

goal of isolating the programmer of the complexity of NP hardware architecture. Two main approaches are highlighted. The first one is headed by IP Fabrics [10] and its virtualization software called Packet Processing Language (PPL) [10]. PPL defines a type of virtual machine which is executed on the NP system. PPL offers a non-hardware dependent specific language to program the packet processing, based on rules, events, and policies. A rule lists one or more conditions under which a set of specified actions are performed. An event causes a designated set of rules to be processed. The typical event is the arrival of a packet. A policy is a complex function, which often manages an internal state that must be tracked from packet to packet.

The second approach is based on newer hardware designs like the Xelerated NPs [11]. The x10q and x11 products integrate 200 and 800 engines respectively, running in parallel. Thanks to the 100% deterministic nature of the data flow architecture, programmers avoid to manage complex tasks such as processor load balancing, synchronization and accesses to common resources. The programmer simply writes linear code sequences with no loops, no context switches and no temporal dependencies. Furthermore, the synchronous operation of the architecture and the absence of shared resources bring major productivity gains in development, testing and optimization tasks.

The simplification in the programming effort obtained with the above two options implies an additional cost. In the case of IP Fabrics, the execution of the PPL virtual machine reduces the system performance. In the case of Xelerated alternative, an enormous increase in the number of engines is required to achieve the same processing capacity as other NP architectures.

4. CONCLUSIONS

In this paper, the wire-speed packet processing issue is examined 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 units devoted to this type of processing. Among the technological options that can be considered, this paper highlights NP based-units, as the most promising solution. NPs are programmable, and provide the required flexibility to shorten the time-to-market cycle.

The comparison between the processing requirements and the processing performance available, show that packet processing speed is not a challenge in OPS and OBS nodes. Furthermore, it will not be a challenge in the future: the expected increase in the transmission binary rates does not affect the forwarding decision requirements. 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 the 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

[1] T. S. El-Bawab, "Optical Switching", Ed., New York: Springer, 1st edition April 5, 2006.
[2] L. Dittman, 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.
[3] www.motorola.com, last consulted on Feb. 4, 2007.
[4] 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.
[5] F. Abel, et al., "A Four-Terabit Packet Switch Supporting Long Round-Trip Times", IEEE Micro Magazine, vol. 23, no. 1, pp. 10–24, Jan./Feb. 2003.
[6] H. Shrikumar, "40Gbps de-layered silicon protocol engine for TCP record", in Proc. IEEE European Design and Automation Association (EDAA), pp. 188–193, Munich (Germany), 2006.
[7] www.baymicrosystems.com, last consulted on Feb. 4, 2007.
[8] F. Ramos, et al., "IST-LASAGNE: Towards all-optical label swapping employing optical logic gates and optical flip-flops," J. Lightw. Technol., vol. 23, no. 10, pp. 2993–3011, Oct. 2005.
[9] M. Murata and K. Kitayama, "Ultrafast photonic label switch for asynchronous packets of variable length," in Proc. IEEE Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2002), pp. 371- 380 vol.1, New York, June 23–27, 2002.
[10] http://www.ipfabrics.com, last consulted on Feb. 4, 2007.
[11] http://www.xelerated.com, last consulted on Feb. 4, 2007.