

A Control and Management Plane for Large Packet Switches

Andrea Bianco, Jorge M. Finochietto *, Guido Gavilanes and Fabio Neri

Dipartimento di Elettronica, Politecnico di Torino
C.so Duca degli Abruzzi 24, 10129, Torino, Italy
Email: {lastname}@polito.it

**CONICET - Universidad Nacional de Cordoba,*
Av. Velez Sarsfield 1611, 5000, Cordoba, Argentina

*jorge.finochietto@ieee.org

Abstract—In this paper we address the problem of designing a control and management plane for large packet switches. In particular, we consider the case where the forwarding plane is distributed among several line cards interconnected through a passive optical fabric. The main contribution of the paper is the discussion of issues arising when designing a unique and consistent control and management plane, and the proposal of a solution for the considered architecture. Finally, we consider the case of building an Ethernet switch, discuss implementation details and describe the experimental setup used to validate the proposed approach.

I. INTRODUCTION

The introduction of new bandwidth-hungry services and the continuous increase of broadband subscribers is driving the exponential growth of Internet traffic and imposing more complexity on packet switches, such as IP routers and Ethernet switches.

Several carriers cope with this growth by simply adding to their existing infrastructure more packet switches interconnected by high-speed ports. This results in a significant increase in the port count, as the number of interconnection interfaces typically increases exponentially with the number of network elements. As a result, in the long-term, this strategy may lead to high capital expenditures (CAPEX) costs. Also operational expenditures (OPEX) costs can become high since more network elements imply an increased complexity in the control and management planes. Indeed, network elements need to exchange and handle more routing information and typically require more advanced configuration from the operator (e.g., configuration of hierarchical routing, different routing protocols and policies). Besides, the more network elements need to be managed, the more manpower is required. Finally, more devices are subject to failures, which means that repair and maintenance costs can increase, and that cost and complexity of fault protection strategies grows significantly.

An alternative approach is the use of high-end packet switches equipped with many high bit rate ports, which can offer large forwarding capacity (currently above Tb/s) and reduce the number of network elements [1]. However, each new generation of packet switches requires more power than the previous one, and it is more difficult to package it in a single rack. Thus, high-end switches often comprise

several racks: one or more racks host the electronic switching fabric and the control logic, while others racks host the line cards. In this multi-rack configuration, optical links are being increasingly used to interconnect the fabric and the line cards.

These solutions occupy valuable space, consume too much power, and pose reliability concerns due to the large number of active components in the switching fabric. Large packet switches with an internal passive optical switching fabric can scale better to high capacities. If all buffering and switching capabilities are implemented electronically inside line cards, no active switching elements are present on the fabric, thus increasing device reliability. In addition, the number of electronic transceivers can be reduced with respect to the case of an electronic switching fabric where optical links are used between each line card and the fabric; therefore, both footprint and power consumption can be significantly reduced.

Typically, switching decisions at line cards are controlled by centralized electronic arbitration schemes, which suffer from limited scalability. The performance of a packet switch can be upper bounded by the complexity of implementing these centralized arbitration schemes as the aggregate packet processing rate increases. As a result, the availability of distributed arbitration schemes, that can be implemented on line cards interconnected through an optical fabric, creates a distributed forwarding plane with optimal scalability properties that can be used in the design of large packet switches. On the other hand, a distributed forwarding plane introduces more complexity in the control and management plane. Indeed, forwarding/routing information needs to be distributed to all line cards in a consistent way, while control and management protocols must be implemented so that the whole system behaves as a single entity. In recent years, standardization efforts have addressed these issues [2], [3] as commercial core routers are based on distributed multi-rack solutions [4], [5].

In this paper we consider the design and implementation of the control and management plane for a packet switch proposed in the Italian project OSATE, which makes use of a passive optical fabric and of line cards that implement (in the electronic domain) a high performance distributed arbitration scheme. The main contribution of this paper is the discussion of the main issues arising on the design of the control and man-

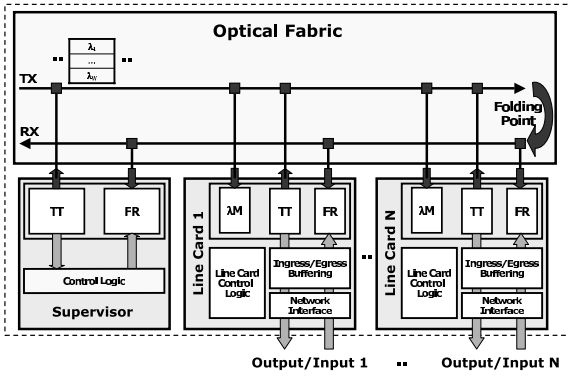


Fig. 1. Optical Fabric Architecture

agement plane, and the proposal of a control and management plane suitable for the considered architecture. In particular, we address and experimentally demonstrate the implementation of an Ethernet switch. The paper is organized as follows. Section II describes the forwarding plane of the packet switch, Section III discusses issues related to control and management plane, and Section IV describes the experimental work done to demonstrate the proposed approach. Finally, Section V concludes the paper.

II. THE FORWARDING PLANE

The packet switch architecture consists of several line cards interconnected by a passive optical fabric. Each line card is equipped with one optical interface attached to the fabric that is responsible for sending/receiving packets to/from other line cards through the optical fabric. Besides, each line card is equipped with at least one network interface that acts as a packet switch port where other network elements can be connected. In this section, we mainly discuss the architecture of the optical fabric and its interfaces, and describe the distributed arbitration scheme used to schedule packets.

A. Optical Fabric

The considered optical fabric [8] is based upon a passive WDM (Wavelength Division Multiplexing) all-optical data path over a folded bus, as depicted in Fig. 1. The folded bus conveys W wavelengths which first traverse the transmission (TX) bus and, after a folding point, the reception (RX) bus. The optical fabric interconnects N line cards and one supervisor card (that will be discussed in Section III). Each line card is equipped with one optical interface made of one transmitter and one receiver operating at the data rate of a single WDM channel. The optical input and output ports of the line card are passively coupled to the TX bus and RX bus respectively. Since full connectivity between all available line cards must be provided on a packet-by-packet basis, fast wavelength tunability at transceivers is required to temporally allocate all-optical single-hop bandwidth. However, due to cost tunability of transceivers, this is limited only to transmitters, while receivers are permanently tuned to a specific WDM channel. When a single receiver per WDM channel is present,

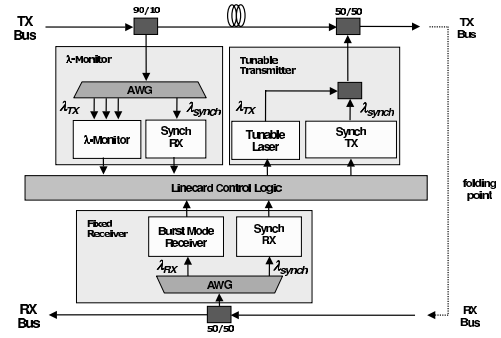


Fig. 2. Optical Interface Architecture

and thus the number of available WDM channels W equals the number of line cards N , the architecture can be shown to be equivalent to a distributed crossbar switch, which is able to connect at every time up to N disjoint input-output pairs.

The architecture is synchronous, with time slotted operation. For this purpose, one additional wavelength of the WDM comb is dedicated to the distribution of synchronization information (bit frequency and slot phase and frequency). This information is generated by the first transmitter on the folded bus, and broadcasted to all receivers along the data path. Each optical interface, as shown in Fig. 2, is equipped with one tunable transmitter (TT) and one fixed receiver (FR) operating in burst mode. A (fixed wavelength) synchronization transmitter is present at least in the first interface on the bus, and a fixed synchronization receiver drops the signal from the synchronization channel to facilitate clock and data recovery in the burst-mode data receiver. The slotted behavior facilitates the use of distributed scheduling mechanisms, which avoid packet collisions by means of a void-detection mechanism called λ -Monitor (λM), by which optical interfaces know which wavelengths were not used by upstream ones in each time slot. Priority is given to in-transit traffic, thus without requiring any packet buffering for in-transit data, as well as no packet switching or stripping in the optical domain. The TT is used to insert packets on free time slots on the wavelength leading to the packet's destination. The FR receives all packets on its assigned wavelength. Since each line cards always receive packets on the same wavelength, thus in a non-overlapping way, receiver contention does not occur, since it is solved at the transmitter side. To avoid Head-of-the-Line (HoL) blocking [9], Virtual Output Queuing (VOQ) is envisioned at inputs: optical interfaces queue packets to be inserted on a per destination-wavelength basis. Since neither packet collisions nor receiver contention occur, transmitted packets are never lost except for transmission errors in the physical layer.

B. Distributed Arbitration Scheme

The folded bus topology imposes line cards to transmit packets to other line cards in a given slot/wavelength sequentially and not in parallel, as in traditional crossbars.

However, a simple empty slot arbitration scheme might lead to fairness problems due to access probabilities depending on the position of the optical interface along the TX bus. Referring to Fig. 1, an upstream line cards can "flood" a given wavelength, reducing (or even blocking) the transmission opportunities of downstream ones competing for access to that channel, thus leading to significant fairness problems. Therefore, suitable scheduling algorithms are required to arbitrate packet transmissions to ensure not only high throughput and bounded delay, but also equal transmission probabilities for all line cards, even when some inputs are heavily loaded.

Arbitration schemes can be either centralized or distributed. The former require the use of an electronic scheduler that, after receiving status information from line cards, defines a new input/output permutation, i.e., input/output port connection pattern, for each time slot. The latter uses only locally available information on interfaces to determine which packet to transfer. Thus, centralized schemes can potentially offer better performance in terms of throughput and latency than distributed ones. However, the electronic complexity of the scheduler implementation and the need for additional signaling bandwidth must be taken into account. Indeed, optimal algorithms such as Maximum Weight Matching (MWM) [10] are impractical because of their complexity, and sub-optimal ones, such as iSLIP [11], are in general preferred.

In this context, the performance of a distributed scheduling scheme becomes a crucial issue to assess the actual efficiency of the considered architecture. One of these schemes is dubbed Fasnet [12], and it was previously proposed and studied to arbitrate access [13]. As a first approximation, the Fasnet scheme behaves as a distributed polling mechanism without the need of a centralized scheduler. Fasnet operates cyclically, and each cycle is associated with a chained transmission of packets by all line cards, named train. A train is composed by a first control packet, dubbed locomotive, transmitted by the first line card, and by all packets transmitted sequentially after the locomotive by the remaining ones. The first line card starts a new cycle, thus transmits a new locomotive, every time it detects the end of a returning train (i.e., an empty slot on the RX bus). Note that the first line card can have its λ -monitor passively coupled to the RX bus instead of the TX one, as this line card always senses empty slots on the TX bus. To limit the maximum number of packets each line cards can transmit on each cycle, each one is assigned a quota Q . When a line card senses the end of a train, i.e., an empty slot on the TX bus, it seizes the channel for a number of packets equal to the minimum between the quota Q and the number of packets in its queue for that channel. Once a line card releases the channel (either by exhausted quota or empty queue), it restores its quota and waits for the next train before attempting to access the channel again. In a WDM multi-channel network, the Fasnet behavior can be easily replicated over all available wavelengths: thus, W trains exist, one for each channel. Performance analysis of the scheme and discussion on the scalability of the optical fabric is addressed in [14].

III. THE CONTROL AND MANAGEMENT PLANE

Despite the distributed nature of the forwarding plane, end users, other network elements, and the operator itself must see the packet switch as a unique entity. As a result, the control and management plane must hide the distributed nature of the forwarding plane to end users. For this purpose, a proper signaling protocol should be used internally to the switch to enable all line cards to exchange control and management information and keep data consistency. In the proposed architecture, it is possible to envision two ways of implementing the signaling protocol: out-of-band and in-band. The former implies that a dedicated channel (wavelength) must be available to exchange signaling messages and that line cards must be equipped with dedicated transceivers. The latter instead makes use of the same channels (wavelengths) and transceivers that line cards use to forward packets. Although out-of-band signaling can offer low latency communication, it increases the complexity (hence the cost) of line cards.

The control and management plane of the switch must implement several protocols to exchange information with other network elements and with end users. In the case of IP routers, control protocols include Routing Information Protocol (RIP), Open Shortest Path Protocol (OSPF), Border Gateway Protocol (BGP), etc; and require routing tables to be consistently available to all line cards. In the case of Ethernet switches, backward address learning must be propagated to all line cards to build a common Media Access Control (MAC) forwarding database; moreover control protocols such as the Spanning Tree Protocol (STP), VLAN Trunking Protocol (VTP), etc. must be supported. For management purposes, protocols such as Telnet and Simple Network Management Protocol (SNMP) must always be implemented.

All these protocols require much complexity to envision a fully distributed implementation on every line card. However, some recent research effort have considered the implementation of some limited functionalities of these protocols in a distributed way [6], [7]. An alternative solution is to introduce a supervisor line card that implements all required protocols and maintain data consistency (see Fig. 1). This solution introduces minimum complexity for line cards that only implement a forwarding mechanism either of data packets to other line cards, or of control and management packets to the supervisor. Besides, the use of a supervisor also facilitates the implementation of the signaling protocol. Indeed, if line cards only exchange signaling messages with the supervisor, a dedicated wavelength can be allocated for signaling, but line cards can still use the same (tunable) transmitter to send both data and signaling packets. On the supervisor side, a dedicated receiver is required for the signaling channel, while a single transmitter implemented as an array of lasers can be used for sending (possibly in broadcast) messages to line cards. Note that this solution represents a hybrid one among the previously described out-of-band and in-band mechanisms, and can significantly reduce the latency of signaling messages by increasing only the complexity of the supervisor.

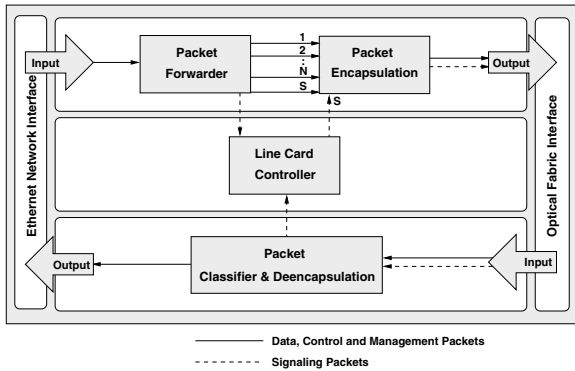


Fig. 3. Line Card Functional Diagram

Since the proposed architecture introduces different access priorities, the position of the supervisor on the optical folded bus can have some impact in the control and management plane performance. Indeed, to guarantee the minimum possible latency in exchanging messages, the supervisor should be positioned before the first line card. In that position, the supervisor can always send packets immediately since all time slots are empty. At this point, it is also possible to move the generation of locomotive packets for the distributed arbitration schemes to the supervisor. Note that both unicast and broadcast packets can be sent if the transmitter is built using an array of lasers. Broadcast messages can be used to update/ask control information (e. g., routing/forwarding tables, statistics) on all line cards, while unicast messages can be used to transport both control and management protocols implemented on the supervisor.

IV. ETHERNET SWITCH PROTOTYPE

In this section we describe implementation details of the control and management plane when considering the described architecture to build an Ethernet switch. In particular, we discuss the implementation of the backward address learning mechanism on line cards, and the support of control (STP, VTP) and management (Telnet, SMTP) protocols on the supervisor. The main design goal is to introduce the minimum complexity on line cards that enables the implementation of a control and management plane based on the principles sketched on the previous section. By minimizing the line card complexity, we claim that the scalability properties of the proposed architecture can be maintained.

A. Design Considerations

The line card architecture is shown in Fig. 3. It is equipped with two interfaces: an Ethernet interface that can be connected to standard Ethernet devices, and an optical interface, attached to the optical fabric, that implements the distributed access scheme as described in Section II. The behavior of an Ethernet switch states that incoming packets must be forwarded to the corresponding output interface. In our architecture this implies that incoming packets to a line card must be forwarded to the corresponding one through the optical fabric. For this purpose, packets first traverse a Packet Forwarder module that decides to which line card the packet must be forwarded. As the Packet Forwarder stores a MAC forwarding table

containing Ethernet MAC addresses and the line card where each address is reachable, it reads the destination address of incoming packets and decides its destination line card. If the destination address is not available on the MAC forwarding table, then the packet is sent in broadcast to all line cards. Besides the destination address, the source address is also read by the Packet Forwarder module for backward address learning. Indeed, unknown host addresses are dynamically learnt as packets traverse the line card.

To forward packets to others line cards through the optical fabric, each packet is first encapsulated over a “transport” packet that includes a checksum field and the address of the optical interface of the line card where the packet is forwarded to. Note that an internal addressing scheme is required for the optical interfaces since in a general configuration more than one (fixed) receiver could be tuned to the same wavelength (sharing the wavelength capacity). As a result, packets are encapsulated either with line cards addresses ($1, \dots, N$) or the supervisor address (S) (see Fig. 3). After the packet has been encapsulated, the tunable transmitter (TT) sends it over the proper destination wavelength. Finally, outgoing packets are de-encapsulated before leaving the line card to preserve the original format of the packet.

As discussed in Section III, a signaling protocol is needed to let line cards exchange information with the supervisor, and vice versa. For this purpose, a Controller module is added to the line card. This module is able to generate (receive) signaling messages to (from) the supervisor and to read (write) data from (to) the different modules available on the line card. For example, if the supervisor request information about packet counters, then the line card Controller must first receive and process this message, read the information from the counters, and generate a response message to the supervisor. Besides, the Controller can receive a message from the Packet Forwarder module each time a new MAC address is learnt. In this case, the Controller forwards the message to the supervisor that broadcasts the message to all line cards to update the MAC forwarding table with the new address. As a result, the supervisor can keep the MAC forwarding table updated and consistent over all line cards.

The supervisor line card is only equipped with the optical interface. Besides the internal address of this interface (used for forwarding purposes), an Ethernet MAC address is required to implement control and management protocols. In the case of an Ethernet switch, the supervisor uses this MAC address as the switch base MAC address when running both the control protocols (STP) and the management ones (SNMP, Telnet). All these protocols are run in a centralized fashion, and by means of a proper signaling protocol the supervisor can interact with line cards if required. As a result, the Packet Forwarder module must send control and management packets to the supervisor. This does not require any additional complexity, since control and management packet forwarding simply exploits the MAC address of the supervisor. In other words, sending packets to the supervisor is a special case of the general forwarding mechanism already available on line cards.

B. Experimental Demonstration

To demonstrate the feasibility of the proposed control and management plane, a proof-of-concept implementation has been built and tested. The prototyped Ethernet switch consists of three line cards and one supervisor. To simplify the implementation, the interconnection of the line cards and the supervisor has been done using an Ethernet network instead of the optical fabric. Note that actually the implementation of the control and management plane are independent of the nature of the interconnection fabric. As a result, each line card was built using a PC equipped with two Ethernet interfaces, while the supervisor used just one. Line cards use one interface as the switch port while the other one is used for interconnection purposes.

The required processing logic on line cards was implemented using available Click [15] modules when possible and developing new ones when needed. The Click framework allowed to build and interconnect packet processing modules that perform all the previously described functionalities. On the supervisor card, the use of Click was limited to low level functionalities such as packet classification and new addresses broadcasting (backward learning), while the actual implementation of control and management protocols was left to the Linux operating system. Indeed, control and management packets arriving to the supervisor are forwarded to the operating system after a classification done by a Click module.

Finally, to validate the Ethernet switch prototype, a traffic generator was used to run several functional tests and verify the proper behavior of the switch. All test results showed that the proposed approach is feasible. As a result, we envision in the future to integrate the control and management plane on the actual switch prototype of the OSATE project that implements line cards on FPGA boards equipped with one Gigabit Ethernet interface and one optical fabric interface.

V. CONCLUSION

Scaling switching capacity may take two main paths: increasing the number of “small” switching devices, or building “large” switches. The first approach is less demanding in terms of technology, but requires more components and more complex control and management procedures at the network level. The second approach is much more challenging in switch technological design, and may suffer from the limited scalability of current centralized resource allocation and contention resolution procedures internally to the switch. Distributed control of resource allocation in “large” switches equipped with intelligent line cards improves switch scalability, but requires extra effort to guarantee consistency of packet filtering and forwarding information, and to handle in a switch-wide consistent manner control and management information. It is indeed important, to guarantee good network scalability properties, that the switch, albeit internally realized in a distributed fashion, i.e., similarly to an interconnection of “small” switches, externally appears to other network elements and to end users as a unique entity.

In this paper, we described how these ideas were instantiated on a specific switching architecture based upon a passive optical interconnection of line cards. Resource allocation (i.e., switch control for packet forwarding) is fully distributed across intelligent line cards for better scalability, while control plane and management plane protocols are handled in a centralized manner (through internal signaling) by a supervisor. By so doing, the bottleneck of a centralized packet-by-packet switch control is avoided, and the (internally distributed) switch appears externally to other network elements as a single entity, thereby avoiding scalability problems at the network level.

Our design choices lead to good scalability properties (currently being addressed by our research group in a quantitative way), and enable the effective realization of multi-terabit-per-second packet switches. The proposed architecture is being prototyped in the PhotonLab and LIPAR experimental facilities of Politecnico di Torino in the framework of the Italian research project OSATE.

ACKNOWLEDGMENT

The authors would like to thank the support of the Italian project OSATE and the European Network of Excellence e-Photon/ONe.

REFERENCES

- [1] H. J. Chao, “Next Generation Routers [Invited]”, Proceedings of the IEEE, Vol.90, Iss.9, Pages: 1518- 1558, Sep 2002
- [2] L. Yang, et al., ForCES Forwarding Element Model, IETF - Network Working Group, July, 2006.
- [3] A. Doria, et al., ForCES Protocol Specification, IETF - Network Working Group, December, 2005.
- [4] Cisco, “CRS-1 Carrier Routing System”, White paper, October 2006
- [5] Juniper, “T-series Routing Platforms: T320, T640, T1600, and TX Matrix”, Datasheet, June 2007
- [6] M. Deval; H. Khosravu; R. Muralidhar; S. Ahmed; S. Bakshi; R. Yavatkar, “Distributed Control Plane Architecture for Network Elements”, Intel Technology Journal, Vol. 7, Iss. 4, Pages 51-63, 2003.
- [7] K. Nguyen; H. Mahkour; B. Jaumard; C. Assi; M. Lanoue, “Toward a Distributed Control Plane Architecture for Next Generation Routers”, Universal Multiservice Networks, 2007. ECUMN '07. Fourth European Conference on, Pages:173-182, Feb. 2007
- [8] A. Carena; V. De Feo; J. M. Finochietto; R. Gaudino; F. Neri, C. Pignone, P. Poggiolini, ‘RingO: An Experimental WDM Optical Packet Network for Metro Applications,’ IEEE Journal on Selected Areas in Communications, vol. 22, no. 8, , pp. 1561-1571, Oct. 2004
- [9] M. Karol, K. Eng, H. Obara, “Improving the performance of input-queued ATM packets switches”, IEEE Infocom 92, May 1992, pp. 110-115, Firenze, Italy
- [10] N. McKeown, et al.; “Achieving 100% throughput in an input-queued switch”, IEEE Transactions on Communications, Vol. 47, No. 8, pp. 1260-1267, Aug. 1999.
- [11] N. McKeown, “The iSLIP Scheduling Algorithm for Input-Queued Switches”, IEEE/ACM Transactions on Networking, Vol. 7, No. 2, April 1999
- [12] J. O. Limb and C. Flores “Description of Fasnet – A Unidirectional Local–Area Communication Network”, *The Bell System Technical Journal*, Vol. 61, No. 7, September 1982.
- [13] A. Bianco, D. Cuda, J. M. Finochietto, F. Neri, C. Pignone, “Multi-Fasnet Protocol: Short-Term Fairness Control in WDM Slotted MANs”, IEEE Globecom 2006, San Francisco, CA, USA, 27-30 November 2006
- [14] A. Bianco, E. Carta, D. Cuda, J. M. Finochietto, F. Neri, “An Optical Interconnection Architecture for Large Packet Switches”, International Conference on Transparent Optical Networks (ICTON 2007), Rome, Italy, June 2007
- [15] E. Kohler, R. Morris, B. Chen, J. Jannotti, M. Kaashoek, “The click modular router”. ACM Trans. Comput. Syst. 18, 3, Pages 263-297, August 2000.