

Improving Energy Efficiency in IEEE 802.3ba High-Rate Ethernet Optical Links

P. Reviriego, *Member, IEEE*, B. Huiszoon, *Member, IEEE*, V. López, *Member, IEEE*, R. B. Coenen, J. A. Hernández, and J. A. Maestro, *Member, IEEE*

Abstract—Energy efficiency in communications, and particularly Ethernet, is becoming an important research area. The IEEE 802.3az energy-efficient Ethernet standard has already proposed a mechanism to reduce energy consumption on copper-based physical layer devices. This mechanism exploits the well-known fact that end-user links are usually lightly loaded, and that important energy savings can be achieved, if the device is put into a low-power sleep mode whenever no frames are pending for transmission. This study pursues a two-goal purpose: firstly, it aims to evaluate whether such an active/sleep dual-state mode is suitable for high-rate optical Ethernet links as defined in the upcoming IEEE 802.3ba amendment that specifies the lower layers of 40- and 100-Gigabit Ethernet. Secondly it proposes to use an algorithm that dynamically exploits the multilane architecture of the high-speed optical transport layer. As it is studied throughout the paper, the two-state active/sleep-based mechanism may not achieve energy savings, while a load-based dynamic lane management enables energy reductions for a wide range of input traffic loads. A commercially available 100GBASE-SR10 module is used as case study for the analysis.

Index Terms—Networks, network interfaces, optical communication equipment, optical fiber communication, reconfigurable architectures.

I. INTRODUCTION

AT PRESENT, Ethernet is the most widely deployed technology on wired LANs, apart from an already large installed base, with over a hundred million devices shipped every year [1]. Ethernet supports a wide variety of physical media ranging from unshielded twisted pairs (UTP) commonly used in office buildings to optical fibers used in high-performance applications and backplanes to provide connectivity among different

modules of a large system. It also supports different link speeds, which are currently standardized at bitrates from 10 Mb/s going up to 10 Gb/s per order-of-magnitude. That traditional ten-fold increment is continued with the upcoming IEEE 802.3ba amendment; however, an intermediate line rate of 40 Gb/s is also supported, next to the expected 100 Gb/s [2]. The 100-Gigabit Ethernet (GbE) line rate targets network aggregation applications, while 40-GbE targets computer and server applications. It has become apparent that both segments show different growth rates, bandwidth requirements, market potential, and cost/performance balance, and therefore, the necessity for two bitrates is justified in order to provide a significant bandwidth while maintaining maximum compatibility with the installed base of 802.3 interfaces and the earlier investment made in research and development [3].

The evolution toward higher bitrates comes with a cost in terms of increasing circuit complexity which, intuitively, impacts the total energy consumption of the system. In general, the power usage tends to increase as the line rate grows. A recent publication by The Climate Group reported the carbon footprint of the whole information and communication technology (ICT) sector, as divided into PCs and Peripherals, Telecoms Networks and Devices, and Data Centers, to be 2% of the estimated total emissions by human activity in 2007 [4]. This number could grow to 6% in 2020 covering both the embodied carbon and the footprint from use. Given the widespread adoption of Ethernet in the ICT sector, significant energy is readily consumed by Ethernet modules worldwide. Hence, if the energy consumption per operational Ethernet device is reduced, regardless how small the amount, large aggregated energy savings are obtained [5]. This reduction is possible, since most existing Ethernet physical layers (PHYs) continuously transmit signals on the link to keep the receivers synchronized and adapted to channel conditions, even if there is no data to send. Thus, Ethernet devices typically consume the same amount of energy, no matter how much traffic load they carry. It is clear that this mechanism can be improved. Another issue is that the power consumption of Ethernet PHYs increases with speed. This is due to increased transceiver complexity with each new speed. This results in a larger energy consumption when higher speeds are used [5].

The IEEE 802.3az energy-efficient Ethernet (EEE) standard aims to make the consumption of energy over a link proportional to the amount of traffic exchanged [6]. To this end, EEE defines a low-power mode referred to as low-power idle (LPI) mode. The PHY is put into this sleep mode when no frames are pending for transmission and awakes very quickly upon data arrival without changing the line rate. The LPI mode stops

Manuscript received March 1, 2010; revised April 16, 2010; accepted April 28, 2010. Date of publication June 27, 2010; date of current version April 6, 2011. This work was supported by BONE (Building the Future Optical Network in Europe), a Network of Excellence funded by the European Commission through the 7th ICT-Framework Programme, by a research award from Google, and by the “Juan de la Cierva” postdoctoral research grant from The Spanish Ministry Ministerio de Ciencia e Innovación (MICINN).

P. Reviriego and J. A. Maestro are with the Universidad Antonio de Nebrija, Madrid 28040, Spain (e-mail: previrie@nebrija.es; jmaestro@nebrija.es).

B. Huiszoon and V. López are with the High Performance Computing and Networking Group, Universidad Autónoma de Madrid, Madrid 28049, Spain (e-mail: bas.huiszoon@uam.es; victor.lopez@uam.es).

R. B. Coenen is with Reflex Photonics Inc., Sunnyvale, CA 94085 USA (e-mail: rcoenen@reflexphotonics.com).

J. A. Hernández is with the Universidad Carlos III de Madrid, Madrid 28911, Spain (e-mail: jahgutie@it.uc3m.es).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/JSTQE.2010.2050136

transmission and enables elements in the receiver to be frozen. When new frames arrive, the link is activated within a few microseconds. Such sleep/active operation requires minor changes to the receiver elements, since the channel is quite stable. Additionally, periodic refresh intervals are scheduled in such sleep mode to keep the receivers synchronized and adapted to channel conditions at all times. The EEE standard currently focuses on the most common Ethernet PHYs, covering those that use UTP as a physical media, namely, 10GBASE-TX, 100BASE-T, and 10GBASE-T. Also backplane PHYs, the XG media-independent interface (MII) extended sublayer and 10-Gb attachment unit interface are included in the standard. It is worth mentioning that both the IEEE 802.3az and IEEE 802.3ba study groups have readily started exchanging information and the identification of key considerations to improve energy efficiency in 40 and 100 GbE [7].

In EEE, significant energy consumption is only attributed to the active mode and while making a transition between the active and the sleep modes. The usage is significantly reduced when the transmitter is in the sleep mode and savings up to about 90% can be achieved. Accordingly, significant savings will be achieved for links that remain in the sleep mode most of the time. However, the transition times between the two power modes are in the order of microseconds, and therefore, comparable or even larger than frame transmission times. In [8], an initial evaluation of EEE is presented showing that it achieves significantly lower energy consumption in very lightly loaded links, i.e., significantly less than one percent, while the savings are smaller for links with a load above a few percent of the link capacity. The first situation is common in links that connect desktop computers, while the second are typically encountered in server networks. Finally, it was concluded that mode transitions could result in a large overhead when EEE is introduced [8].

As discussed earlier, the higher Ethernet line rates will most likely infer a higher energy consumption, and therefore, it seems interesting to explore options to improve the energy efficiency of this new generation of Ethernet PHYs. Among the PHYs specified in 802.3ba, several options for an optical transport layer are proposed. One common denominator is that it has been decided to use multiple lanes, which operate at a lower baud rate instead of having a single serial connection at the line rate. Such consensus was reached because currently low-cost serial architectures at 40 and 100 Gb/s are not yet available, and therefore, existing 10-Gb/s technology can be leveraged [9]. However, opting for multiple lanes may in turn slow down the development and deployment of serial architectures. To that respect, it was recently announced that a separate study group (IEEE 802.3bg) is established to design the 40 Gb/s serial physical medium dependent (PMD) sublayer in order not to slow down progress on the IEEE 802.3ba Standard, which is to be delivered in mid-2010.

In this study, the multilane approach in 802.3ba is exploited, for the first time, in order to reduce the energy consumption of the transmission system. Two alternatives for the implementation of energy-efficient mechanisms for the 802.3ba optical PHYs are evaluated on their feasibility and impact via numerical simulations. First, the extension of the approach used in EEE is evaluated showing that it would provide a poor performance.

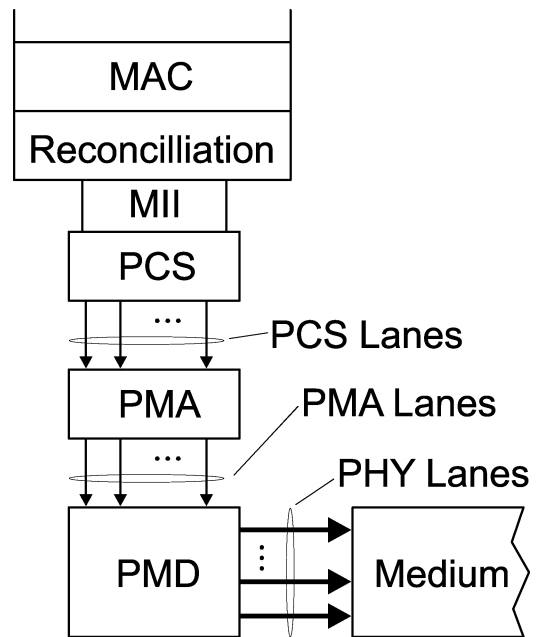


Fig. 1. Generalized IEEE 802.3ba architecture. MAC: medium access control; MII: media-independent interface; PCS: physical-coding sublayer; PMA: physical-medium-attachment sublayer; PMD: physical-medium-dependent sublayer; PHY: physical layer.

Then, a new approach is presented that dynamically activates the lanes of 802.3ba PHYs depending on the experienced load. This alternative is analyzed using real traffic traces. Measured data on the power consumption of a commercially available 100-GbE module are presented. In conclusion, load-based multilane handling is a straightforward technique to potentially provide energy savings in a wide range of 802.3ba-based networking scenarios.

The remaining of this paper is organized as follows: Section II provides a brief introduction to the IEEE 802.3ba architecture. In Section III, the use of a two-mode approach as the one used in EEE is evaluated in case of 40 Gb/s and 100 Gb/s. Then in Section IV, the proposed adaptive multilane approach is described and its performance is evaluated in Section V. Finally, the conclusions are presented in Section VI.

II. IEEE 802.3BA 40-GB/S AND 100-GB/S ETHERNET

A. Overview

A generalized overview of the IEEE 802.3ba architecture is shown in Fig. 1. The 40-Gb/s and 100-Gb/s PHY implementations are generally referred to as 40GBASE-R and 100GBASE-R [9]. The MII provides the logical interconnection between the medium access control (MAC) and the PHY sublayers, and either an XLGMII or the CGMII is present in Fig. 1 for 40GBASE-R or 100GBASE-R, respectively. A downstream implementation is considered in the following to explain the principle of operation of the physical-coding sublayer (PCS), physical medium attachment (PMA), and PMD sublayers. Generally speaking, 40GBASE-R and 100GBASE-R only differ in the number of lanes and the lane baud rate.

The PCS handles the data coming from the MAC by encoding eight data octets into 66-bit 64-B/66-B blocks. These are then distributed toward the PMA via a round-robin principle [10], that is, a first 64-B/66-B block is sent to PCS lane 1, a second to lane 2, etc., and over again starting with lane 1. At the receiver side, the blocks may be received in any order, and skew is resolved by using lane markers. The 40GBASE-R PCS has four lanes at a baud rate of 10.3125 Gb/s, while the 100GBASE-R PCS has 20 lanes at 5.15625 Gb/s.

The PMA layer allows the PCS to be interconnected in a media independent way with a range of PHYs. It can be implemented in one or more segments to allow for different configurations at the PMD. However, the 802.3ba Standard explicitly mentions that the number of input and output lanes of each PMA segment should always be a divisor of the number of PCS lanes (N_{PCS}). Consequently, the number of PMA lanes, and thus the number of PHY lanes, is limited to $N_{PHY} = \{1, 2, 4, 5, 10, 20\}$ with $N_{PHY} \leq N_{PCS}$. In case of the commercially available 100GBASE-SR10 C Form-factor Pluggable (CFP) module considered in this paper, N_{PHY} equals to 10; therefore, for each two ingoing PCS streams, the PMA layer has an outgoing 10.3125 Gb/s stream [11]. In the remaining, this module is referred to as module A, where each lane has its own fiber.

Finally, the PMD maps the incoming PMA data streams onto a matching amount of PHY lanes. The PMD architecture may be implemented in eight different ways depending on the line rate, distance, and transmission medium [10]. Here, the focus is on PMDs with optical PHYs.

B. Selective Lane Activation

All eight different PMDs defined by the 802.3ba amendment have an option termed “PMD lane by lane transmit disable function” which is an optional functionality depending on the implementation of the management data input/output (MDIO) interface. The MDIO specification of the CFP multisource agreement (MSA) provides more information on how such functionality is handled in their hot pluggable optical transceiver modules [12]. Module A does support selective lane activation according to the CFP MSA.

The status of individual PHY lanes can be controlled by adjusting the entries of the “individual network lane TX_DIS control”-register in volatile register 1 (VR1). The latter is read when the CFP module is driven into the “TX-turn-on State” which is a final transition to the “ready state.” Furthermore, it is mentioned in [12] that any lanes that are disabled shall remain disabled after the module enters that state, and that any changes in TX_DISs shall be ignored by the register. In other words, the CFP MSA MDIO specification allows selective lane de-activation, but only when the module initializes and not during the uptime. It can be concluded that, at the time of writing, a *dynamic* lane handling is not supported by the MSA. Draft version D3.2 of the 802.3ba amendment (March 24, 2010) does not specify such support during operation, because it is unambiguously mentioned in Clause 45.2.1.8: “Disabling the transmitter on one or more lanes stops the entire link from carrying data”.

The reader should take note that selective lane activation or de-activation during the uptime of module A is technologically not prohibited. It is, therefore, that in the remaining of this study, it is assumed that the aforementioned flexibility is available at the optical PHY. The incentive of allowing this by the CFP MSA and the 802.3ba amendment is provided in terms of significant improvements in energy efficiency. A first suggestion was recently made by authors in [20].

III. EVALUATION OF EEE APPLIED TO 802.3BA HIGH-SPEED OPTICAL LINKS

The most obvious approach to improve energy efficiency in the 802.3ba PHYs is to extend the mechanism defined in EEE. That is to employ a low-power mode such that when there is no data to transmit, the link is put into that mode. Then, when data arrives, the link is activated again. The effectiveness of this approach is limited by the overhead involved in the transitions between modes, as discussed earlier. For optical PHYs, the transition times are generally much larger than for copper PHYs due to a more complex circuitry to stabilize the channel. Values between 1 and 2 ms are reported in [13] and [14] while transition times well below a millisecond are considered for all copper PHYs in EEE, that is, until a bitrate of 10 Gb/s [6]. As the frame transmission time will be smaller in the high-rate 802.3ba links, the relative overhead of the transitions would be even larger. The CFP MSA MDIO specification defines preliminary maxima of 150 ms for the response speed of the MDIO interface, if a single-state change occurs. Module A has a maximum “power on time” of 100 ms, which involves three transitions, if the device goes from the “reset state” to the “ready state”. In any case, a state transition can be realized much faster on a hardware level. Finally, the expected load for these high-rate links is higher than observed in low-rate links that connect desktop computers. This means that transitions will occur frequently thus wasting a lot of energy. This reasoning is confirmed in the following by the analysis of different trace measurements. The aforementioned discussion assumes that significant energy is consumed during the active/sleep mode transitions in optical PHYs, which is motivated in [8].

To show that the EEE approach would provide poor performance in 802.3ba optical links, a number of traces are now analyzed. The frame interarrivals and lengths are used to compute the amount of time at which the link would stay in the low-power mode, if the EEE approach was used. To do so, it is assumed that as soon as there are no frames to transmit the link is put into low-power mode and when a frame arrives for transmission, it is awoken to send the frame. This minimizes the additional frame delay due to EEE. The energy consumption is then calculated assuming that the energy consumption in low-power mode is a small percentage of that of the active mode. For example, Reviriego *et al.* [8] report a fraction of 10% for all PHYs in [6], while the CFP module in this study consumes around 1.3 W in the “low-power state,” which corresponds to about 20% of the “ready state” having all ten lanes active. Furthermore, the energy consumption during state transition is assumed equal to the usage in the active mode [8]. The sleep/active transition time

TABLE I
ENERGY CONSUMPTION ESTIMATES FOR DIFFERENT SCENARIOS

Scenario	Direction	Energy (% of peak)	Link load	Aver. frame size (Bytes)
Data center: File and search server	Input	81.68	1.22	87
	Output	82.58	52.21	1497
Data center: Search server	Input	81.33	8.51	945
	Output	79.25	7.23	934
Data center: File and app. server	Input	90.01	0.65	130
	Output	90.73	4.02	749

is taken 2 ms for the optical transceivers in EEE. Such value is also expected for the *lane assertion time* on a hardware-level of the CFP MSA modules [15].

In the first numerical experiments, traces from 1-Gb/s links will be used to measure the energy consumption when the larger transition times are used, that is, compared with the earlier analysis done in [8] for copper PHYs. The 1-Gb/s traces are the same as used in the EEE performance evaluation and correspond to different types of servers in a commercial data center [8]. The results obtained in this numerical experiment are shown in Table I. It can be observed that for low loads, power consumption is close to that of the active mode. This means that even for lightly loaded 1-Gb/s links, poor energy efficiency will be achieved, if EEE mechanisms are applied to optical links due to the larger transition times between modes. For links at 40 Gb/s or 100 Gb/s, the energy consumption would be even worse as for the same load level many more frames will be sent, causing the link to change modes continuously.

To confirm the results, a second analysis was done using traces from a 10-Gb/s optical link. The monitoring equipment is located at Equinix data center in San Jose, CA and is connected to an OC192 backbone link of a Tier1 internet service provider (ISP) between San Jose and Los Angeles. This ISP has multiple OC192 links between these cities and the load balancing is done per flow. The traces are described in [16] and available through the cooperative association for internet data analysis (CAIDA). The link load in the traces varies from 10% to 15% of the link capacity. The maximum frame interarrival time has been measured and it was observed that in all traces, there was not a single millisecond without a frame being transmitted. This means that if EEE mechanisms were used, the link would never be in low-power mode. The link rates in 802.3ba are 40 Gb/s or 100 Gb/s, so the situation will be even worse. In Fig. 2, the average frame interarrival time in the analyzed traces are shown. It is interesting that the average frame interarrival time is a few microseconds in all traces. This means that even transition times in the order of microseconds would achieve poor energy efficiency. As an example, if the transition times specified in the EEE standard draft for 10GBASE-T (below 5 μ s for activation and deactivation) are used, a consumption of at least 80% that of the active mode is obtained for all the traces. Therefore, even if the transition times for the optical PHYs could be reduced by two orders of magnitude, the performance of the EEE approach would still be poor. From the simple numerical experiments that have been discussed, it becomes apparent that the approach

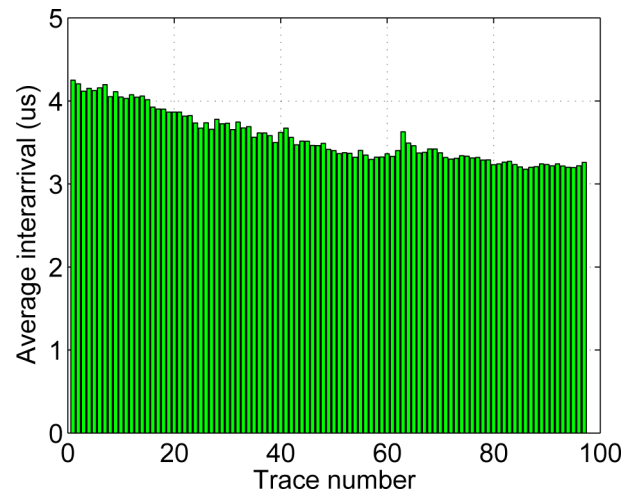


Fig. 2. Average frame inter arrival times in a 10-Gb/s optical link for different traces.

used in the EEE standard will not be effective for 802.3ba high-rate links. In the following section an alternative approach is proposed.

IV. MULTILANE APPROACH FOR ENERGY-EFFICIENT IEEE 802.3BA HIGH-RATE OPTICAL LINKS

An alternative to achieve good energy efficiency in high-speed optical links is to exploit the multilane architecture of 802.3ba PHYs. This can be done by selectively activating some lanes while others are kept in low power mode. This is similar to the techniques proposed in [17]–[19] to improve energy efficiency in link-aggregated groups using the IEEE 802.3ax Standard [1]. In that case, links in a link-aggregated group are active or idle depending on the traffic load, and the aggregation is done above the MAC layer, while in the case considered here, the lanes are inside the PHY layer. For link aggregations, transitions between active and standby modes of the links require coordination between the link partners using link aggregation control protocol. This means that transitions would require a relative long time. Additionally, links are managed as a whole, that is, both link directions are active or idle, but there is no possibility to have a link active in one direction and idle in the other. The proposed approach does not have those limitations, as transitions can be done faster and for each link, directions are managed independently. Finally, in a link aggregation, traffic distribution among the links is quite complex [1], while in the approach in this study, the distribution is done in a simple round-robin fashion, as described in the IEEE 802.3ba Standard.

This study assumes that the CFP MSA and IEEE 802.3ba have adopted dynamic lane handling during uptime. Then, the proposed approach is schematically introduced in Fig. 3 with a flowchart. A traffic measurement/estimation algorithm is used to estimate the traffic load, and from that, the required number of active lanes is determined. A wide number of alternatives can be used to estimate the traffic load. This study proposes a solution for this problem. The optimal solution for all traffic situations is out of the scope of this paper. To determine the number of active lanes, thresholds can be used such that when the load

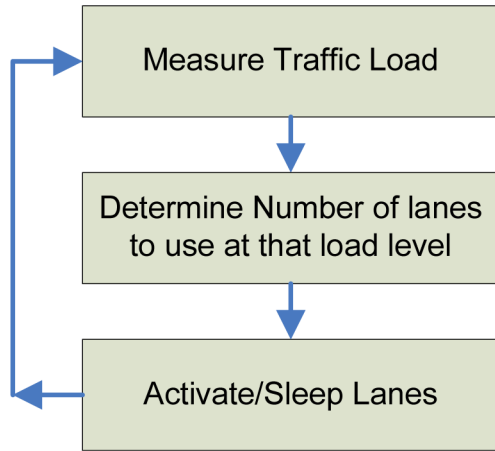


Fig. 3. Flowchart of the proposed technique.

exceeds a threshold, a lane is activated, and conversely, when the load is lower than a threshold, a lane is put in idle mode. This activation/deactivation process reduces the total amount of power consumed by the devices. To select the thresholds, the implications on the frame delay and discard probability should be considered. For example, the number of active lanes should never exceed a given percentage of load (say in the range of 60%–90%), otherwise, the delay experienced by the packets may be excessive, according to basic queuing theory [21].

As shown in Fig. 3, the proposed method has three main steps: 1) measure the traffic load; 2) determine the number of lanes; and 3) update the number of required active lanes. Regarding the first step, the well-known exponential-weighted moving-average (EWMA) algorithm is considered in order to estimate λ , which represents the number of packets per second. Essentially, the EWMA algorithm proceeds as follows: for every new packet arrival with interarrival time x_n with respect to the earlier packet $n \geq 1$, the average interarrival time is estimated as follows:

$$\hat{x}_n = \frac{W}{W+1} \hat{x}_{n-1} + \frac{1}{W+1} x_n \quad (1)$$

where W is the weighting factor. Then, the estimated packet arrival rate $\hat{\lambda} = (\hat{x}_n)^{-1}$ is calculated after choosing a suitable value for W . The reader should note that $\lambda = 1/E[x_n]$. The choice of parameter W influences the “memory” of the packet arrival rate estimation, that is, small W gives more importance to new samples than large W . On the other hand, a large W provides a smoother estimate of λ . An estimate of the average frame length is required in order to estimate the load next to the interarrivals. This is also done using an EWMA algorithm.

The second part of the algorithm in Fig. 3 determines how many lanes should be active. Accordingly, a threshold-based mechanism is proposed in which the decision to activate a lane is based on the amount of incoming traffic that must be handled. A certain percentage of the lane occupation th_{UP} should be exceeded before a lane activation command is passed on to the final step of the algorithm. Similarly, a lane deactivation threshold is defined th_{DOWN} , in case the load drops to a certain level with $\text{th}_{\text{UP}} > \text{th}_{\text{DOWN}}$. A double threshold is used to avoid instability when the load is close to th_{UP} . Finally, the algorithm

then decides to keep the same number of lanes, or to de-activate a lane. The latter involves a lane assertion time, which is expected not to be larger than 2 ms on a hardware level for CFP MSA modules [15]. Recall from Section III that the MDIO response speed on state transitions is typically much slower due to latency of the bus and controls. It is worth mentioning that, for example, the MSA of the small form factor pluggable hot pluggable 10-Gb/s module also recommends a 2 ms value for the maximum turn-ON time of the optical channel [22]. Such value is used in this study to represent a worst case scenario. The lane deassertion time is expected not to be longer than 100 μs on the hardware level [15]; however, longer times may be expected for long-range dense wavelength division-multiplexing configurations.

V. PERFORMANCE EVALUATION

A. Simulation Parameters

A Poisson model is assumed for the frame arrivals, which is consistent with existing studies of high-rate links that carry aggregated traffic [23]. Analyzing the complementary cumulative distribution function of the interarrivals, the publicly available network traces of CAIDA [16] used in the earlier section have been confirmed to show approximately exponentially distributed interarrivals so that the Poisson model is a reasonable approximation. Accordingly, the model-based evaluation is complemented with simulations done with the CAIDA traces in order to countercheck the results. The frame lengths are taken as exponentially distributed with an average length of 600 B. This frame length corresponds approximately to the average frame length observed in the CAIDA traces. The parameter W is set to 1/512 and 1/1024 for the 40 Gb/s and 100 Gb/s, respectively. The average frame length estimator employs a much lower value, namely, $W = 1/16384$. The lane deactivation threshold th_{DOWN} is set to 60% in all simulations, while the lane activation threshold th_{UP} is varied between 70 and 90% in steps of 10%. Regarding the CFP module of [11], the energy consumption in the sleep mode and the lane de-activation times have been reported in earlier sections. Deasserting lanes for symmetric traffic saves around 500 mW per lane, which includes the transmitter and receiver.

B. Energy Savings

To estimate the energy savings, a simple analysis is presented that assumes a static link load and the number of active lanes is set ideally by the algorithm. Using these assumptions, Fig. 4 compares the power usage when the traffic load is increasing for the legacy Ethernet and the adaptive multilane solution. It is shown that the threshold value th_{UP} greatly impacts on the number of lanes used and, consequently, the total energy consumption of the system. Depending on such value, the system can save more or less energy, but significant savings are achieved in most cases compared to the existing standard.

To further evaluate the potential savings of the multilane approach, a scenario with asymmetric traffic is considered. In this case, the load in one direction is assumed to be twenty times less than in the other such that for that direction only one lane is

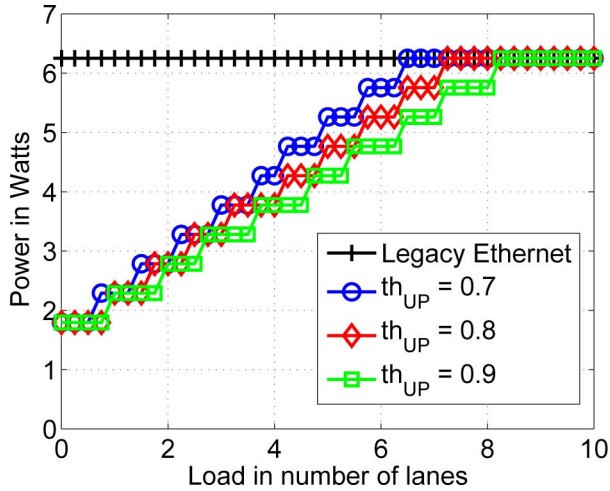


Fig. 4. Energy consumption comparison between the legacy Ethernet solution and the adaptive multilane algorithm for different th_{UP} values in case of symmetric traffic and characteristics of module A.

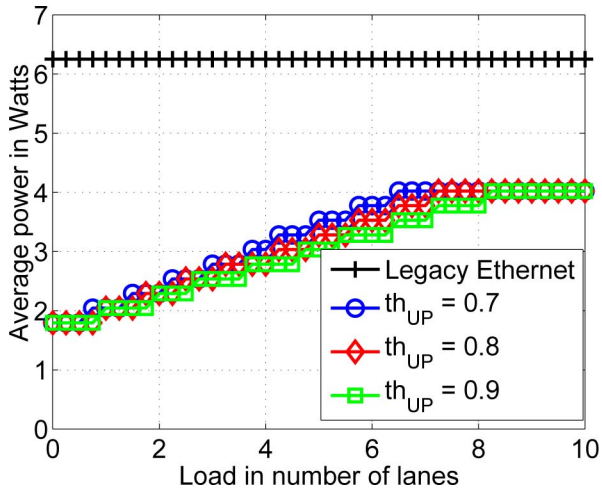


Fig. 5. Average energy consumption comparison between the legacy Ethernet solution and the adaptive multilane algorithm for different th_{UP} values in case of asymmetric traffic and characteristics of module A.

active. Regarding the CFP module in this study, it would merely have to deassert some lanes because the incoming and outgoing traffic are space-division multiplexed [11]. The average power consumption for both link directions is shown in Fig. 5. In this case, larger energy savings are obtained for all loads (note that the x -axis refers to the load on the direction with more traffic). This is an interesting result as many links carry asymmetric traffic. In those links, large data frames (~ 1500 B) are sent in one direction and the much smaller acknowledgment (ACK)-packets in the other (~ 70 B). This means that if a two-state active/idle approach is used, as in EEE, poor energy savings would be obtained, as ACKs would activate the link frequently even if the load is low. This is seen in Table I for the first type of server. In fact, the case shown in Fig. 5 represents the most asymmetric case such that the energy savings of any other traffic scenario will be in between the results shown in Figs. 4 and 5.

When estimating the energy savings, the additional power consumption needed to implement the proposed algorithm has not been included. This functionality should ideally be implemented in some of the existing components to reduce complexity. The main elements needed are the EWMA algorithm and some additional memory to buffer frames while lanes are activated. The additional memory requirements are small, as the latency increase is bounded to $10 \mu s$, as discussed in the following. This means a worst case requirement of an additional 1 Mb of memory. Given the simplicity of the EWMA algorithm and the small memory increase, their power usage should be a small fraction of the module's consumption, and therefore, can be neglected as a first approximation.

C. Impact on Delay

It is well known that as link load approaches the link capacity, the waiting time of the frames grows exponentially. This means that if th_{UP} is close to unity, frames can suffer a large delay possibly impacting on the performance of applications and services or causing buffer overflows and even frame loss. On the other hand, if lanes are added when the traffic is well below the capacity of the currently active lanes, then energy efficiency will be poor, as is shown in the earlier section.

The next numerical experiment shows how the algorithm adapts to incoming traffic and the performance in terms of delay of the multilane algorithm. We assume a 100-Gb/s link using a configuration of ten lanes of 10 Gb/s. Fig. 6(a) illustrates the traffic pattern injected to the system, where λ is varying according to a sin function. The multilane algorithm is changing the number of active lanes when the traffic is changing. Depending on the th_{UP} value, the algorithm makes the decision sooner or later. For the sake of completeness, the results for a 4×25 Gb/s configuration are shown in Fig. 7, where now lane activation/deactivation is less frequent due to the reduced number of lanes. Current traffic patterns for core networks do not change as fast as this example pattern [16]. Depending on the traffic pattern, the parameters of the EWMA-algorithm can be tuned to follow the incoming traffic variations. This EWMA algorithm can be easily implemented and it is used in other networking systems such as in the round-trip time-estimation algorithm of the original specification of TCP [24]. Fig. 8 shows the queuing delay when the packets traverse a 10×10 Gb/s link using the multilane algorithm. In light of the results, the queuing delay is higher than the existing Ethernet solution that uses all lanes, but it remains in the same order of magnitude at least when th_{UP} is 0.7 or less [see Fig. 8(b)]. The waiting time values are in the range of a few microseconds; therefore, it would be reasonable for most applications. In fact, they are comparable to the mode transition time specified in EEE for 10GBASE-T, which is around $5 \mu s$ [6]. This means that the proposed scheme would introduce at most a delay similar to that suffered in EEE at 10 Gb/s when the link has to be activated. An incorrect number of lanes can raise the queuing delay degrading the network performance. When the th_{UP} threshold is set to 0.9, the delay increases by one order of magnitude. Similar results are obtained for the queuing delay in the case of the other two lane configurations proposed by the

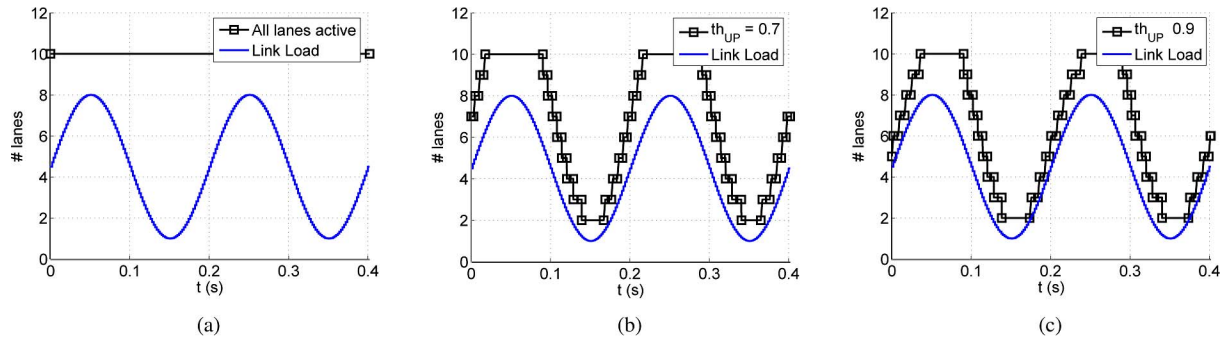


Fig. 6. Pattern traffic in the numerical experiment and number of lanes active when using the algorithm in a 10×10 Gb/s configuration. (a) All lanes active. (b) $th_{UP} = 0.7$. (c) $th_{UP} = 0.9$.

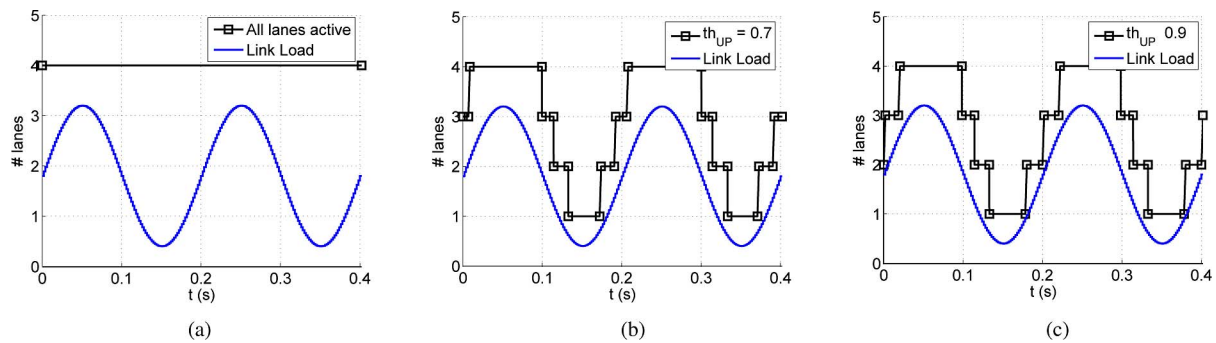


Fig. 7. Pattern traffic in the numerical experiment and number of lanes active when using the algorithm in a 4×25 Gb/s configuration. (a) All lanes active. (b) $th_{UP} = 0.7$. (c) $th_{UP} = 0.9$.

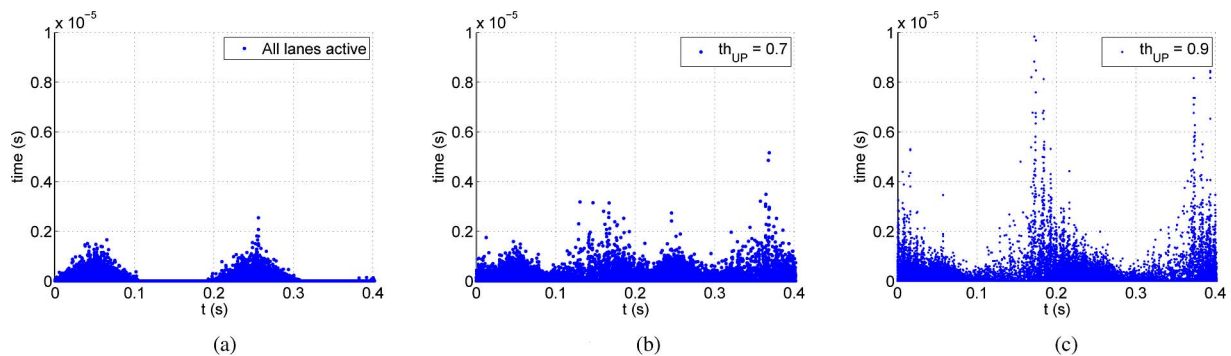


Fig. 8. Queue waiting time for different th_{UP} values when using the algorithm in a 10×10 Gb/s configuration. (a) All lanes active. (b) $th_{UP} = 0.7$. (c) $th_{UP} = 0.9$.

802.3ba amendment. This is the tradeoff, meaning that larger values of th_{UP} achieves larger energy savings at the expense of increased frame delay.

D. CAIDA Trace Validation

To validate the achieved results for Poisson traffic, Fig. 9 depicts the delay experienced by the packets when following the dynamic lane switch-OFF mechanism using CAIDA traces [16]. In the simulations, a hypothetical 100-Gb/s Ethernet switch is fed with ten measured OC192 traffic traces (that is, ten traces from 10-Gb/s links) destined to its 100-Gb/s output port, as an

initial approximation of the traffic expected on 100-Gb/s links. The load of each trace is determined to be about 10%–15%. If the 100-Gb/s output port has the 4×25 Gb/s 100GBASE-R configuration, the number of active lines shall be constant at 1 because a 15% traffic load on a 100-Gb/s link does not exceed the lowest considered threshold of 70% for a single lane. In terms of delay, it can be said that the results are similar to the Poisson model. Visually, there is a greater difference between the results in Fig. 9 than in Fig. 8. The reason is that in this case, the all-lanes active scenario is using four lanes, while the EWMA algorithm requires just one lane. However, the delay remains in the order of few microseconds. Therefore,

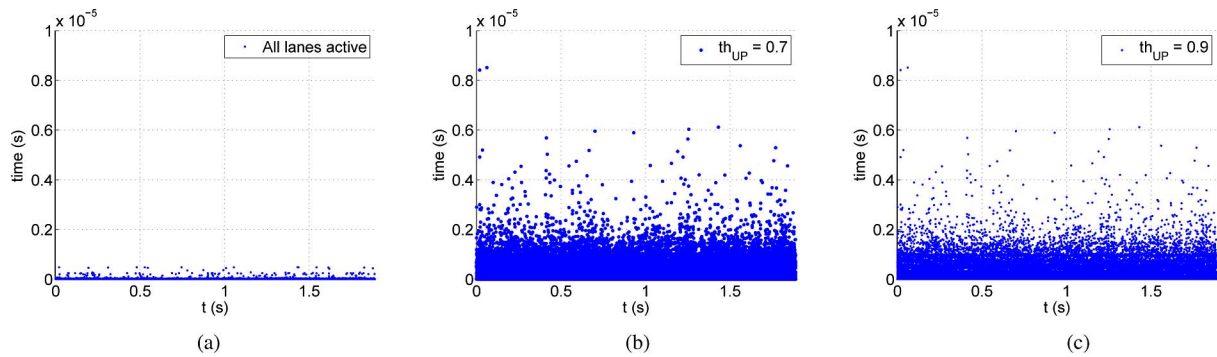


Fig. 9. Queue waiting time for an emulated 100-Gb/s trace based on 10-Gb/s CAIDA measurements. (a) All lanes active. (b) $th_{UP} = 0.7$. (c) $th_{UP} = 0.9$.

it may be concluded that the multilane algorithm can work in a real-life scenario. The numerical experiments with real traces also demonstrate that lane activation/deactivation may not be as frequent as in the model-based simulations that were designed to force rapid and frequent changes in the number of lanes to test the algorithm. In any case, if the traffic pattern changes, there are several estimators that can be used to adapt to the new traffic conditions and follow its variation. Also different thresholds can be used for lane activation/deactivation.

VI. CONCLUSION

In this paper, the improvement of energy efficiency in 802.3ba high-rate Ethernet optical links has been studied. The initial analysis shows that a two-state active/idle strategy would not provide significant energy savings due to the large transition times and small times between frame arrivals. Based on this result, an alternative approach that exploits the multilane architecture of 802.3ba Ethernet PHYs has been proposed. The approach adaptively changes the number of active lanes to the link load to achieve energy savings without impacting performance. Simulations with a Poisson model and with real traces have then been presented to evaluate the performance of the proposed scheme. Measured numbers on the energy consumption of a commercial 100-GbE module have been used. The results indicate that energy savings may be achieved with little expected impact on performance in terms of frame delay. An implementation at the hardware level should be considered to avoid timing delays of the MDIO interface.

ACKNOWLEDGMENT

The authors would like to thank D. Larrabeiti for his contribution to the simulation experiments reported in this paper.

REFERENCES

- [1] R. Seifert and J. Edwards, *The All-New Switch Book, The Complete Guide to LAN Switching Technology*. New York: Wiley, 2008.
- [2] Home page of the IEEE 802.3ba 40Gb/s and 100Gb/s Ethernet Task Force. (2010, Apr.). [Online]. Available: <http://grouper.ieee.org/groups/802/3/ba/public/index.html>
- [3] C. Cole, J. D'Ambrosia, C. DiMinico, H. Frazier, A. Healey, J. Jaeger, J. Jewell, M. Nowell, and S. Trowbridge, (2007, Nov.). An overview: The next generation of ethernet, *IEEE 802.3-HSSG Meeting* [Online]. Available: <http://www.ieee802.org/3/hssg/public/nov07/index.htm> (Accessed Apr. 2010).
- [4] GeSI Report. (2008, Jun.). *SMART 2020: Enabling the low carbon economy in the information age* [Online]. Available: <http://www.gesi.org/ReportsPublications/tabid/60/Default.aspx> (Accessed Apr. 2010).
- [5] C. Gunaratne, K. Christensen, B. Nordman, and S. Suen, "Reducing the energy consumption of ethernet with adaptive link rate (ALR)," *IEEE Trans. Comput.*, vol. 57, no. 4, pp. 448–461, Apr. 2008.
- [6] Home page of the IEEE 802.3az Energy Efficient Ethernet task force. (2010, Apr.). [Online]. Available: <http://grouper.ieee.org/groups/802/3/az/public/index.html>
- [7] D. Law and J. D'Ambrosia. (2009, May). "IEEE P802.3ba: Architecture overview," *IEEE 802.3-EEE Meeting* [Online]. Available: http://www.ieee802.org/3/az/public/apr09/dambrosia_04_0509.pdf (Accessed Apr. 2010).
- [8] P. Reviriego, J. A. Hernández, D. Larrabeiti, and J. A. Maestro, "Performance evaluation of energy efficient ethernet," *IEEE Commun. Lett.*, vol. 13, no. 9, pp. 697–699, Sep. 2009.
- [9] O. Ishida, "40/100GbE Technologies and related activities of IEEE standardization," presented at the OFC 2009, San Diego, CA, Mar., Paper OWR5 (tutorial).
- [10] White paper of the Ethernet alliance. (2008, Nov.). *40 Gigabit Ethernet and 100 Gigabit Ethernet technology overview* [Online]. Available: <http://www.ethernetalliance.org/> (Accessed Apr. 2010).
- [11] Reflex Photonics 100GBASE-SR10. (2010, Apr.). 100 Gbps Ethernet, CFP fiber optic transceiver module, part number CF-X12-C11801-02 [Online]. Available: <http://www.reflexphotonics.com/>
- [12] CFP MSA Management interface specification, version 1.2. (2009, Sep.). [Online]. Available: <http://www.cfp-msa.org/documents.html> (Accessed Apr. 2010).
- [13] O. Haran. (2007, Sep.). "Applicability of EEE to fiber PHY's," *IEEE 802.3-EEE Meeting* [Online]. Available: http://www.ieee802.org/3/eee_study/public/sep07/haran_1_0907.pdf (Accessed Apr. 2010).
- [14] R. Kubo, J. Kani, Y. Fujimoto, N. Yoshimoto, and K. Kumozaki, "Proposal and performance analysis of a power-saving mechanism for 10-Gigabit-class passive optical network systems," in *Proc. Netw. Opt. Conf. Opt. Cabling Infrastructure 2009*, Valladolid, Spain, Jun., pp. 87–94.
- [15] Private communication with Matt Traverso (Opnext), January 15, 2010.
- [16] C. Walsworth, E. Aben, K. C. Claffy, and D. Andersen. (2009, Jan.). *The CAIDA anonymized 2009 Internet traces* [Online]. Available: http://www.caida.org/data/passive/passive_2009_dataset.xml (Accessed Apr. 2010).
- [17] Method for Ethernet power savings on link aggregated groups, by D. J. Koenen and M. Chuang. (2008, Dec.). *Patent* US2008/0 304 519 A1 [Online]. Available: <http://www.wipo.int/>
- [18] Y. Fukuda, T. Ikenaga, H. Tamura, M. Uchida, K. Kawahara, and Y. Oie, "Performance evaluation of power saving scheme with dynamic transmission capacity control," in *Proc. GLOBECOM 2009*, Honolulu, HI, Dec., pp. 1–5.
- [19] H. Imaizumi, T. Nagata, G. Kunito, K. Yamazaki, and H. Morikawa, "Power saving mechanism based on simple moving average for 802.3ad link aggregation," in *Proc. GLOBECOM 2009*, Honolulu, HI, Dec., pp. 1–5.
- [20] H. Imaizumi, T. Nagata, G. Kunito, K. Yamazaki, and H. Morikawa, "Power saving technique based on simple moving average for multi-channel Ethernet," in *Proc. OECC 2009*, Hong Kong, P.R. China, Jul., pp. 1–2.
- [21] L. Kleinrock, *Queueing Systems: Volume I: Theory*. New York: Wiley Interscience, 1975.

- [22] Home page of the XFP MSA group. (2010, Apr.). [Online]. Available: <http://www.xfpmsa.org/>
- [23] T. Karagiannis, M. Molle, M. Faloutsos, and A. Broido, "A non-stationary Poisson view of internet traffic," in *Proc. INFOCOM 2004*, Hong Kong, China, Mar., pp. 1–12.
- [24] J. Postel. (1981, Sep.). Transmission control protocol. *IETF RFC 793* [Online], pp. 1–85. Available: <http://tools.ietf.org/html/rfc793> (Accessed Apr. 2010).

P. Reviriego (M'03–A'03–M'04) received the M.Sc. and Ph.D. degrees (Hons.) in telecommunications engineering from the Technical University of Madrid, Madrid, Spain, in 1994 and 1997, respectively.

From 1997 to 2000, he was an R&D Engineer at Teldat, Madrid, where he was involved in router implementation. In 2000, he joined Massana, where he was involved in the development of 1000BASE-T transceivers. During 2003, he was a Visiting Professor at University Carlos III, Leganés, Madrid. During 2004–2007, he was a Distinguished Member of Technical Staff at LSI Corporation, where he was involved in the development of Ethernet transceivers. He is currently at Universidad Antonio de Nebrija, Madrid. He has authored numerous papers in international conferences and journals. He has also participated in the IEEE 802.3 standardization for 10GBASE-T. His research interests include fault-tolerant systems, performance evaluation of communication networks, and the design of physical layer communication devices.

B. Huiszoon (S'03–M'09) was born in Vlissingen, The Netherlands, in 1978. He received the M.Sc. and Ph.D. degrees in electrical engineering from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 2003 and 2008, respectively.

He won a "Juan de la Cierva" contract from the Spanish Ministry Ministerio de Ciencia e Innovación (MICINN), and then joined the Universidad Autónoma de Madrid (UAM), Madrid, Spain, in December 2008, where he is currently a Postdoctoral Researcher. He is also involved as a Project Leader and a member in contract research at UAM. He has authored the book *Optical Code Communication in Local Networks*, holds a Dutch patent, and co-authored more than 30 papers in peer-reviewed journals and international conferences. His research interests include fiber-optic access networks and systems, optical code division multiple access, energy-efficiency of telecom networks, and optical-wireless convergence.

Dr. Huiszoon is a recipient of Mignot prize 2004 for the results of his M.Sc. thesis, and the results of his Ph.D. thesis were awarded with the IEEE LEOS Graduate Student Fellowship 2008, and the Dutch Veder prize 2008. He was a Technical Program Committee (TPC) Member at the IEEE Lasers and Electro-Optics Society Summer Topicals 2009 and Optical Networking Design and Modeling 2010 conferences.

V. López (M'10) received the M.Sc. (Hons.) degree in telecommunications engineering from Universidad de Alcalá De Henares, Spain, in 2005 and the Ph.D. (Hons.) degree in computer science and telecommunications engineering from Universidad Autónoma de Madrid (UAM), Madrid, Spain, in 2009.

In 2004, he joined Telefónica I+D as a Researcher, where he was involved in next generation networks for metro, core, and access. He was involved with several European Union projects Next Generation Optical Network for Broadband European Leadership (NOBEL), Multi Service Access Everywhere (MUSE), multi-partner European test beds for research networking (MUPBED). In 2006, he joined the High-Performance Computing and Networking Research Group (UAM) as a Researcher in the ePhoton/One+ Network of Excellence. Currently, he is an Assistant Professor at UAM, where he is involved in optical metro-core projects Building the Future Optical Network in Europe (BONE) and Metro Architectures enabling Sub-wavelengths (MAINS). His research interests include the analysis and characterization of services, design, and performance evaluation of traffic monitoring equipment, and the integration of Internet services over optical networks, mainly Optical Burst Switching (OBS) solutions and multilayer architectures.

R. B. Coenen received the B.S. degree in computer engineering from the University of Victoria, Victoria, BC, the M.A.Sc. degree in electrical engineering from the University of Waterloo, Waterloo, ON, and the Ph.D. degree in electrical engineering from the University of British Columbia, Vancouver, BC.

He was a Senior Design Engineer at Gtran Corporation, where he was the Leader of various optical and electrical design teams. He was a Senior Applications Engineer at Redfern Integrated Optics and at Scintera Networks. He was a Senior Design Engineer at Finisar Corporation, where he was involved with product development efforts and provided key customer support in the company's optical transceiver product line. He is currently the Director of Product Management at Reflex Photonics Inc., Sunnyvale, CA, where he is responsible for overseeing global marketing, business development and customer initiatives related to the company's product lines, as well as managing original equipment manufacturer (OEM) and partner application engineering support. He was involved with various standards committees and MSA organizations such as: the Optical Internetworking Forum, IEEE 802.3ba, 802.3aq and 802.3ae, Small Form Factor (SFF) Multi-Source Agreement (MSA) Committee, T11, and Infiniband.

J. A. Hernández received the M.Sc. degree in telecommunications engineering from Universidad Carlos III de Madrid, Madrid, Spain, in 2002, and the Ph.D. degree in computer science from Loughborough University, Leics, U.K., in 2005.

From 2005 to 2009, he was a Postdoctoral Research and Teaching Assistant at Universidad Autónoma de Madrid, where he was involved in a number of both national and European research projects concerning the modeling and performance evaluation of communication networks, and particularly the optical burst switching technology. In 2009, he joined Universidad Carlos III de Madrid, where he is currently a Visiting Lecturer and a Senior Researcher. He has authored or coauthored more than 40 articles published in international journals and conference proceedings. His research interests include the mathematical modeling and computer networks overlap.

J. A. Maestro (M'07) received the M.Sc. degree in physics and the Ph.D. degree in computer science from Universidad Complutense de Madrid, Madrid, Spain, in 1994 and 1999, respectively.

He was both a Lecturer and a Researcher at several universities as Universidad Complutense de Madrid, Universidad Nacional de Educación a Distancia (Open University), Madrid, Saint Louis University, Madrid, and Universidad Antonio de Nebrija, Madrid. He is currently at Universidad Antonio de Nebrija, where he is the Head of the Computer Architecture and Technology Group. His current activities has been concerned with the Space field, with several projects on reliability and radiation protection, and collaborations with the European Space Agency. He is the author of numerous technical publications, both in journals and international conferences. Besides from this, he was involved with several multinational companies, managing projects as a Project Management Professional, and organizing support departments. His research interests include high-level synthesis and cosynthesis, signal processing and real-time systems, fault-tolerance, and reliability.