Received <day> <Month>, <year>; | Revised <day> <Month>, <year>; | Accepted <day> <Month>, <year>

DOI: xxx/xxxx

SPECIAL ISSUE PAPER

# Modeling and Analysis of the Performance of Exascale Photonic Networks

José Duro*[1] | Jose A. Pascual[2] | Salvador Petit[1] | Julio Sahuquillo[1] | María E. Gómez[1]

[1]Departamento de Ingeniería de Sistemas y Computadores, Universitat Politècnica de València, Valencia, Spain
[2]Advanced Processor Technology (APT) Research Group, The University of Manchester, Manchester, United Kingdom

**Correspondence**
*Email: jodugo1@gap.upv.es

**Summary**

Photonics technology has become a promising and viable alternative for both on-chip and off-chip interconnection networks of future Exascale systems. Nevertheless, this technology is not mature enough yet in this context, so research efforts focusing on photonic networks are still required to achieve realistic suitable network implementations. In this regard, system-level photonic network simulators can help guide designers to assess the multiple design choices. Most current research is done on electrical network simulators, whose components work widely different from photonics components. In this work, we summarize and compare the working behavior of both technologies which includes the use of optical routers, wavelength-division multiplexing and circuit switching among others. After implementing them into a well-known simulation framework, an extensive simulation study has been carried out using realistic photonic network configurations with synthetic and realistic traffic. Experimental results show that, compared to electrical networks, optical networks can reduce the execution time of the studied real workloads in almost one order of magnitude. Our study also reveals that the photonic configuration highly impacts on the network performance, being the bandwidth per channel and the message length the most important parameters.

**KEYWORDS:**
Interconnection networks, Photonic technology, Simulation framework

## 1 | INTRODUCTION

The most powerful supercomputers in the world (1) are ranked by their computational power in terms of floating-point operations executed per second (FLOPS). The Sunway TaihuLight, the recent supercomputer leading the list at November 2017, realizes by 93 PetaFlops ($10^{15}$) with 10,5 million cores. The Top500 list tracks the computational power of supercomputers since 1971, and according to the current growing computational trend, it is expected that supercomputers will break the ExaFlop ($10^{18}$) barrier by 2020. Reaching this target, however, is challenging and requires from multiple simultaneous solutions addressing, among others, computation at chip level (nodes of the system), data movement across the system, distributed storage, energy management, etc.

From the aforementioned challenges, the data movement challenge is probably the most critical to be achieved, mainly due to the increasing number of computing nodes and, therefore, the increasing communication requirements. *Exascale* networks will count with thousands of computing nodes, so

[0]This work was supported by the ExaNeSt project, funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 671553, and by the Spanish Ministerio de Economía y Competitividad (MINECO) and Plan E funds under Grant TIN2015-66972-C5-1-R. Jose A. Pascual was supported by a HiPEAC Collaboration Grant.

data transmission among them becomes a major design concern, and new requirements rise, not only in terms of throughput, but also in energy demands. In such systems, the underlying network technology (2, 3) is a critical design choice and this is the focus of the European ExaNeSt project which is currently being developed (see Section 2.1 for more details).

In this regard, photonics interconnects –both on-chip and off-chip– have emerged as a worth alternative technology addressing the key constraints of traditional electrical networks. This technology provides much more bandwidth than electrical technology with much less energy consumption (4, 5). Optical Networks on-Chip (ONoCs) (6, 7) will become a viable option for the growing demand of high performance computing (HPC) applications that electric networks cannot efficiently deal with. On the other hand, off-chip photonics technology can provide what is required to cover the rising requirements in Exascale computing, contributing additional advantages over electrical technology such as the volume of the interconnection links (inter- and intra-cabinet) or the power savings. Depending on the technology node (from 90 nm to 22 nm), photonics technology is from 7 to 4 orders of magnitude smaller than electrical technology (8).

The development of such a system requires from multiple performance evaluation studies in order to guide the system construction. Regarding interconnects, simulation frameworks are being used to assess network bandwidth and network latency. Original ExaNeSt simulation frameworks (see Section 2.1) model electrical interconnection networks and they are not prepared to simulate networks with photonic technology.

Modeling photonics networks is not a straightforward process and requires a sound knowledge on the basics of the state of the art photonic technologies. Many aspects are widely different from the traditional electrical networks since photonics offers new possibilities but also prevents the use of some others. For instance, photonics provides the capability of sending multiple messages concurrently on a given link, while it does not support flit storage at the intermediate network routers. The key objective of this work is the identification of the components and technologies required to model a photonic interconnect and its implementation into a simulation framework that will help on the study and evaluation of network techniques for such future networks.

In order to validate the simulation framework and perform an initial evaluation of photonic technologies we have performed an extensive simulation study using three traditional network topologies using both synthetic and traffic extracted from real parallel applications. In particular, we have used applications traces obtained by ExaNeSt partners, which are being used to design and evaluate the ExaNest network architecture. These traces perform a high amount of communications compared to computation, which is key to evaluate the network. As a result of this study in which we have tested different photonic configurations, we have identified the main factors affecting the performance of the interconnect. Moreover, we have also compared the performance delivered by this new technology with traditional electrical networks. Initial results show the potential of this technology to execute parallel applications achieving orders of magnitude lower execution times.

The remainder of this paper is organized as follows. After presenting in Section 2 the ExaNeSt project and some photonics background we continue in Section 3 describing the simulation framework and discussing the photonic technologies required to model a photonic network. In Section 4, the experimental setup is described. Section 5 discusses the simulation results. Section 6 summarizes the related work, finally, Section 7 presents concluding remarks and future work.

## 2 | BACKGROUND

As mentioned above, the requirements for Exascale computation over the current decade are expected to scale the network performance. In this section, after motivating the present work in the context of the ExaNeSt project, we summarize recent advances in silicon photonics technology and its current state on both off-chip and on-chip interconnects.

### 2.1 | Motivation

The European Exascale System Interconnect and Storage project (ExaNeSt) (9) is currently designing and building a prototype architecture capable of reaching Exascale computation. The aim of ExaNeSt is to develop a system that can be scaled up to the tens of millions of interconnected low-power consumption ARM cores to solve large-scale scientific and big data problems. In order to support a system of this size, ExaNeSt is confronted with the huge challenge of designing an interconnect able to meet very strict performance, resilience, and cost constraints for a range of computational challenges. The ExaNeSt interconnect is a multi-tier interconnect which can be divided into two distinct parts. The lower tiers, which are physically fixed by means of boards and backplanes, and the higher tiers which are fully reconfigurable using custom-made FPGA-based (10) routers. This flexibility allows to build any network topology, i.e. direct, indirect or hybrid, or even the use of standard off-the-shelf commodity switches. In order to meet the requirements of such demanding interconnect, we explore in this work the use of photonic technology within the higher tier. In particular, we evaluate whether the use of traditional topologies and parallel scientific applications can take advantage of the benefits provided by this new technology such as high network bandwidths and very low latencies.

## 2.2 | Off-Chip Silicon Photonics

Silicon photonics-based interconnects are being widely deployed in data communication (datacom) systems due to their potential to achieve large scale and low cost integration together with low power operation. This potential relies on advantages like higher bandwidth capability, better distance/speed trade-off and easier cable management. Regarding bandwidth, achieving more than 10 Gbps with conventional copper wires remains a challenge, while a single optic fiber can offer bandwidths in the Terahertz range. With respect to the distance/speed trade-off, optic fibers are able to transmit data along several kilometers without bandwidth penalties. Finally, due to their inherent lightness and thinness, using optic fiber cables instead of copper ones highly reduces cable density. Thus, it considerably eases cable management.

Current state-of-the-art photonics technologies, however, require from pluggable transceivers to transform electrical signals to optical signals and vice versa, which limits the potential of a full silicon photonics system. To deal with this shortcoming, current research aims to integrate optical devices with logic chips. This goal has not been reached yet and it would make a significant impact on datacom network topology.

The bandwidth limit of current Quad Small Form-factor Pluggable (QSFP) transceivers based on Vertical-Cavity Surface-Emitting Laser (VCSEL) technology is by 40 Gbps. Nevertheless, silicon photonic interconnects are expected to reach the 100 Gbps mark and beyond in the near future. For instance, Intel Corporation and other big companies such as IBM or Cisco Systems have moved their silicon photonics efforts beyond research and development, and have produced engineering samples that run at speeds of up to 100 Gbps. Moreover, Intel and Corning are currently developing the MXC connector, which supports up to 64 fibers communicating at 25 Gbps, reaching an unprecedented data transmission capacity by 1.6 Tbps over a 300 meters distance.

## 2.3 | On-Chip Silicon Photonics

The need of low latency and high bandwidth multicore data transmissions has led CMOS-compatible photonic interconnects as an alternative technology to address these design issues in on-chip networks. Moreover, silicon photonics-based on-chip networks enable the implementation of silicon photonic routers, which are a key development for inter-rack and intra-rack *full*-optical networks based only on optical components –i.e. all-optical networks– for Exascale systems. In this regard, current efforts have concentrated on the realization of reliable hybrid silicon lasers, electro-optic modulators, ring resonators and receivers; the most critical building components of photonic circuits.

Laser sources inject light into the chip's waveguides. Laser sources are probably the most difficult devices to be integrated on silicon. *Duan et al.* have developed hybrid silicon/III-V lasers with less power consumption than previous works (11, 12), although not yet achieving ultra-low power consumption, which will significantly reduce packaging costs.

Electro-optical modulators establish the switching capacity, that is, the operation bandwidth of any photonic integrated circuit (PIC). High bandwidth modulation can be realized in silicon with free-carrier induced index change (13), using biased pn structures (carrier depletion) achieving up to 30-50 Gbps data rates (14, 15).

Optical ring resonators are the key component to leverage wavelength division multiplexing (WDM) (16) technology. WDM allows splitting up the optical signal into multiple independent wavelengths. A ring resonator captures specific optical wavelengths; thus, it can redirect these wavelengths to other waveguides and receivers, so enabling the implementation of complex optical on-chip networks and photonic routers.

Finally, optical coherent receivers (also known as photodetectors), which convert the amplitude, phase, and polarization of an optical signal into the electrical domain have already been integrated, and provide very high data conversion rates (up to 224 Gb/s with PDM-16-QAM signals) (17, 18).

# 3 | SIMULATION FRAMEWORK AND PHOTONIC IMPLEMENTATION

This section presents key features of the INSEE (19), the simulation framework used to evaluate the use of photonic technologies. INSEE was originally developed with the aim of modeling electrical networks and implements multiple, direct and indirect, network topologies (e.g. cubes, dragonfly and trees) and multiple traffic generation methods (e.g. synthetic, traces and architectural simulators). However, it lacks of most of the required components to simulate a photonic-based interconnect. The objective of this section is to discuss the differences between both electrical and photonic technologies and to identify the required components and techniques which need to be implemented into the simulation framework.

## 3.1 | Optical Routers versus Electrical Routers

Electrical routers implement internal buffers that provide local temporal storage for in-transit packets (or a smaller data unit, depending on the switching technique). Packets are kept in these buffers in case they cannot advance due to traffic or network constraints.

On the contrary, all-optical networks do not provide buffering capacity at network routers. This fact is illustrated in Figure 1 , which presents an example of an optical switch with no buffering. In particular, it is shown an 8x8 (8 inputs and 8 outputs) reconfigurable optical switch based on a Benes architecture. The minimal building block is a 2x2 Mach-Zehnder interferometer (MZI) switch element and each optical path goes through 5 stages of those elements. As we can see, waveguide crossings are required for the two-dimensional connection of the 20 switching elements. More details about this kind of optical switches can be found in (20).

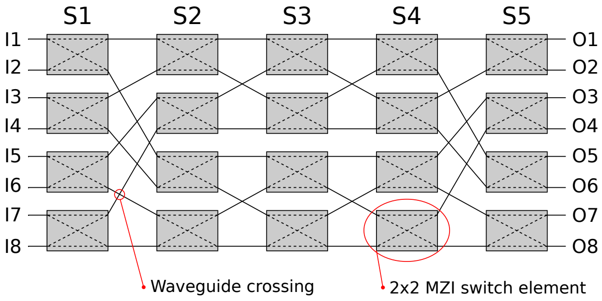


**FIGURE 1** 8x8 optical switch based on a Benes architecture.

The lack of buffering capacities means that once the data is injected in the network it must travel without waits, that is, without being blocked all along its path. An interesting attempt to deal with this drawback could be the use of hybrid electro-optical routers. This approach, however, requires from electro-optical and opto-electrical conversions to write and read data into and from electric buffers, respectively, limiting the achievable bandwidth and latency improvements. As discussed below, the fact that all-optical networks do not provide buffering support has important consequences for the switching technique used in all-optical networks.

## 3.2 | Circuit Switching versus Packet Switching

c

## 3.3 | Wavelength-Division Multiplexing

Although circuit switching can be considered a rather old switching method, it can bring important advantages in optical networks combined with wavelength-division multiplexing (WDM). This technique, used in optical communications, that consists in multiplexing in frequency a number of wavelengths onto the same optical cable. The amount of multiplexed wavelengths depends on the *separation* between them (e.g. as standard, with 100, 50 or 25 GHz there may be up to 40, 80 or 160 wavelengths respectively). When using WDM, the total link bandwidth (i.e. considering all the wavelengths), known as *aggregated bandwidth*, is given by the sum of the bandwidths provided by each individual wavelength.

To model both WDM and circuit switching together in the baseline simulator, several design issues have been considered. First, since there are multiple wavelengths in the same link, circuit switching needs to be adapted since more than one path per link can be defined at the same time; that is, each channel[1] (i.e. set of wavelengths) can be part of an eligible path. Therefore, when a message is ready to be injected into the network, the number of possible paths that it can reserve is much higher than in electric networks.

In a previous work (21), we used a static version of WDM in which once a channel is selected to transmit a packet, this is maintained for the whole path. As a result, the utilization of the photonic channels is severely degraded due to the lack of adaptivity. To deal with this shortcoming, we have implemented the Tunable Wavelength Converter (TWC) (22) component in the router. This technology allows to change the wavelengths used to transmit a packet between two routers, that is, a packet injected into a given channel can make a transition to another channel in the next hop if the current one is not available. As a result, an increase in the channel utilization will rise which is expected to translate in performance improvements.

## 3.4 | Photonic Links versus Electrical Links

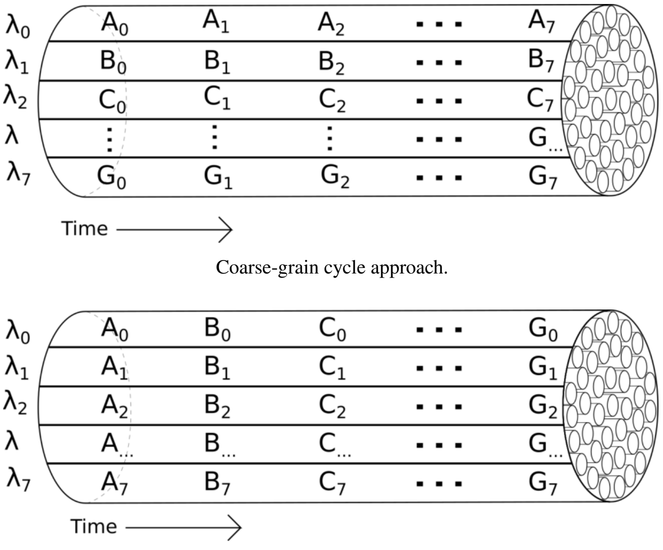
As explained above, electrical links only allow to send information of a single message or packet in a network cycle. In contrast, optical links are split in channels each one using a different set of wavelengths.

To exploit this issue, the proposal provides support to configure the amount of wavelengths per optical link and to group wavelengths in independent channels, each one handling the transmission of a different message. In short, a first configuration option allows specifying the link bandwidth, and a second the channel bandwidth.

## 3.5 | Transmission units: Phits versus Bits

Virtual Cut-Through and Wormhole are the most widely used switching techniques in electrical networks. These techniques split the packet in small *flits*, which are the units of flow control. Flits are in turn divided in *phits* (physical units). A phit is the

[1] Note: The term channel has been used in the literature also to refer to a single wavelength in optical technology. This paper uses this term from a *computer perspective* to refer to a set of wavelengths used to transfer the same message.



Coarse-grain cycle approach.

Packed wavelengths approach.

**FIGURE 2** Example of using eight wavelengths to transmit eight phits, referred to from A to G, with the studied transmission approaches.

amount of bits that can be transferred in a single network cycle. In contrast, optical networks only can transfer one single bit per wavelength and per network cycle (note that optical cycles are much smaller than electrical cycles).

In general, electrical network simulators define the phit size as an integer amount of bytes (8 bits). Therefore, when such a kind of simulator is used as a basis of an optical simulator, the byte is kept as the minimal transference unit per cycle. However, as mentioned above, optical links transfer one bit per wavelength in a given network cycle. Therefore, a new approach is required to fulfill this mismatch.

Two main approaches, *coarse-grain cycles* and *packed wavelengths*, have been devised to model the transmission of bits instead of phits in INSEE. The former approach defines coarse-grain cycles consisting of 8 *small* simulation cycles as the working cycle unit, which allows submitting 8 bits (i.e. 1 byte, the minimum phit size in the baseline simulator) per cycle using the same wavelength. The latter groups 8 wavelengths, which acts as a single transmission unit; that is, 8 wavelengths are used to transmit 8 bits of the same message in a single photonic cycle, which implies that the minimum channel size is 8 wavelengths. Figure 2 presents an example where 8 phits (labeled as A to G) are transmitted in both approaches. In the coarse-grain cycles approach, each phit is transmitted in a different wavelength while in the packed wavelengths approach several wavelengths cooperate to transmit the same phit in parallel.

| Technique | # Channels | Channel Bandwidth |
|---|---|---|
| Coarse-grain cycles | 40 | 40 Gbps |
| Hybrid | 20 | 80 Gbps |
| Hybrid | 10 | 160 Gbps |
| Packed wavelengths | 5 | 320 Gbps |

**TABLE 1** Trade-off between the studied transmission approaches for an optical link populated with 40 wavelengths.

As shown in Table 1 , choosing between both approaches presents a trade-off in the network features. For an optical link composed of 40 wavelengths, the coarse-grain cycle approach can provide up to 40 parallel channels, but each one of these channels only can offer the bandwidth of a single wavelength (40 Gbps). In contrast, the packed approach offers a limited maximum number (i.e. 5) of channels but each one aggregates the bandwidth of 8 wavelengths ($40 \times 8 = 320$ Gbps).

Note that between both approaches there are several possible *hybrid* approaches. For instance, the amount of channels can be halved with respect to the coarse-grain cycles approach (second and third line of the table). Then, instead of transmitting 1 byte in 1 network cycle using 1 wavelength, the byte can be divided in 2 nibbles that are transmitted by 2 wavelengths (doubling the network frequency).

## 3.6 | Topologies and Routing

One of the objectives of this work is to investigate if traditional topology/routing combinations are able to take advantage of the benefits of using photonic technology. To this end, we have evaluated the three most widely used network topologies for HPC together with the most used routing policies:

- Torus: The modeled topology uses Dimension Order Routing (DOR) in which in order to establish a path between each source-destination pair only paths using the minimum number of hops are considered, crossing the network dimensions always in the same order.
- Fat-tree: This topology uses an adaptive version of the Up/Down routing. In this case, among the possible alternatives in the Up stage, we select the links which have the lowest channel utilization aiming at balancing the use of the network links.
- Dragonfly: Finally, this topology uses the Valiant routing algorithm (23). This policy selects randomly an intermediate switch (proxy) and performs minimal routing from the source to the proxy and from the proxy to the destination. Although the maximum length path of Valiant is longer than when using minimal routing (from

5 to 7 hops), it is known that it provides higher load balancing and avoids bottlenecks for specific types of traffic.

We want to remark that, although well-known topologies have been used, each one has its own routing policy, thus, these routing algorithms have been adapted to photonics to select the proper optical channel at each hop along the path.

## 4 | EXPERIMENTAL SETUP

This section presents the experimental setup used to evaluate the performance of the photonic interconnect. After describing the system configuration for both electric and photonic networks, we discuss the characteristic of the traffic. we discuss the characteristic of the traffic.

### 4.1 | System Configuration

The network configuration includes the bandwidth, the number of wavelengths per link, the network topologies and the types of traffic that have been considered to obtain the results.

The experiments consider a 10 Gigabit Ethernet electrical network. In the case of the photonics network 1.6 Tbps is considered as the aggregated bandwidth provided by each photonic link. The electrical bandwidth was chosen because an important set (by 40.8%) of the supercomputers ranked in Top500 (1) list implement this network technology, while the optical was selected according to the current VCSEL technology commented in Section 2.

On the other hand, photonic networks are limited by the optic communication band (24) and, as explained in Section 3.3, the amount of wavelengths depends on the spacing between them. Nowadays, 100 GHz channel spacing is typically used, which gives 40 wavelengths per single optical fiber or link (25), but this spacing can be reduced in order to populate more wavelengths per single fiber or optical link. For instance, 50 GHz spacing allows to populate the link with 80 wavelengths, or recently 160 wavelengths are allowed with a 25 GHz spacing (26).

Table 1 summarizes the main design choices of the studied configurations for a photonic link populated with 40 optical wavelengths. In addition to the two main transmission approaches, labeled as *packing wavelength* and *coarse grain*, hybrid schemes combining both approaches have been studied. All the configurations present an aggregate bandwidth of 1.6 Tbps with 40 wavelengths per link.

### 4.2 | Network Traffic

To evaluate the impact of the number of photonic channels on the execution time two types of traffic have been taken into account, synthetic and real from applications.

Regarding synthetic traffic, we analyze the impact of the packet length in the execution time using uniform traffic, which is related to the ability of establishing a photonic route between the source and destination, in particular we perform several experiments sending 10 GB of data varying the average length of the packets, from 1 KB to 128 KB. These experiments have been carried out using a (8x8x8) 3D torus, a 8-ary 3-fat-tree and a (4,6,4) dragonfly topologies.

With respect to real traffic, two traces collected from the execution of two ExaNeSt MPI based applications, `Gadget`(27) and `Lammps`(28) have been studied. We want to remark that the execution model of INSEE maintains the causality among the messages in the traces, and that the collectives have been translated into point-to-point messages following the algorithms of the Open MPI[2] implementation of the MPI standard.

As these applications are composed of different number of tasks, different network configurations have been used to simulate each trace. First, for `Gadget` which is composed of 72 tasks, we have used a (4x6x3) 3D torus, a 3-ary 4-fat-tree and a (3,4,3) dragonfly. `Gadget` performs short computations, and it mainly uses barriers for synchronization and broadcasts for data exchanges. On the other hand, to simulate `Lammps`, which is composed of 192 tasks, we have used a (4x8x6) 3D torus, a 6-ary 3-fat-tree and a (6,3,4) dragonfly. In this case, several types of collectives are used to perform the data exchange but mainly point-to-point messages. As in the previous case, the amount of computation is negligible compared to amount of data sent through the network. Notice that it is not possible to match the number of nodes of the topologies and the number of tasks of the traces. For that reason, the size of the topologies has been approximated as much as possible to the number of tasks of the applications.

## 5 | EXPERIMENTAL RESULTS

This section analyzes the results obtained for several photonic and electric network configurations with both our modified INSEE[3] and the original one [4]. First, a detailed analysis of the results obtained using synthetic traffic and several photonic configurations. The, we analyze the performance achieved by two real applications from the ExaNeSt project in both electric

[2] https://www.open-mpi.org

[3] https://bitbucket.org/joseduro/photonic-insee

[4] https://gitlab.com/ExaNeSt/insee

and photonic networks. The section concludes with some general conclusions remarking the configuration parameters that most affected the obtained performance results.

## 5.1 | Synthetic traffic

First we analyze the execution time obtained with synthetic traffic using the three aforementioned topologies with the four channel splitting techniques described in Table 1 . From the results, depicted in Figure 3 , it is clear that the most important parameter affecting the performance is the number of channels used to split the photonic link. In fact, a closer look to the results, shows an almost linear relation between the number of channels and the execution time. Each time we reduce (by half) the number of channels the execution time is also reduced (almost by half). The rationale behind this behavior is that photonic channels are underutilized and assigning the maximum bandwidth to each transmission is the most effective way to take advantage of the high bandwidth provided by the photonic interconnect.

Regarding the message size, the results clearly show that the longer the message the shorter the execution time. The reason for this behavior is the overhead in a photonic network significantly rises with the reservation of the paths, hence it reduces when long messages (i.e. less amount) are used. Surprisingly, the network topology employed does not affect the results. This is clearly an indication that the network is underused due to the high amount of bandwidth provided by the photonic interconnect.

As shown, the best performance is achieved using the photonic configuration with fewer number of channels. Figure 4 depicts the average time required by a packet to travel from source to destination (injection + transmission). As expected, results show that the 5-channel configuration is the most efficient for all topologies and packet sizes due to the higher bandwidth utilization. Furthermore, we can also see that as the message size increases, the average time required to send a packet is slightly decreased (notice that when all packets are considered this increment has a high impact on the performance). This overhead, depicted in Figure 5 , is caused by the path reservation stage (i.e. injection delay) which is up to a 30% of the total time for 1KB messages and just a 0.3% for 128 KB messages.

The injection delays shown in Figure 5 rise as a consequence of the reservation of the channel to perform the transmission. As photonic interconnects use circuit switching, the network is congestion-free, meaning that no interference will occur once a packet is injected. In order to analyze the impact of the network topology on the time required to reserve a channel, Figures 6 and 7 represent the number of retries (failed attempts to establish a route for a given message). Results for the 5-channel photonic configuration are only shown because no retries appear with a greater number of channels. Figure 6 plots the total number of retries. If we focus on the fat-tree topology we can see that the number of retries is negligible regardless of the packet size. The reason is the high number of alternative paths provided by this network that makes the reservation of photonic paths successful almost always. In case of the torus topology the number of retries slightly decreases as the size of the packet increases. The reason for this behavior is the use of static routing, that even when we reduce the number of packets sent, these will use the same set of paths to reach destination. A complete different picture happens for the dragonfly topology in which the increase of the packet size reduces remarkably the number of retries. In this case, reducing the number of packets sent also reduces the utilization of links that connect the local group with the intermediate proxies mainly due to Valiant routing, thus reducing the number of failures to reserve a photonic path.

This reduction of the number of retries was expected because we are sending the same amount of data using longer packets. In order to check the real impact of the use of longer messages we represent in Figure 7 the number of retries per packet. In this case,it can be observed that the number of retries increases with the packet length. This is the expected behavior because the transmission of the packets will require a longer time, thus maintaining the photonics channels occupied for longer. As a result, subsequent transmission attempts will be delayed more often. In any case, we have to consider that, although the maximum number of retries per packet is obtained with the longest packets, the overhead introduced is negligible with respect to the overall transmission time (see Figure 5 ).

## 5.2 | Real application traffic

So far we have analyzed the photonic configurations using synthetic traffic. However, this kind of traffic does not represent the interactions that occur in real applications such as the causality between messages. For that reason, now we evaluate the performance that real parallel applications can achieve when executed over photonic interconnects. Furthermore, we also compare the performance of those applications when executed in traditional electrical networks. Results are depicted in Figure 8 . For the sake of clarity, due to the huge differences in terms of execution time, we represent the results using two axes (the left one for electrical networks and the right one for photonic configurations).

First let us analyze the time required to execute both applications in the electrical network. The results clearly show the impact of the topology on the execution time. We would like to remark the higher performance achieved by *Gadget* when executed in torus topology due to the match between the

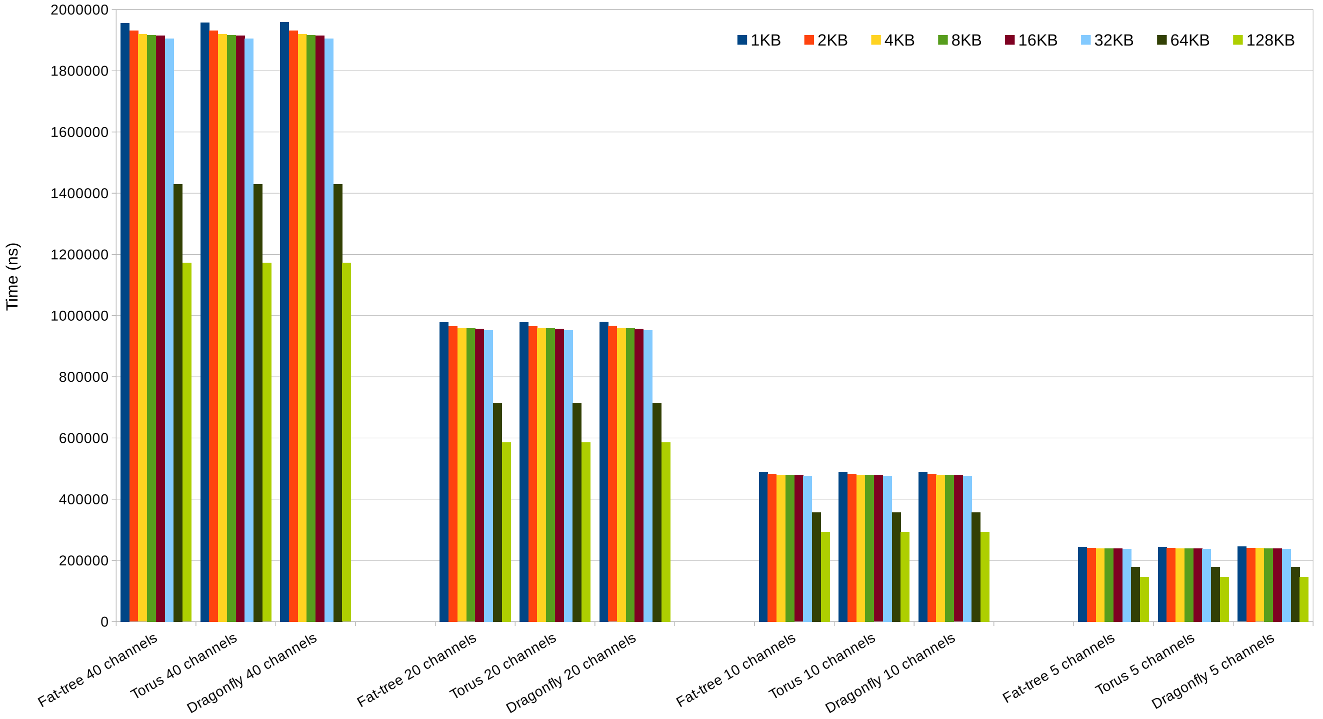


**FIGURE 3** Execution time (in ns) sending 10 GB of synthetic traffic using 8 packet lengths from 1 KB to 128 KB in the studied network topologies. Photonic links are configured using 5, 10, 20 and 40 channels.

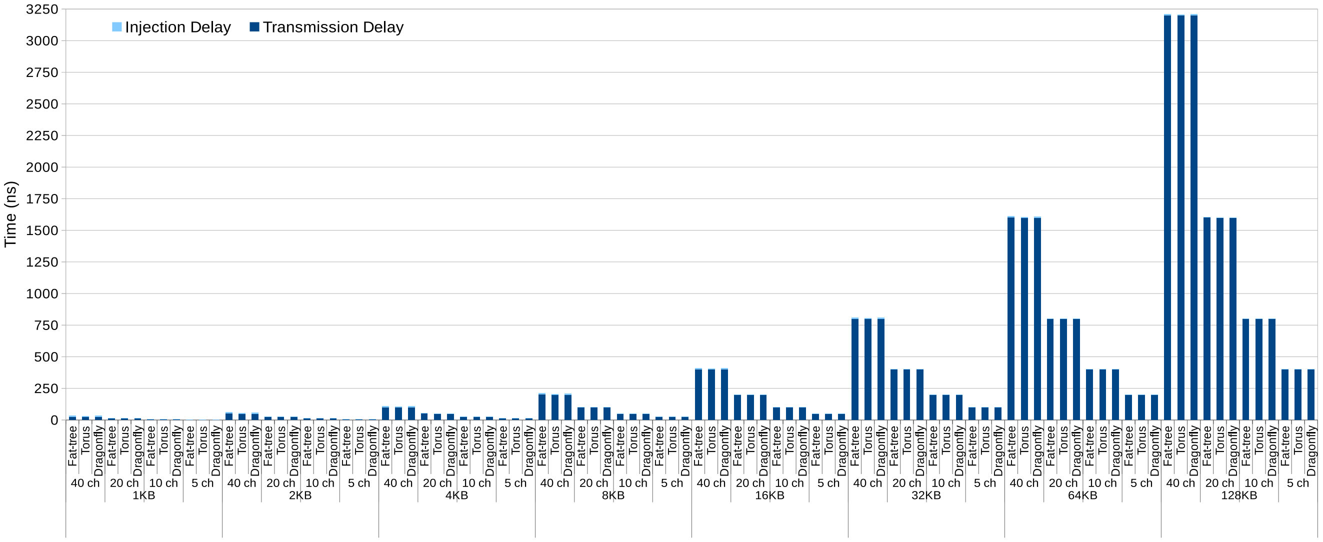


**FIGURE 4** Average sending time (in ns) per packet showing the injection delay and transmission time using 8 packet lengths from 1 KB to 128 KB in the studied network topologies. Photonic links are configured using 5, 10, 20 and 40 channels.

communication pattern of the application and the underlying physical arrangement of the nodes. When we compare these results with those obtained using photonic configurations, we can see that, as with synthetic traffic, the network topology does not impact on the performance. However, if we compare both network technologies, we can see that the applications executed in photonic configurations deliver executions times around one order of magnitude lower.

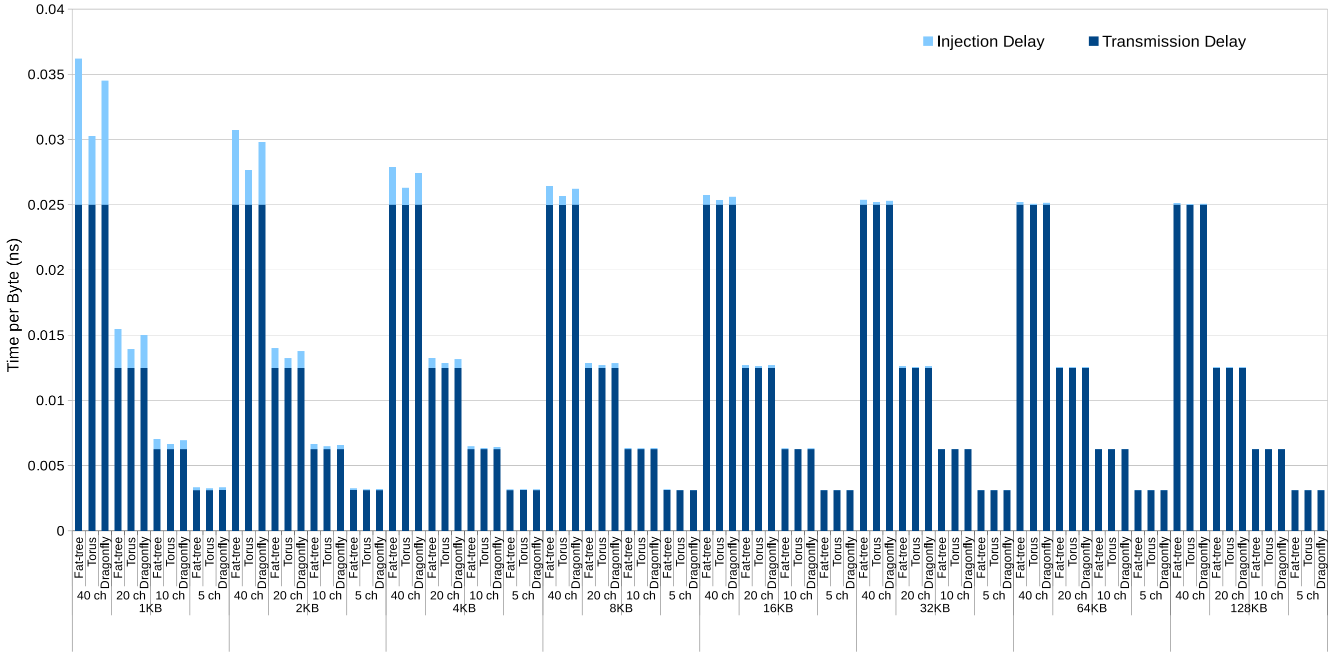


**FIGURE 5** Average sending time (in ns) per byte showing the injection delay and transmission time using 8 packet lengths from 1 KB to 128 KB in the studied network topologies. Photonic links are configured using 5, 10, 20 and 40 channels.

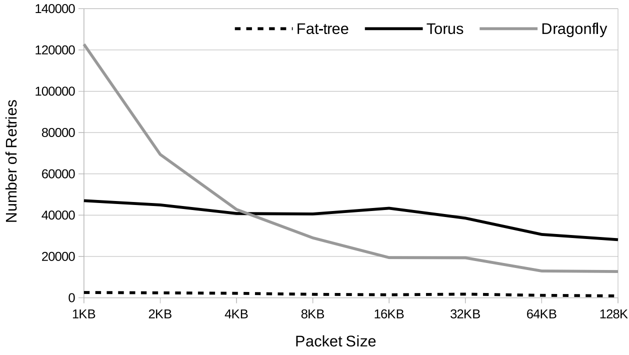


**FIGURE 6** Retries to establish the photonic route with 5 photonic channels.

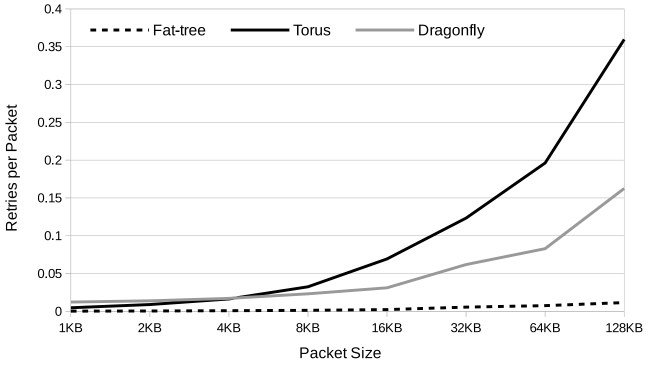


**FIGURE 7** Retries per packet to establish the photonic route with 5 photonic channels.

Regarding the photonic configurations, we can see that the less the number of photonic channels the higher the performance, especially for Gadget. These results corroborate the findings achieved with synthetic traffic, clearly showing that the assigned bandwidth is key to take advantage of the photonic interconnect.

## 5.3 | Final remarks

This section summarizes the finding using both synthetic and applications traffic. The results using the photonic interconnect has shown that the most important performance factor is the number of channels used to split the photonic link. This implies that assigning more bandwidth to the channels makes applications perform better. The downsize of this conclusion reveals that current topologies/routings and channel reservation strategies are not able to take advantage of the high

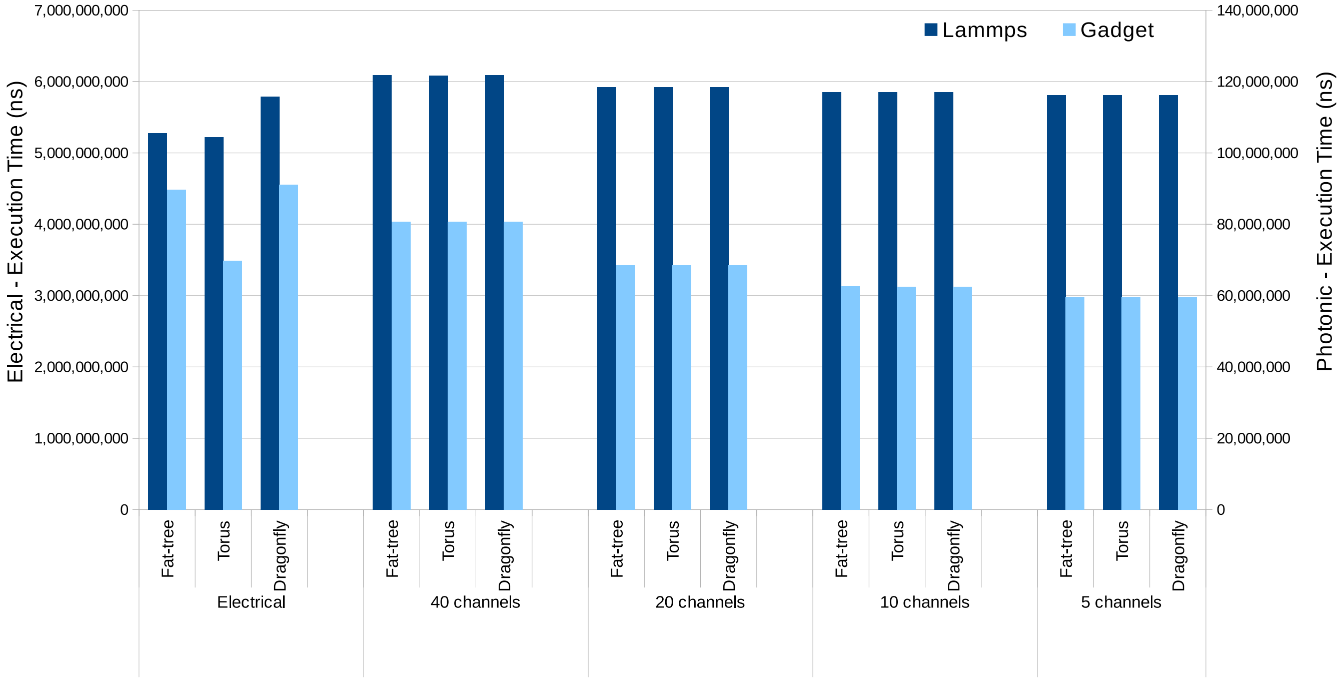


**FIGURE 8** Execution time (in ns) for 10 Gigabit Ethernet (electrical) and photonic network configured using 5, 10, 20 and 40 channels with two ExaNeSt traces.

amount of resources (i.e. bandwidth) provided by a photonic interconnect.

On the other hand, we have also seen that the use of long messages makes applications perform better. In photonic interconnects, due to the use of circuit switching, once the sequence of photonic channels has been reserved the transmission is performed without any interference. For this reason, the main performance drawback occurs in the injection phase when there is not available channel to perform the transmission, situation that is less frequent with long packets.

Regarding the performance delivered by the electrical and photonic interconnects, it is clear that the new technology outperforms, around one order of magnitude in terms of execution time, to the traditional electrical version. These results were expected due to the higher bandwidth provided by the photonic technology.

## 6 | RELATED WORK

The interest of the academia and industry communities in the development of Exascale systems, and in improving on-chip architectures, has fostered the research on photonics technology. In order to study these systems, novel simulation and estimation tools are required.

Current network simulators (29, 30, 31) focus on packet-switching electrical networks. These tools can be easily adapted to model packet-switching hybrid electro-optical networks. However, in order to adapt them to model the circuit switching capabilities of all-optical networks, a significant amount of programming effort is required. Due to this fact, some tools have been recently proposed designed from the ground up to support all-optical networks. In this regard, a well known simulator is PhoenixSim (32, 33). This framework models multiprocessor systems that use electrical networks, optical networks, and hybrid networks. PhoenixSim is based on the OMNeT++ simulation environment (34) and allows the analysis of interconnection networks from both the physical level (e.g. optical insertion loss, crosstalk, energy dissipation) and the system level (e.g. latency, performance, execution time).

The Design Space Exploration of Networks Tool (DSENT) (35) improves the PhoenixSim model of electro-optical interface circuitry such as modulators, receivers, and thermal tuning, capturing trade-offs among photonic devices and modulator/receiver specifications that can be exploited to reach optimal configurations in terms of area and power. DSENT is designed to enable fast area and power evaluation of multiple optical network configurations and, when coupled with an architectural simulator, to obtain power and area estimations for the simulated network. However, DSENT does not model photonic switches so it cannot be used to simulate circuit-switched networks such as the evaluated in this work. In addition, DSENT does not support traffic patterns and workload traces, so it cannot provide the details of a system-level simulation.

LioeSim (36) is a electrical and optical network simulator that uses Orion (37) for the models of electrical routers and links. Unlike DSENT, it models photonic switches and allows analyzing both physical level (optical insertion loss, crosstalk, optical power budget, energy dissipation) and system level (latency, energy delay product) performance metrics of interconnection networks. Unfortunately, LioeSim is focused on on-chip networks. In contrast, in this work we simulate off-chip all-optical networks for intra-rack and inter-rack communications.

Finally, there is also a need for aiding designers in layout tasks such as visually placing photonic devices, connecting waveguides, etc. To this end, in (38), Hendry et al. introduce the Visual Automated Nanophotonic Design And Layout (VANDAL), which also can be interfaced with industry-standard software tools for chip fabrication processes.

## 7 | CONCLUSIONS AND FUTURE WORK

Photonic networks simulation tools are required to guide designers in decision making when designing future Exascale supercomputers. In this work we identified the key components and discussed technologies to build a simulator featuring only optical (e.g. WDM with TWC) interconnect. We have implemented these features as an extension of an existing simulation framework.

This framework has been used to carry out an extensive simulation study to check the impact on performance of a wide set of photonic networks. For this purpose three main design parameters have been explored, the amount of wavelengths, the number of possible channels and the bandwidth per channel. These results have been compared against a traditional electrical network. To this end, three well-known topologies (3D torus, fat-tree and dragonfly) have been used.

Experimental results, obtained with synthetic traffic and excerpts of real applications from the ExaNeSt project, show that the optical network configuration has a great impact on the execution time of the applications, even when the same optical network technology is used (i.e. an aggregated bandwidth of 1.6 Tbps). In general, the parameter that impacts on performance the most is the bandwidth per channel, achieving the best results with 320 Gbps per channel (i.e. 5 channels). In addition, we also found that applications using long messages require up to 40% less time to deliver the same amount of data than using shorter messages. Regarding the performance achieved using electrical networks with real applications traffic, the use of photonic interconnects greatly reduces the execution time, one order of magnitude for the configurations evaluated in this work.

A surprising result from this work was the lack of impact on the results of the network topology, which mainly rises due to the underutilization of the photonic network capacity which leaves open a possible future line of work. Traditional network topologies are not design to deal with the technologies required to implement photonic interconnects. For this reason specific topologies should be designed to take advantage of the characteristics of such network such as the high bandwidth and the lack of congestion. In addition, new strategies to dynamically select the number of photonic channels that need an specific message must be developed. We also plan to analyze the power consumption of photonic network configurations and compare them with traditional electrical versions.

## References

[1] *Top500 Website.* 2018, Jan.

[2] Kodi Avinash Karanth, Neel Brian, Brantley William C. Photonic interconnects for exascale and datacenter architectures. *IEEE Micro.* 2014;34(5):18–30.

[3] Rumley Sébastien, Nikolova Dessislava, Hendry Robert, Li Qi, Calhoun David, Bergman Keren. Silicon photonics for exascale systems. *Journal of Lightwave Technology.* 2015;33(3):547–562.

[4] Shacham Assaf, Bergman Keren, Carloni Luca P. Photonic networks-on-chip for future generations of chip multiprocessors. *IEEE Transactions on Computers.* 2008;57(9):1246–1260.

[5] Batten Christopher, Joshi Ajay, Orcutt Jason, et al. Building many-core processor-to-DRAM networks with monolithic CMOS silicon photonics. *IEEE Micro.* 2009;29(4).

[6] Werner Sebastian, Navaridas Javier, Lujan Mikel. Designing Low-power, Low-latency Networks-on-chip by Optimally Combining Electrical and Optical Links. *The 23rd IEEE Symposium on High Performance Computer Architecture.* 2016;.

[7] Puche José, Lechago Sergio, Petit Salvador, Gómez María E, Sahuquillo Julio. Accurately modeling a photonic NoC in a detailed CMP simulation framework. *High Performance Computing & Simulation (HPCS), 2016 International Conference on.* 2016;:387–394.

[8] Chen Guoqing, Chen Hui, Haurylau Mikhail, et al. On-chip copper-based vs. optical interconnects: delay uncertainty, latency, power, and bandwidth density comparative predictions. *Interconnect Technology Conference, 2006 International.* 2006;:39–41.

[9] Katevenis M, Chrysos N, Marazakis M, et al. The ExaNeSt Project: Interconnects, Storage, and Packaging for Exascale Systems. *Digital System Design (DSD), 2016 Euromicro Conference on.* 2016;:60–67.

[10] Concatto Caroline, Pascual Jose A, Navaridas Javier, et al. A CAM-Free Exascalable HPC Router for Low-Energy Communications. 2018;:99–111.

[11] Duan G. .. H., Jany C., Liepvre A. Le, et al. 10 Gb/s integrated tunable hybrid III-V/Si laser and silicon Mach-Zehnder modulator. *2012 38th European Conference and Exhibition on Optical Communications.* 2012;:1-3.

[12] Duan G. H., Jany C., Liepvre A. Le, et al. Integrated hybrid IIIâĂŞV/Si laser and transmitter. *2012 International Conference on Indium Phosphide and Related Materials.* 2012;:16-19.

[13] Soref R., Bennett B.. Electrooptical effects in silicon. *IEEE Journal of Quantum Electronics.* 1987;23(1):123-129.

[14] Liu Ansheng, Liao Ling, Rubin Doron, et al. High-speed optical modulation based on carrier depletion in a silicon waveguide. *Opt. Express.* 2007;15(2):660–668.

[15] Thomson D. J., Gardes F. Y., Hu Y., et al. High contrast 40Gbit/s optical modulation in silicon. *Opt. Express.* 2011;19(12):11507–11516.

[16] Bergman Keren, Carloni Luca P, Biberman Aleksandr, Chan Johnnie, Hendry Gilbert. *Photonic network-on-chip design.* Springer; .

[17] Dong Po, Chen Long, Xie Chongjin, Buhl Lawrence L., Chen Young-Kai. 50-Gb/s silicon quadrature phase-shift keying modulator. *Opt. Express.* 2012;20(19):21181–21186.

[18] Dong Po, Liu Xiang, Sethumadhavan Chandrasekhar, et al. 224-Gb/s PDM-16-QAM Modulator and Receiver based on Silicon Photonic Integrated Circuits. *Optical Fiber Communication Conference/National Fiber Optic Engineers Conference 2013.* 2013;:PDP5C.6.

[19] Navaridas Javier, Miguel-Alonso Jose, A. Jose, Ridruejo Francisco J.. Simulating and evaluating interconnection networks with {INSEE}. *Simulation Modelling Practice and Theory.* 2011;19(1):494 - 515. Modeling and Performance Analysis of Networking and Collaborative Systems.

[20] Lu Liangjun, Zhao Shuoyi, Zhou Linjie, et al. 16 x 16 Non-blocking silicon optical switch based on electro-optic Mach-Zehnder interferometers. *Optics Express.* 2016;24:9295.

[21] Duro José, Petit Salvador, Sahuquillo Julio, Gómez María E. Modeling a Photonic Network for Exascale Computing. *High Performance Computing & Simulation (HPCS), 2017 International Conference on.* 2017;:511–518.

[22] Xi Kang, Kao Yu-Hsiang, Chao H Jonathan. A petabit bufferless optical switch for data center networks. In: Springer 2013 (pp. 135–154).

[23] Kim John, Dally William J., Scott Steve, Abts Dennis. Technology-Driven, Highly-Scalable Dragonfly Topology. *35th International Symposium on Computer Architecture (ISCA 2008), June 21-25, 2008, Beijing, China.* 2008;:77–88.

[24] Alwayn Vivek. *Optical network design and implementation.* Cisco Press; 2004.

[25] Essiambre René-Jean, Tkach Robert W. Capacity trends and limits of optical communication networks. *Proceedings of the IEEE.* 2012;100(5):1035–1055.

[26] Temprana E, Myslivets E, Kuo BP-P, et al. Overcoming Kerr-induced capacity limit in optical fiber transmission. *Science.* 2015;348(6242):1445–1448.

[27] Springel Volker. The cosmological simulation code GADGET-2. *Monthly notices of the royal astronomical society.* 2005;364(4):1105–1134.

[28] Plimpton Steve. Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics.* 1995;117(1):1–19.

[29] Ben-Itzhak Yaniv, Zahavi Eitan, Cidon Israel, Kolodny Avinoam. Hnocs: Modular open-source simulator for heterogeneous nocs. *Embedded Computer Systems (SAMOS), 2012 International Conference on.* 2012;:51–57.

[30] Hossain Hemayet, Ahmed Mostak, Al-Nayeem Abdullah, Islam Tanzima Zerin, Akbar Md Mostofa. Gpnocsim-a general purpose simulator for network-on-chip. *Information and Communication Technology, 2007. ICICT'07. International Conference on.* 2007;:254–257.

[31] Jain Lavina, Al-Hashimi BM, Gaur MS, Laxmi V, Narayanan A. NIRGAM: a simulator for NoC interconnect routing and application modeling. *Design, Automation and Test in Europe Conference.* 2007;:16–20.

[32] Chan Johnnie, Hendry Gilbert, Biberman Aleksandr, Bergman Keren, Carloni Luca P.. PhoenixSim: A Simulator for Physical-layer Analysis of Chip-scale Photonic Interconnection Networks. *Proceedings of the Conference on Design, Automation and Test in Europe.* 2010;:691–696.

[33] Rumley Sébastien, Bahadori Meisam, Wen Ke, Nikolova Dessislava, Bergman Keren. Phoenixsim: Crosslayer design and modeling of silicon photonic interconnects. *Proceedings of the 1st International Workshop on Advanced Interconnect Solutions and Technologies for Emerging Computing Systems.* 2016;:7.

[34] Varga András, Hornig Rudolf. An Overview of the OMNeT++ Simulation Environment. *Proceedings of the 1st International Conference on Simulation Tools and Techniques for Communications, Networks and Systems & Workshops.* 2008;:60:1–60:10.

[35] Sun Chen, Chen Chia-Hsin Owen, Kurian George, et al. DSENT-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling. *Networks on Chip (NoCS), 2012 Sixth IEEE/ACM International Symposium on.* 2012;:201–210.

[36] Ma X., Yu J., Hua X., et al. LioeSim: A Network Simulator for Hybrid Opto-Electronic Networks-on-Chip Analysis. *Journal of Lightwave Technology.* 2014;32(22):4301-4310.

[37] Kahng Andrew B, Li Bin, Peh Li-Shiuan, Samadi Kambiz. ORION 2.0: a fast and accurate NoC power and area model for early-stage design space exploration. *Proceedings of the conference on Design, Automation and Test in Europe.* 2009;:423–428.

[38] Chan J., Hendry G., Bergman K., Carloni L. P.. Physical-Layer Modeling and System-Level Design of Chip-Scale Photonic Interconnection Networks. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.* 2011;30(10):1507-1520.

**How cite this article:** José Duro, Jose A. Pascual, Salvador Petit, Julio Sahuquillo and María E. Gómez (2017), Modeling a Photonic Network for Exascale Computing, , *2017;00:1–6*.