Design Exploration of Optical Interconnection Networks for Chip Multiprocessors

Michele Petracca*
Dept. of Computer Science
Columbia University
petracca@cs.columbia.edu

Benjamin G. Lee Keren Bergman
Dept. of Electrical Engineering
Columbia University
{benlee,bergman}@ee.columbia.edu

Luca P. Carloni
Dept. of Computer Science
Columbia University
luca@cs.columbia.edu

Abstract

The Network-on-Chip (NoC) paradigm has emerged as a promising solution for providing connectivity among the increasing number of cores that get integrated into both systems-on-chip (SoC) and chip multiprocessors (CMP). In future high-performance CMPs, however, the high bandwidth requirements will not be adequately provided by electronic NoCs without dissipating large amounts of power. Previously, we have made the case for the photonic NoC as a unique interconnect solution for delivering scalable bandwidth-per-watt *performance that surpasses equivalent electronic NoCs. Building on this work, we study the adoption of photonic communication for CMPs and we present three main contributions: (1) we propose two non-blocking topologies for photonic NoC designs and we assess both qualitatively and quantitatively the pros and cons that they offer with respect to the original (blocking) topology, (2) we show how a photonic NoC is better suited for a CMP made of complex multi-threaded cores, and (3) we present the first simulation-based assessment of the benefits of using a photonic NoC for a real application, i.e. computing a large FFT.*

1 Introduction

The new trend of integrating an increasing number of processing cores into a single die raises the importance of designing an efficient communication infrastructure among them. Consequently, substantial research has recently focused on packet-switched networks-on-chip (NoC) designs for both general purpose chip multiprocessors (CMP) and application-specific systems-on-chip (SoC) [2, 5, 15, 22]. Many studies have been presented on the optimization of the NoC bandwidth and latency, which directly impact the system application performance. However, since packaging constraints will continue to impose strong limitations on the maximum on-chip temperature for the foreseeable future, the analysis and optimization of the NoC power dissipation becomes increasingly important as the number of cores on the chip grows [7]. If fact, current prototypes of future CMPs with tens of cores show that the power dissipated by the NoC accounts for over 25% of the overall power [16]. Moreover the power of a NoC implemented with current circuit techniques is estimated to be too high (by a factor of 10) to meet the expected needs of future CMPs [14]. Consequently, the limited on-chip power budget will have to be carefully distributed between computation and communication activities. Clearly, a reduction of the power dissipated by the NoC enables a larger portion of the limited power budget to be devoted to the cores, which directly improves the performance-per-watt of the overall system.

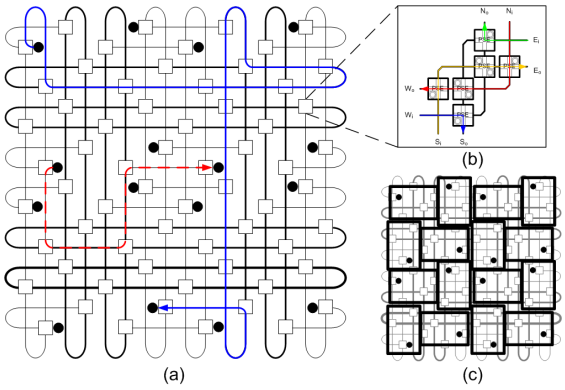


Figure 1. 16-core *Blocking Mesh* photonic NoC from [9]: (a) shortest (longest) path is marked dashed (solid); (b) basic non-blocking switch from [10]; (c) core layout over the NoC.

In this context, photonic communication holds the promise of providing a mechanism for both intra-chip and inter-chip large data transfers with minimal power dissipation. In particular, a NoC with photonic communication links offers two main advantages:

1) the achievable communication bandwidth on a single waveguide (or link) can approach multiple terabits-per-second with limited power dissipation;

2) the power dissipation to first order is independent of the distance covered by the optical signal across the system and scales only with the link transmission interface circuitry (modulators, drivers and receivers).

*M. Petracca is also with the Dipartimento di Elettronica, Politecnico di Torino, Italy.

1550-4794/08 $25.00 © 2008 IEEE
DOI 10.1109/HOTI.2008.20

The effective lack of optical memories or equivalent optical RAM and the impracticality of processing directly in the optical domain will force designers to combine photonic communication with electronic computation. However, while the integration of optical devices on a chip still presents many difficult challenges, remarkable breakthroughs have been made in the field of CMOS-compatible silicon photonics in recent years [6, 17–19, 24–30, 32, 34–38]. In particular, an innovative small-footprint device based on *silicon micro-ring resonator* has been shown to deliver ultrafast data modulation [28,37,38] and switching capabilities [26,34] with limited power dissipation.

By relying on this result, we have proposed an innovative photonic NoC for CMPs that is based on a hybrid approach: a high-bandwidth circuit-switched photonic network is combined with a low-bandwidth packet-switched electronic network [9]. While the electronic network carries small-size control (and data) *packets*, the photonic network transfers large-size data *messages* between pairs of cores. The NoC operates as follows: (1) a photonic circuit is reserved through the exchange of a *path-setup* packet over the electronic network between the source and the destination, followed by a short *Ack* pulse over the photonic network (*path-setup process*); (2) a large data transfer is completed on the photonic circuit, which offers up to 960*Gbps* of photonic transmission *line rate* per core by combining time-division and wavelength-division multiplexing (TDM-WDM), and (3) at the end of the communication the photonic circuit is released by the source through the transmission of a *tear-down* packet (*path-teardown process*). The physical implementation and performance of this photonic NoC is discussed in [10], and its power consumption is analyzed in [8]. Its main organization for the case of a 16-core CMP is illustrated in Fig. 1(a): the black circles represent the cores' network interfaces (*gateways*) and the white squares represent the *photonic switches*. As shown in Fig. 1(b), each switch is composed of a set of *Photonic Switching Elements* (PSE) and one *Electronic Router* (ER). A PSE is a single (or double) silicon micro-ring resonator element that can deflect/pass the light according to its polarization. The ER not only switches the electronic packets but also polarizes the PSEs based on the value of the control packets that are exchanged during the setup and tear-down processes. The bold lines in Fig. 1(a) represent the *transport matrix* of switches, while each switch that is not placed on a junction between a bold column and a bold row regulates the messages injection/ejection from/into the *gateway* of a core into/from the NoC. The NoC contains four kinds of switches:
- a *gateway switch* connects the gateway to the NoC;
- an *injection switch* deflects the traffic from a gateway switch into the NoC row;
- an *ejection switch* deflects the traffic from the NoC column into the gateway switch;
- a *transport switch* forwards the traffic over the transport matrix.

The injection/ejection switches require a smaller number of PSEs than the other switches.

The injection/ejection policy allows every core to access the network, but the network itself cannot simultaneously sustain all possible communications among distinct cores due to the internal congestion that can occur during the set-up of the photonic paths. In fact, the network proposed in [9] has a blocking topology and therefore offers limited connectivity. Blocked communication flows must be delayed until an open path is available resulting in some degradation to the network throughput and message latency. As shown in [10] it is possible to reduce the blocking probability, and consequently improve the NoC performance, by over-provisioning the network. This over-provisioning is accomplished by increasing the number of rows and columns of the transport matrix while keeping unchanged the number of cores, so that more paths are available for each source-destination couple. In [9] the best over-provisioning trade-off is obtained by doubling the number of rows and columns. For a 36-core network, this results in a 18×18 mesh of switches, including those necessary for injection and ejection. In the sequel we refer to this topology as the *Blocking Mesh*.

We consider here the combined layout of the CMP and their supporting NoC. In order to optimize the fabrication process, it is reasonable to expect that the NoC's photonic devices and the CMP cores will be located on different planes by taking advantage of progress in *3D Integration* (3DI) [20]. In Fig. 1(c) we show a possible layout for a 16-core CMP that we derived by assuming: (1) that all cores are identical, (2) that the network interface of each core must match the assigned position on the network plane, which is dictated by the network topology, and (3) that only 90°-multiple rotations and vertical/horizontal flips of the cores are allowed. The same assumptions are made to derive possible layouts for the other NoC topologies that we propose in the following pages.

In this paper we advance the idea of using silicon photonics to address the on-chip communication requirements of future high-performance CMPs. First, we analyze the physical layer of our system by discussing the most recent advances in silicon photonic device integration. Then, we consider two new alternative implementations based on non-blocking topologies and we complete a comparative analysis with respect to the Blocking Mesh from the performance and power viewpoints. Finally, we present a case study where

we analyzed the total execution time and power dissipation necessary to compute a double-precision 2^{29}-element Fast Fourier Transform (FFT) on a hypothetical 36-core CMP fabricated in a future $22nm$ technology process.

2 Physical device progress

Before continuing with the description of the model and the discussion of the topologies, we devote a short section to the review of the photonic device building blocks which will be required in order to realize such a network and the associated recent advancements in the field of silicon photonics. These components comprise (1) optical modulators, at the beginning of the photonic pathway, alongside (2) high-bandwidth links, (3) routing switches, and (4) optical receivers. Other devices, such as lasers, amplifiers, and (de)multiplexers could be used in certain implementations, but are not required in every case, and so will not be discussed here despite significant progress [1].

In the past year, since Xu *et al.* reported the successful operation of silicon micro-ring resonator modulators operating at $12.5Gbps$ [37], a number of advancements have occurred. These easily cascadable devices, as shown in [38], were demonstrated in parallel creating an excellent form factor match between multilane electronic busses and wavelength-parallel photonic links [28]. Here, four rings were used to modulate four separate lightwaves of different wavelengths copropagating along the same photonic link. Each modulator was driven simultaneously at $4Gbps$ with decorrelated pseudo-random data, while error-free operation (Bit-Error Rates [BERs] below 10^{-12}) was observed. Additionally, a cascade of five micro-rings with radii of $1.5\mu m$ (a reduction of more than a factor of 3 from the previous devices) was reported in [36]. Moreover, the electronic structure of these devices still has room for improvement. Manipatruni *et al.* proposed an alteration in the doping configuration which is expected to reduce the carrier injection and extraction times, providing 40-$Gbps$ modulation using a single ring [30]. Alternatively, other silicon optical modulators based on the Mach-Zehnder interferometer (MZI), rather than the ring resonator, have also seen improvement. After a 30-$Gbps$ silicon modulator was implemented in [29] using a somewhat lengthy and power hungry MZI, Green *et al.* reduced the length requirements to $100\mu m$ and $200\mu m$ for $5Gbps$ and $10Gbps$, respectively, and the power requirements to $5pJ/bit$ at $10Gbps$ [18]. Although the power and footprint are still not as attractive as ring-based modulators, which can theoretically operate with less than $0.1pJ/bit$ occupying areas less than $10\mu m^2$, the newly reported MZI provides significant improvement in thermal stability and presents a plausible alternative to micro-rings if further gains are realized in the future.

[1] As discussed in [10], we advocate using off-chip laser sources to further minimize the on-chip power requirements of the photonic network.

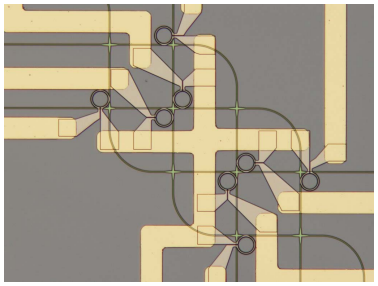


Figure 2. Fabricated non-blocking 4-way silicon photonic switch with individual tuners on each ring (image courtesy of Michal Lipson).

Waveguide losses below $4dB/cm$ and coupling losses below $0.5dB$ per facet [35] have been reported for some time now in silicon waveguides. Although these values are more than sufficient for chip-scale links, improvements continue to be made with some links currently reporting less than $1dB/cm$ [24]. Furthermore, a recent demonstration has shown the ability to propagate over $1Tbps$ of optical data through a 5-cm photonic link while performing system-level measurements on the received data. [27].

Perhaps the most explosive progress in the past year has occurred in the proposed micro-ring routing switches. The first broadband multiple-wavelength silicon micro-ring switch was demonstrated in [6], and soon thereafter, BER measurements were reported on actively switched optical messages at near-GHz switching speeds using the same device [26]. Furthermore, Lee *et al.* demonstrated the ability of the device to switch 20 wavelength channels at once. Building on a similar design, Vlasov *et al.* demonstrated a fifth-order resonator device, which operates on up to 9 wavelengths with resonator passbands wide enough to allow each wavelength to be modulated at $40Gbps$ [34]. Although the above switches are all 1×2 devices, representing the functionality of half of each of the generic PSE blocks in Fig. 1(b), more complex structures are being implemented as well. A structure comprising the entire photonics of the 4-way non-blocking switch in Fig. 1(b) has been fabricated by the research group headed by Professor Michal Lipson at Cornell University (microscope image shown in Fig. 2).

Silicon receivers employing SiGe or Ge photodiodes in order to realize optical absorption have been monolithically integrated with CMOS amplifier circuitry to

achieve a wide variety of results targeted for different applications requiring both low sensitivities and large RF bandwidths. In [25], a sensitivity of $-7.4dBm$ at a BER of 10^{-12} was achieved for a bit rate of $15Gbps$, requiring $7pJ/bit$ for the entire receiver. The work reports other configurations resulting in receiver power consumption as low as $1.1pJ/bit$. Alternatively, recent developments have shown that silicon with damaged crystalline structure may be employed to achieve the necessary absorption, avoiding the use of high concentrations of Ge in CMOS process lines. In this new research, RF bandwidths between 10 and $20GHz$ have already been demonstrated [17]. Finally, efforts to combine transmitters and receivers together forming complete optical links have been fruitful as well. In [32], 16 links enabling a system bandwidth of $160Gbps$ were demonstrated with a total link power dissipation of $4.65pJ/bit$ in a 130-nm CMOS process.

3 Multi-thread core model

The introduction of photonic NoCs aims at providing high-bandwidth low-latency communication channels for large data transfers between cores. A single core can be a multi-threaded processor, where many threads are executed in parallel and each thread can independently request a data transfer to another core. Fig. 3 illustrates the core model that we developed for our simulator. It consists of three main blocks:

• The *Traffic Generator* simulates the behavior of the core threads that request data transfers during their processing time. The number of threads per core is a simulation parameter. Each thread can request one connection at a time so that the number of simultaneous requests from a core never exceeds its number of threads. Communication requests are generated according to a Poisson process with uniformly-distributed destinations. The message length can be fixed to emulate constant size transfers, or randomly set with an exponential probability distribution. The requests are stored into a finite-size back-pressuring FIFO queue.

• The *Scheduler* extracts the requests from the FIFO to generate the relative path-setup packets and attempts to inject/eject packets into/from the network through the *Electronic Network Interface*. Blocked requests are re-enqueued into the FIFO. The main goal of the scheduler is to avoid head-of-line (HoL) blocking, a well-known problem in switching networks. Since communication requests are generated independently from the network status and enqueued into the FIFO, it is possible that the request at the head of the FIFO cannot be served immediately because of an internal block of the network (or, simply, because the destination node is already busy receiving another communi-

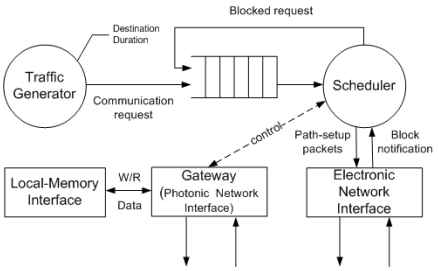


Figure 3. Model of multi-threaded processing core.

cation). Therefore, it must wait for the resolution of the congestion together with all the following enqueued packets, even though there may be good chances to setup successfully a connection for one of these. To avoid HoL blocking, a *Scheduler* monitors a window of packets at the head of the FIFO and attempts to establish the connection for one of them based on their arrival time. After a successful communication is completed the procedure is restarted from the oldest packet.

• The *Gateway* is simply the *Photonic Network Interface*, that is able to send/receive photonic messages to/from the NoC by reading/writing the data from/into the *Local-Memory Interface*.

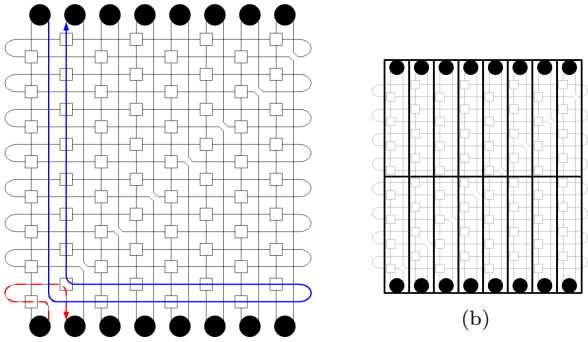
4 Non-Blocking Topologies

In this section we discuss the pros and cons of using non-blocking topologies for photonic NoC from a conceptual viewpoint. In the next section we present quantitative analysis based on experimental results.

A strictly non-blocking network is able to simultaneously handle the maximum number of connections. We propose two non-blocking topologies: a *Crossbar* and a *Non-Blocking Mesh*. Both topologies are strictly non-blocking with a $O(N^2)$ complexity in terms of the number of switches, where N is the number of cores. We considered also Clos-like topologies, that are always strictly non-blocking with complexity $O(N \log N)$, but it is not easy to effectively employ them under the layout constraints imposed by the need to uniformly distribute the cores over the chip area.

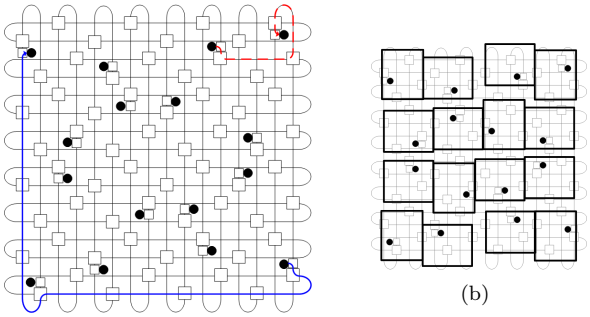
Crossbar. Fig. 4(a) illustrates a *Crossbar* for a 16-core CMP: the black circles represent the core gateways and the white boxes represent the switches. The switches are organized in a 8×8 matrix and connected by bidirectional links. For figure clarity, 16-core CMPs are considered for the topology pictures while the performance analysis is performed over 36-core CMPs.

The switches on the diagonal are the input switches for the gateways. Each pair of gateways that face each other on the same column share the same row for injection and the same column for ejection, thereby exploiting the bi-directionality of the 4×4 switches. Once



(a) (b)

Figure 4. (a) 16-core *Crossbar*; (b) core layout over the NoC.



(a) (b)

Figure 5. (a) 16-core *Non-Blocking Mesh*; (b) core layout over the NoC.

a packet is injected into the row, it passes straight through all the switches on that row until it reaches the column of the destination gateway. The switch located at the crossing of the "transmitter row" and the "receiver column" deflects the message from the row to the column. Then, the message proceeds by going straight through the column until the destination gateway. Hence, only one turn is needed to reach a destination from any transmitting gateway.

This topology uses a simple 4×4 internally-blocking switch as proposed in [9]. Since only horizontal-to-vertical turns are possible, it is not necessary to use (more complex) non-blocking switches. In fact, simpler switches with just 4 ring resonators are sufficient.

Crossbars have limited scalability both in terms of resources needed to build the network and in terms of maximum (and average) path length. The maximum path length has an impact on the maximum attenuation experienced by the photonic signal. This must be taken into account while designing the optical output power for the lasers and the sensitivity of the optical receivers. The average path length affects the average duration of the path-setup process, thus impacting the average throughput and latency performance.

Another critical issue is the electronic connection between each gateway and its own injection switch. The length of these metal wires can affect the performance from the viewpoint of both power dissipation and path-setup overhead, because of the energy/time needed by the electronic signal to propagate.

Non-Blocking Mesh. A *Non-Blocking Mesh* can be built by using the non-blocking switches of Fig. 1(b) and simplifying the injection/ejection policy. Since the links and switches are bidirectional, in order to have a non-blocking topology it is sufficient to have just two cores injecting on each row and two cores ejecting from each column. Fig. 5 shows a Non-Blocking Mesh for a 16-core CMP. The white boxes represent switches - the small ones are simpler injection/ejection 3×3 gateway switches - while the black circles represent the gateways. A gateway is connected to a gateway switch through the only horizontal port. For a 36-core network this topology consists of a 18×18 mesh of switches plus 36 gateway switches.

The implementation of this topology is based on dividing the chip into 4 quadrants. Each quadrant is a square where $N/4$ cores are placed in an interleaved fashion, so that one core is placed on each row and one core on each column. The value of $N/4$ must be a square number. The two horizontally contiguous quadrants are identical. The difference between the quadrants of the upper/bottom half is that the gateway switches are placed above/below the corresponding row, based on the injection rule. The resulting topology is a mesh with $N^2/4$ switches and N gateway switches. Fig. 5 is the folded version of this network, where every column and row has constant distance between the switches.

During the ejection, a message passing through a column can go into a gateway switch from either one of the vertical ports. If this message is for the attached core, it is deflected toward the gateway. For the injection, a packet is sent by the gateway to the gateway switch that forwards it to the closest row. Once the packet is on the row, it follows simply an XY minimum-distance routing algorithm: it reaches the right column passing through the "input" row and then reaches the destination gateway switch passing through the "output" column. This algorithm avoids the risk of blocking the core when injecting a message. This blocking condition would happen if the output port of the gateway switch that must be used to inject the packet was busy holding a connection that passes through that column.

While the complexity in terms of the asymptotic number of switches grows still as $O(N^2)$, the Non-Blocking Mesh offers a remarkable improvement with respect to the Crossbar by reducing the average path length between cores. Table 1 reports the number of

cores	Blocking Mesh switch count	Blocking Mesh avg. path	Crossbar switch count	Crossbar avg. path	Non-Blocking Mesh switch count	Non-Blocking Mesh avg. path
16	144	8	64	12	80	6
36	324	12	324	27	360	11
64	576	16	1024	48	1088	18

Table 1. Comparison of the 3 topologies as the number of cores scales.

switches and the average electronic path lengths, expressed in number of switch-to-switch hops, for the three alternative NoC topologies. For the Crossbar each long electronic connection is taken into account as an integer multiple of one hop. Even if the maximum and average paths are shorter than in the Crossbar, they still scale linearly with the number of cores. Hence, as this number grows a Non-Blocking Mesh also faces problems with optical signal integrity and the amount of necessary resources. The Blocking Mesh, instead, scales better because the number of switches grows linearly, and the path length scales as $\sqrt{N}$, where N is the number of cores. The lack of scalability impacts also the performance gain of the non-blocking networks with respect to the blocking ones. The larger is the number of switches, the higher is the time needed to set-up the path. Hence, when the network size is very large, the time needed to set-up the average path over the Non-Blocking Mesh can be several times longer than for the Blocking Mesh.

5 Experimental results

In this section, first we present a comparative analysis of the three photonic NoC topologies discussed in the previous pages. Then, we discuss a case study that shows the benefit in terms of performance-per-watt that on-chip photonic transmission can provide to a communication-intensive application.

Performance Analysis of the NoC Topologies. The *throughput-per-core* is evaluated as the ratio of the time when a core is transmitting photonic messages on the NoC over the total simulation time. This metric is a function of the average path-setup overhead, which depends on the NoC topology, and of the average duration of a photonic message, which is the ratio between the average message size and the photonic transmission line rate. The *offered load* is the ratio of the time when a core is ready to transmit at least one message and the total simulation time. In a non-congested network the throughput-per-core matches the offered load. In [10] we assumed to have 36 cores exchanging DMA transfers of fixed size, equal to $16kBytes$, with a line rate of $960Gbps$. This corresponds to a photonic message with a duration of $134ns$. Under the same assumption, we report in Fig. 6 the throughput-per-core as a function of the offered load for four distinct scenarios:

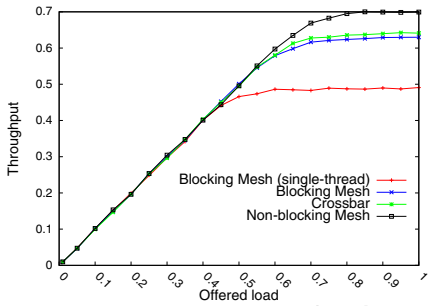


Figure 6. Throughput-per-core of various 36-core NoC topologies.

- *Blocking Mesh* with single-thread cores (as in [10]);
- *Blocking Mesh* with multi-threaded cores;
- *Crossbar* with multi-threaded cores;
- *Non-Blocking Mesh* with multi-threaded cores.

The first observation is that using multithreaded cores allows us to better exploit the high bandwidth offered by a photonic NoC, thus leading to a gain of more than 26% in throughput-per-core. In fact, whenever a path-setup request gets blocked into the NoC, a single-thread core cannot make other requests that could be more successful if addressed to other destinations located in less congested parts of the network.

Also, the analysis of the relationship among different topologies, using the same core model, surprisingly shows that the performance of the Crossbar and of the Blocking Mesh are very similar. Generally a non-blocking topology does not achieve a near-100% maximum throughput-per-core, because the overhead introduced by the path-setup process is not negligible for short-duration messages. On the other hand, by definition, a non-blocking topology guarantees the delivery of a message to every free destination. This advantage, however, gets partially neutralized in the Crossbar because the long electronic paths increase considerably the propagation time of the path-setup packet over the control network. In the Non-Blocking Mesh, instead, the distance between two gateways is comparable with the corresponding distance in the Blocking Mesh, thus leading to a throughput gain of about 13%.

The *workload* of a core is evaluated as the ratio of the time when at least one thread is active within the core over the total simulation time. As discussed in Section 3, the number of communication requests enqueued, plus the one injected into the network, cannot exceed the number of threads. When the number of outstanding requests matches the number of threads, each thread is suspended and, therefore, the core is stalled. Fig. 7 reports the workload per node as a function of the offered load for the four scenarios. While the throughput-per-core is a measure of the network resources that are effectively offered to a core, the work-

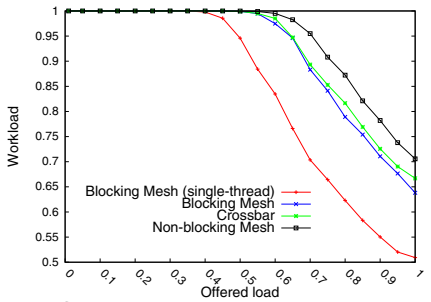


Figure 7. Core workload of various 36-core NoC topologies.

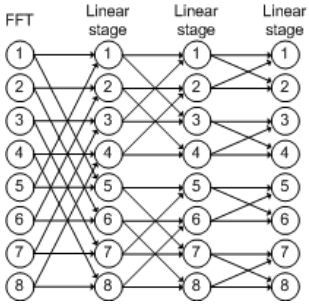


Figure 8. Butterfly scheme of Cooley-Tukey's algorithm for $M = 8$.

load is a metric of its computational efficiency that accounts for all the stalling periods due to network congestion. Observe how the network becomes the bottleneck of the system when the offered load reaches a certain threshold value, which depends on the topology and in our case is about 0.6. In other words, the threads generate more requests than what the network is able to satisfy.

When the time for a thread to generate a new request is shorter, the traffic load is higher. If the traffic load increases, however, the time that a thread must wait to see its request satisfied increases too, due to network congestion. Ultimately, this may lead to a stalling of the core and a reduced workload.

Even though these results were obtained with a simple model of stochastic uniform traffic, which assumes that all threads have the same probability to request a connection with any core, they clearly show the direct relationship between the resources offered by the NoC and the service experienced by a generic application.

Case Study: FFT Computation. The Fast Fourier Transform (FFT) is an important application that can take advantage of the large bandwidth offered by photonic on-chip communication. Using the same simulator, we analyzed the execution of the classic Cooley-Tukey FFT algorithm [4] running on 32 processing cores in a future 36-core CMP.

In the first phase, each core processes $k = m/M$ sample elements, where m is the size of the array of input samples and M is the number of cores. After this phase, the algorithm proceeds with a sequence of $\log M$ iterations. At each iteration, a computation step follows a communication step, when the processors exchange data according to a butterfly scheme. Specifically, at each iteration a core executes the following actions: first, a copy of the sub-array resulting from the previous computation is sent to another core X (a local copy is kept) while, simultaneously, another sub-array is received from core X; when both transfers are complete, the local copy and the newly-received sub-array are linearly combined. Fig. 8 illustrates the butterfly scheme for $M = 8$. At the end of all iterations, the result of the entire FFT on the original m-elements input vector is the merge of the k-elements portions of sub-arrays resulting from the local computation in each core. The time to perform the FFT is the sum of the time for the computation, which depends on the core architecture, plus the time needed to move the data among the cores. The last component depends on the line rate and on the topology. The line rate influences the message duration, while the topology influences the average number of attempts to deliver a data sub-array.

For our experiment, we first assumed to have an hypothetical CMP built in a future $22nm$ technology process with a chip size of about $625mm^2$ and, by using *3D Integration* (3DI) [20], to combine a processing core plane with an optical NoC plane and various on-chip memory planes. Under the assumption of classic scaling, in a $22nm$ we should be able to integrate 36 cores as complex as the first generation of the IBM Cell multi-core processor [12] into our CMP. Under the assumption of combining classic scaling and 3DI, we estimated that each core will have a local memory of about $0.5GBytes$ [33].

Then, we took the result presented by Chow *et al.* in [3], where a Cell processor is reported to compute a large single-precision FFT (2^{24} samples) in $43ms$ using Bailey's FFT algorithm [1]. Based on the Cell roadmap, we assumed that each core in our CMP corresponds to a future version of the Cell whose internal processing units (today's SPEs [12]) have twice the amount of local-store memory and a double precision floating-point unit. This would allow us to scale the same result described in [3] to a $256MBytes$ array of 2^{24} *double-precision* sample elements and, therefore, to use Bailey's FFT algorithm within each core to complete the first phase of the Cooley-Tukey algorithm in about $43ms$. Starting from this number and knowing that the Bailey algorithm requires $5k \log k$ floating point operations, we estimated the duration of the computation step in each of the subsequent iterations

as $1.8ms$, for $k = 2^{24}$.

Finally, assuming a $960Gbps$ photonic transmission line rate as done above, we obtained that our CMP equipped with the Non-Blocking Mesh would execute a 2^{29} double-precision sample FFT in about $66ms$, where $14ms$ are needed for the butterfly data exchanges. This value of $66ms$ will be considered the *reference value* of the total execution time for the rest of our analysis.

This FFT implementation assumes a message-passing MPI protocol and relies on high-bandwidth inter-core transfers. Each core computes on a portion of data, and the resulting data are exchanged among the local memories. Data transfers can be predicted in advance and there is no overhead due to memory coherence mechanisms.

For an application with such a regular communication pattern, a non-blocking topology allows all the transfers to take place simultaneously, because in each butterfly stage each core communicates with a different destination. A blocking topology, instead, presents some conflicts within the network, thus forcing some communications to wait for the completion of others.

In our simulations the same CMP equipped with a Blocking Mesh takes $74.6ms$ to complete the FFT computation due to an increment of $8.6ms$ for the butterfly data exchanges with respect to the non-blocking topology. Notice that the ratio between the two communication times is less than two because at each iteration part of the communication latency is hidden by the local computation.

From now on, as a reference topology we choose the Non-Blocking Mesh with a $625mm^2$ square die. Hence, a hop between two switches spans about $2.78mm$ and the average path between two gateways is 11 hops with 4 turns.

Because silicon photonics represent a new technology, and as such, roadmaps are not yet well established, it is difficult to predict the future scaling of device power consumption. Therefore, we consider three scenarios for the photonic link power requirements. First, based on the performance of the photonic transceivers that are available today in $130nm$ CMOS technology [32], the energy consumption of a complete photonic connection in our $22nm$ chip can be reasonably estimated to be $0.8pJ/bit$. This represents our primary consideration. However, expecting some further improvements in the electronic and photonic designs, we also consider the scenarios with $0.4pJ/bit$ and $0.2pJ/bit$.

Hence, to transfer simultaneously 32 blocks of $256MBytes$ at $960Gbps$, we need less than $24.5W$ (drawing on the conservative prediction), i.e. about $770mW$ per connection. The total power value remains the same for the Blocking Mesh, where the number of

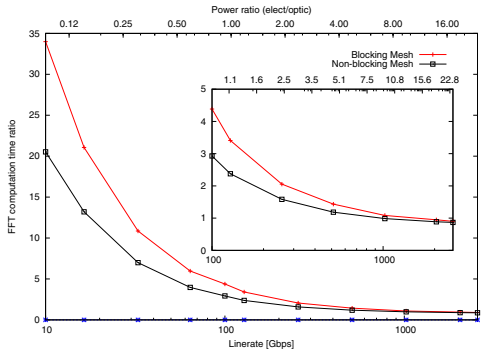


Figure 9. FFT-computation- time ratio and power ratio as a function of line rate for the electronic implementation of two 36-core topologies (blocking and non-blocking) with respect to an equivalent photonic Non-Blocking Mesh, which takes $66ms$ and dissipates $24.5W$ to complete the same task with a $960Gbps$ line rate.

hops is 12 and the number of turns is 3 because most of the power dissipation is in the optical interfaces rather then in the polarized rings.

Comparing Photonic and Electronic NoCs. After evaluating the performance of the CMP with a photonic network, we consider the idea of replacing it with an equivalent electronic network. Due to the persistence of channel utilization for the transfers of the FFT sub-arrays across the cores, a circuit-switched data network achieves better performance than a packet-switched NoC. Hence, we suppose we have the same organization as in the photonic NoC, but we replace each PSE with a device that is functionally equivalent and is implemented electronically. Further, we conservatively assume that it is ideal, i.e. without any delay and power consumption. In order to evaluate delay and power consumption of a communication over the equivalent electronic circuit-switched NoC, we consider the length of the optimally repeated wire in the given $22nm$ technology and like in [14] we assume an energy consumption of $0.25pJ/bit/mm$.

Considering the amount of data to be moved during the transfer stages, the message duration is at least of the order of milliseconds. Since the path-setup time and the light propagation are respectively tens and fractions of nanoseconds, they can be considered negligible. These considerations are valid for an electronic data plane as well: even if the signal propagation is slower in the copper than in an optical waveguide, the latency is still around few nanoseconds. In summary, the computation time does not depend on the media, but just on the line rate: for a given topology, it is the same for both the photonic network and an electronic equivalent network as long as the line rate is the same.

Fig. 9 captures the difference in computing the 2^{29}-

samples FFT with an electronic NoC as a function of the core's line rate and for the two different topologies. The y-axis reports the ratio of the total execution time over the reference time of $66ms$, spent by the photonic Non-Blocking Mesh operating with a line rate of $960Gbps$. The top x-axis reports the ratio of the power dissipated by the electronic NoC over the reference value of $24.5W$, which is dissipated by the same photonic Non-Blocking Mesh.

To assess the gain in performance-per-watt offered by the photonic NoC we consider two special cases:

1. to achieve the same execution time (*time ratio = 1*) as the photonic NoC, the electronic NoC must operate at the same line rate, but in doing so it dissipates $7.6W$ per connection, a value about 10 times higher. This leads to an overall power dissipation for the electronic NoC of about $244W$, a value that alone would exceed the total power budget for the CMP.

2. to achieve the same power dissipation (*power ratio = 1*) the electronic NoC must operate at a line rate of $100Gbps$, a reduction of 90%, thereby taking about $190ms$ to complete the FFT computation. This is 3 times more than the reference network because the computation time remains the same.

While in the first case the total time is dominated by the computation, in the second case it is dominated by the communication. The relation between line rate and total execution time is not linear, as shown in Fig. 9. Also, as the line rate increases the total computation time decreases, but the ratio between the power dissipation for the two networks grows.

The chart of Fig. 9 helps the electronic designer to: (a) evaluate the power budget for the NoC, (b) estimate the line rate that its links can sustain, and (c) determine the time necessary to compute the FFT. For instance, if the power budget for the electronic NoC is $100W$, i.e. about 4 times the reference dissipation value, then the corresponding line rate is about $400Gbps$. At that speed the evaluation can be performed in less then $85ms$. With a Blocking Mesh at the same line rate the necessary time would be almost $100ms$. Using the approach discussed above, Table 2 shows the scaling of the power and performance gain with different projections of the future power dissipation of photonic transceivers.

Discussion. While semi-custom NoCs that dissipate low power (i.e. hundreds of mW) can be efficiently built for SoCs used in embedded applications [13], the bandwidth requirements and die area of future high-performance CMPs cannot be satisfied with NoCs based on traditional circuit techniques. A reduction in power dissipation of at least a factor of ten is needed for high-performance CMPs [14]. This could be possibly achieved with some new promising circuit techniques that reduce the power of an electrical NoC with limited additional design complexity such as pulsed current-model signaling [23] and links based on low-leakage repeaters [31].

Photonic power $[pJ/bit]$	Power efficiency gain	Performance gain
0.8	10×	3×
0.4	20×	4.5×
0.2	40×	8.5×

Table 2. Performance and power gains as function of the photonic power consumption projections.

Our experimental results show that photonic communication has the potential to deliver the necessary reductions in power consumption in a scalable fashion particularly for critical applications that require massive data transfers at high bandwidths over a large-die CMP. Admittedly, the complexity of photonic integration remains high as the technology is significantly less mature that electronics. On the other hand, photonics may find its way into the chip in the nearer term as the best solution to bridge the growing gap between on-chip and off-chip communication bandwidths and to address the vastly increasing power/area costs of the chip I/O [21]. In fact, photonic communication across multiple chips and to DRAM memories can provide the same bandwidth-per-watt as on-chip communication in a manner that is independent of the distances spanned connecting elements within an entire multi-blade system [11]. The introduction of photonic I/O circuitry can then pave the way to the introduction of hybrid NoCs for CMPs to simultaneously reduce power consumption on-chip and off-chip.

6 Conclusions

On-chip photonic communication has been recently proposed as a solution to address the communication requirements in future high-performance CMPs. We presented a simulation-based assessment of this idea and we reached the conclusion that a photonic NoC is best suited to connect a limited number of complex multi-threaded cores. Furthermore, as long as the number of cores is limited, e.g. no more than 36, a photonic NoC based on a non-blocking topology provides better performance without additional design overhead than a photonic NoC based on a blocking topology. Finally, and independently from the chosen topology, we found that for those communication-intensive applications that will run on future large CMPs a photonic NoC offers a valid alternative to electronic NoC while seamlessly providing high-bandwidth and low-power connectivity with off-chip devices.

7 Acknowledgment

This research is partially supported by DARPA MTO office under grant ARL W911NF-08-1-0127.

References

[1] D. H. Bailey. A high-performance FFT algorithm for vector supercomputers. *Int. J. of Supercomputer Applications*, 2(1):82–87, 1988.

[2] L. Benini and G. De Micheli. Networks on chip: A new SoC paradigm. *IEEE Computer*, 49(2/3):70–71, January 2002.

[3] A. C. Chow, G. C. Fossum, and D. A. Brokenshine. A programming example: Large FFT on the Cell broadband engine. GSPx 2005.

[4] James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19:297–301, 1965.

[5] W. J. Dally and B. Towles. Route packets, not wires: On-chip interconnection networks. In *Proc. of the Design Automation Conf.*, pages 684–689, June 2001.

[6] P. Dong, S. F. Preble, and M. Lipson. All-optical compact silicon comb switch. *Opt. Express*, 15(15):9600–9605, 2007.

[7] N. Eisley and L.-S. Peh. High-level power analysis for on-chip networks. In *Intl Conf. on Compilers, Architecture, and Synthesis for Embedded Systems*, September 2004.

[8] A. Shacham *et al.* The case for low-power photonic networks-on-chip. In *Proc. of the Design Automation Conf.*, pages 132–135, June 2007.

[9] A. Shacham *et al.* On the design of a photonic network-on-chip. In *Proc. of the The First Intl. Symp. on Networks-on-Chips (NOCS)*, May 2007.

[10] A. Shacham *et al.* Photonic NoC for DMA communications in chip multiprocessors. In *IEEE Symp. on High-Performance Interconnects*, August 2007.

[11] B. E. Lemoff *et al.* MAUI: Enabling fiber-to-the-processor with parallel multiwavelength optical interconnects. *J. Lightwave Technol.*, 22(9):2043, 2004.

[12] D. Pham *et al.* The design and implementation of a first-generation CELL processor. In *Intl. Solid State Circuits Conf.*, pages 184–185, February 2005.

[13] F. Angiolini *et al.* A layout-aware analysis of networks-on-chip and traditional interconnects for MPSoCs. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 26(3):421–434, March 2007.

[14] J.D. Owens *et al.* Research challenges for on-chip interconnection networks. *IEEE Micro*, 27(5):96–108, Sept.-Oct. 2007.

[15] K. Goossens *et al.* Networks on silicon: Combining best-effort and guaranteed services. In *Conf. on Design, Automation and Test in Europe*, 2002.

[16] Y. Hoskote *et al.* A 5-GHz mesh interconnect for a teraflops processor. *IEEE Micro*, 27(5):51–61, Sept.-Oct. 2007.

[17] M. W. Geis, S. J. Spector, M. E. Grein, R. J. Schulein, J. U. Yoon, D. M. Lennon, C. M. Wynn, S. T. Palmacci, F. Gan, F. X. Käertner, and T. M. Lyszczarz. All silicon infrared photodiodes: photo response and effects of processing temperature. *Opt. Express*, 15(25):16886–16895, 2007.

[18] W. M. J. Green, M. J. Rooks, L. Sekaric, and Y. A. Vlasov. Ultra-compact, low RF power, 10 Gb/s silicon Mach-Zehnder modulator. *Opt. Express*, 15(25):17106–17113, 2007.

[19] C. Gunn. CMOS photonics for high-speed interconnects. *IEEE Micro*, 26(2):58–66, March/April 2006.

[20] W. Haensch. Is 3d the next big thing in microprocessors? In *Intl. Solid State Circuits Conf.*, February 2007.

[21] H. Hatamkhani, F. Lambrecht, V. Stojanovic, and C.-K. K. Yang. Power-centric design of high-speed I/Os. In *Proc. of the Design Automation Conf.*, pages 867–872, 2006.

[22] A. Hemani, A. Jantsch, S. Kumar, A. Postula, J. Oberg, M. Millberg, and D. Lindqvist. Network on chip: An architecture for billion transistor era. In *18th IEEE NorChip Conference*, November 2000.

[23] A.P. Jose, G. Patounakis, and K.L. Shepard. Pulsed current-mode signaling for nearly speed-of-light intrachip communication. *IEEE J. of Solid-State Circuits*, 41(4):772–780, April 2006.

[24] D. W. Kim, A. Barkai, R. Jones, N. Elek, H. Nguyen, and A. Liu. Silicon-on-insulator eight-channel optical multiplexer based on a cascade of asymmetric Mach-Zehnder interferometers. *Opt. Lett.*, 33(5):530–532, 2008.

[25] S.J. Koester, C.L. Schow, L. Schares, G. Dehlinger, J.D. Schaub, F.E. Doany, and R.A. John. Ge-on-SOI-Detector/Si-CMOS-Amplifier receivers for high-performance optical-communication applications. *Lightwave Technology, Journal of*, 25(1):46–57, Jan., 2007.

[26] B.G. Lee, A. Biberman, P. Dong, M. Lipson, and K. Bergman. All-optical comb switch for multiwavelength message routing in silicon photonic networks. *Photonics Technology Letters, IEEE*, 20(10):767–769, May, 2008.

[27] B.G. Lee, X. Chen, A. Biberman, X. Liu, I-W. Hsieh, C.-Y. Chou, J. I. Dadap, F. Xia, W. M. J. Green, L. Sekaric, Y. A. Vlasov, R. M. Osgood, Jr., and K. Bergman. Ultrahigh-bandwidth silicon photonic nanowire waveguides for on-chip networks. *Photonics Technology Letters, IEEE*, 20(6):398–400, Mar., 2008.

[28] B.G. Lee, B.A. Small, Q. Xu, M. Lipson, and K. Bergman. Characterization of a 4 × 4 Gb/s parallel electronic bus to WDM optical link silicon photonic translator. *Photonics Technology Letters, IEEE*, 19(7):456–458, Apr., 2007.

[29] A. Liu, L. Liao, D. Rubin, H. Nguyen, B. Ciftcioglu, Y. Chetrit, N. Izhaky, and M. Paniccia. High-speed optical modulation based on carrier depletion in a silicon waveguide. *Opt. Express*, 15(2):660–668, 2007.

[30] S. Manipatruni, Q. Xu, and M. Lipson. PINIP based high-speed high-extinction ratio micron-size silicon electrooptic modulator. *Opt. Express*, 15(20):13035–13042, 2007.

[31] A. Morgenshtein, I. Cidon, A. Kolodny, and R. Ginosar. Low-leakage repeaters for NoC interconnects. *IEEE Intl. Symp. on Circuits and Systems*, pages 600–603, 2005.

[32] C. Schow, F. Doany, C. Chen, A. Rylyakov, C. Baks, Kuchta D., R. John, and J. Kash. A < 5mW/Gb/s/link, 16 x 10Gb/s bi-directional single chip CMOS optical transceiver for board level optical interconnects. In *IEEE Intl. Solid-State Circuits Conference*, February 2008.

[33] P. Singer. 3-D integration enables 1 Gb memory. Semiconductor International, July 2005.

[34] Y. Vlasov, W. M. J. Green, and F. Xia. High-throughput silicon nanophotonic wavelength-insensitive switch for on-chip optical network. *Nature Photonics*, 2:242–246, 2008.

[35] Y. Vlasov and S. McNab. Losses in single-mode silicon-on-insulator strip waveguides and bends. *Opt. Express*, 12(8):1622–1631, 2004.

[36] Q. Xu, D. Fattal, and R. G. Beausoleil. Silicon microring resonators with 1.5-μm radius. *Opt. Express*, 16(6):4309–4315, 2008.

[37] Q. Xu, S. Manipatruni, B. Schmidt, J. Shakya, and M. Lipson. 12.5 Gbit/s carrier-injection-based silicon microring silicon modulators. *Optics Express*, 15(2):430–436, 2007.

[38] Q. Xu, B. Schmidt, J. Shakya, and M. Lipson. Cascaded silicon micro-ring modulators for WDM optical interconnection. *Optics Express*, 14(20):9430–9435, 2006.