Photonic Networks-on-Chip: Opportunities and Challenges

(Invited Paper)

Michele Petracca
Computer Science Department
Columbia University
Email: petracca@cs.columbia.edu

Keren Bergman
Electrical Engineering Department
Columbia University
Email: bergman@ee.columbia.edu

Luca P. Carloni
Computer Science Department
Columbia University
Email: luca@cs.columbia.edu

***Abstract*— As the number of processing cores that are integrated into a chip multiprocessors (CMP) continues to grow, the network-on-chip paradigm has emerged as a promising solution to address the problem of providing a robust interconnect network among them. In future high-performance CMPs, however, the high bandwidth requirements for both intra-chip and off-chip communication are severely challenging the electronic communications infrastructure to meet these demands without consuming a large fraction of the overall on-chip power-dissipation budget. The introduction of photonic technology for on-chip communication holds the promise of delivering scalable *bandwidth-per-watt* performance that cannot be achieved using only electronic communication. After reviewing the key recent technologies advances that are making possible the integration of photonic devices with CMOS processes, we describe a hybrid micro-architecture for NoCs that combines a broadband photonic circuit-switched network with an electronic packet-switched control network and we discuss the pros and cons of using two different network topologies to implement it.**

I. Introduction

The new trend of integrating several 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) [1], [3]. 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, to analyze and optimize the power dissipation of a NoC becomes increasingly important as the number of cores on the chip grows [5]. In fact, 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.

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

- the achievable communication bandwidth on a single waveguide (or link) can approach multiple terabits-per-second, for limited power dissipation;
- the power dissipation does not depend on the distance spanned by the optical waveguide, but scales only with the link transmission interface circuitry (optical modulators drivers and receivers).

On the other hand, 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.

The main advantage of using photonic communication links relies on a property of the photonic medium, known as *bit-rate transparency* [12]: unlike routers based on CMOS technology that must switch with every bit of the transmitted data, leading to a dynamic power dissipation that scales with the bit rate [11], photonic switches switch on and off once per message, and their energy dissipation does not depend on the bit rate. This property facilitates the transmission of very high bandwidth messages while avoiding the power cost that is typically associated with them in traditional electronic networks.

Another attractive feature of optical communications results from the *low loss in optical waveguides*: at the chip scale, the power dissipated on a photonic link is completely independent of the transmission distance. Energy dissipation remains essentially the same whether a message travels between two cores that are $2mm$ or $2cm$ apart. Furthermore, low loss off-chip interconnects enable the seamless scaling of the optical communication infrastructure to multi-chip systems.

Remarkable advances made over the past several years in silicon photonics have yielded unprecedented control over device optical properties. Fabrication capabilities and integration with commercial CMOS chip manufacturing that are now available open new exciting opportunities [6]. The optical NoC building blocks are nanoscale photonic integrated circuits (PICs) that employ optical microcavities, particularly those based on *ring resonator* structures shaped from photonic waveguides which can easily be fabricated on conventional silicon and silicon-on-insulator (SOI) substrates. This new class of small footprint PICs can realize extremely high interconnection bandwidths which consume less power and in-

978-1-4244-1684-4/08/$25.00 ©2008 IEEE

troduce less latency than their contemporary bulk counterparts. Compatibility with existing CMOS fabrication systems and the juxtaposition with silicon electronics enable direct driving, controllability, and the integration of these optical networks with processor cores and other silicon-based systems.

High speed optical modulators, capable of performing switching operations, have been realized using these ring resonator structures [16], [17] or the free carrier plasma dispersion effect in Mach-Zhender geometries [10]. The integration of modulators, waveguides and photodetectors with CMOS integrated circuits for off-chip communication has been reported and recently became commercially available [6]. At the receiver side, SiGe-based photodetectors and optical receivers were fabricated with reported high efficiencies [7]. Finally, low-loss waveguide technology, with crossovers and a fairly aggressive turn radii, has recently made some remarkable progress and the enabling technologies are currently available [2], [9].

These research advances provide the building blocks that are necessary to realize a photonic NoC in a CMOS process.

II. A Hybrid Approach to NoC Design

While photonic technology offers unique advantages in terms of energy and bandwidth, two necessary functions for packet switching, namely buffering and header processing, are very difficult to implement with optical devices. On the other hand, electronic NoCs do have many advantages in flexibility and abundant functionality but tend to consume high power, which scales up with the transmitted bandwidth [11].

An innovative NoC for CMPs that exploits the latest advances in silicon photonics was recently proposed in [14]. It is based on a *hybrid approach*: a high-bandwidth circuit-switched photonic network is combined with a low-bandwidth packet-switched electronic network. 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 $960Gbps$ of photonic transmission *line rate* per core, 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 [15] and its power consumption is analyzed in [13]. Its main organization is illustrated in Fig. 1(a) for the case of a chip-multiprocessor that contains 16 processing cores. The black circles represent the cores' network interfaces (*gateways*) and the white squares represent the *photonic switches*. In fact, 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 is able to

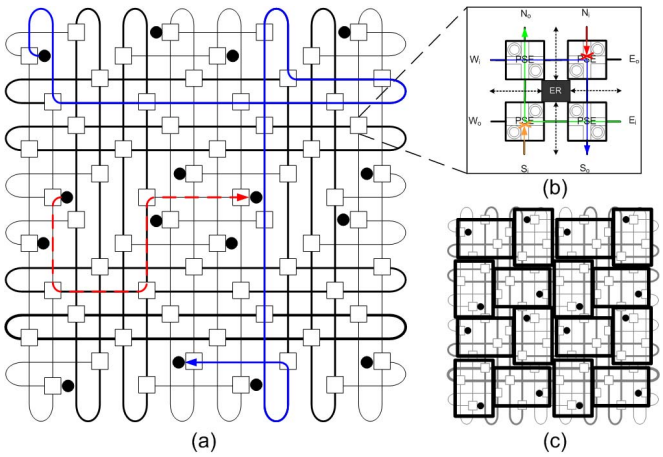


Fig. 1. 16-core *Blocking Mesh* photonic NoC from [14]: (a) shortest (longest) path is marked dashed (solid); (b) basic switch; (c) core layout over the NoC.

deflect the light when polarized. The ER not only switches the electronic packets but also polarizes the PSEs according to the value of the control packets that are exchanged during the setup and tear-down processes.

In order to optimize the fabrication process, it is reasonable to expect that the NoC silicon photonic devices and the CMP cores will be located on different planes by taking advantage of progress in *3D Integration*(3DI) [8].

Fig. 1(c) shows a possible layout for a 16-core CMP that was obtained by assuming that: (1) all cores are identical, (2) the network interface of each core must match the assigned position on the network plane, which is dictated by the network topology, and (3) only $90°$-multiple rotations and vertical/horizontal flips of the cores are allowed.

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:

- *gateway switch* connects the core interface to the NoC;
- *injection switch* receives the traffic from a gateway switch and deflects it toward the row to inject it into the NoC;
- *ejection switch* ejects the traffic from the NoC by deflecting it from the column into to the gateway switch;
- *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 [14] has a blocking topology and, therefore, it offers limited connectivity. Blocked communications flows are delayed until an open path is available resulting in some degradation to the network throughput and message latency.

As shown in [15] it is possible to reduce the blocking

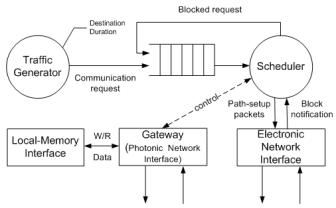


Fig. 2. Abstracted model of multithreaded processing core.

probability, and consequently improve the NoC performance, by *over-provisioning* the network. This 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 [14] 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 all the ones necessary for injection and ejection. In the rest of this paper we refer to this topology as the *Blocking Mesh*.

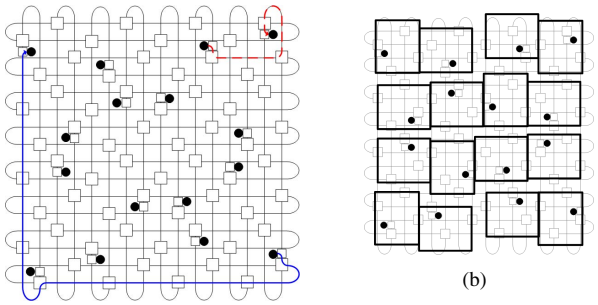
III. Multithreaded Core Model

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 multithreaded processor, where many threads are executed in parallel and each thread can independently request a data transfer to another core. Fig. 2 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 destination. 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.

• 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 communication). 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 set-up 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*.



(a) (b)

Fig. 3. (a) 16-core *Non-Blocking Mesh*; (b) core layout over the NoC.

IV. Non-Blocking Topology

"A *strictly non-blocking n-connector* is an n-connector with input set I and output set O, such that for any $i \in I$, $o \in O$ and a set $\wp$ of vertex disjoint paths from $I \setminus \{i\}$ to $O \setminus \{o\}$, there is a path from i to o which is vertex disjoint from $\wp$ [4]". A strictly non-blocking network can simultaneously handle the maximum possible number of connections, one for each node.

The non-blocking topology we propose is called *Non-Blocking Mesh* and can be built from a Blocking Mesh by using non-blocking switches as proposed in [15]. This allows also a simplification of the injection/ejection policy. Since the links among the 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. 3 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.

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 where the gateway switch injected it, 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 routing algorithm avoids the risk that a core is blocked injecting a message. This block 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.

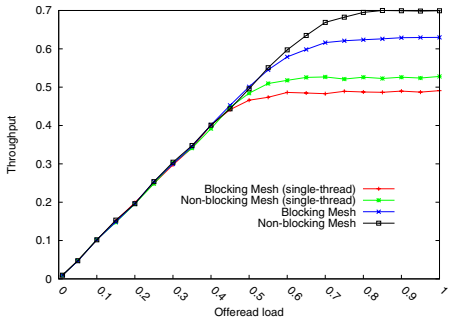


Fig. 4. Comparison of throughput-per-core for 36-core NoCs.

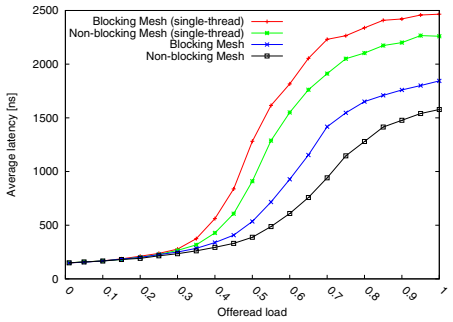


Fig. 5. Comparison of communication latency for 36-core NoCs.

V. Experimental results

Throughput-per-Core. 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. As in [15] we assume 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 $134ns$.

Fig. 4 shows the throughput-per-core as a function of the offered load for four distinct scenarios contrasting a blocking vs. non-blocking topology for the cases of both single- and multithreaded 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 try to make other requests that could be more successful if addressed to other destinations located in less congested parts of the network.

Further, any 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. However, by definition a non-blocking topology guarantees the delivery of a message to every free destination, thus improving the performance with respect to a blocking topology.

Communication Latency. The *communication latency* is measured as the average time needed to transfer all the data, from the generation of the request to the end of the teardown process. As shown in Fig. 5, the average latency is strictly related to the throughput because the higher is the latency of each data transfer the lower is the number of successful data transfers in the unit of time. The value of latency does not diverge because in our multithreaded core model the number of communication requests is limited by the number of threads.

VI. Conclusions

We described the opportunities that silicon photonic devices offers to improve the performance-per-watt of networks-on-chip and explained how a hybrid micro-architecture that combines a circuit-switched photonic network with a packet-switched electronic network can address some of the implementation challenges. We discuss how traffic aggregation at the core level and the employment of non-blocking topologies can better exploit the high bandwidth provided by the photonic data plane.

References

[1] L. Benini and G. D. Micheli, "Networks on chip: A new SoC paradigm," *IEEE Computer*, vol. 49, no. 2/3, pp. 70–71, Jan. 2002.

[2] X. Chen *et al.*, "Demonstration of 300 Gbps error-free transmission of WDM data stream in silicon nanowires," in *Conference on Lasers and Electro-Optics (CLEO'07)*, May 2007, paper CTuQ5.

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

[4] D. Du and H. Ngo, *Switching Networks: Recent Advances*. Springer, 2001.

[5] 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*, Sept. 2004.

[6] C. Gunn, "CMOS photonics for high-speed interconnects," *IEEE Micro*, vol. 26, no. 2, pp. 58–66, Mar./Apr. 2006.

[7] A. Gupta *et al.*, "High-speed optoelectronics receivers in SiGe," in *17th Intl. Conf. on VLSI Design*, Jan. 2004, pp. 957–960.

[8] W. Haensch, "Is 3d the next big thing in microprocessors?" in *Intl. Solid State Circuits Conf.*, Feb. 2007.

[9] I. W. Hsieh *et al.*, "Ultrafast-pulse self-phase modulation and third-order dispersion in si photonic wire-waveguides," *Optics Express*, vol. 14, no. 25, pp. 12 380–12 387, Dec. 2006.

[10] L. Liao *et al.*, "High speed silicon Mach-Zehnder modulator," *Optics Express*, vol. 13, no. 8, pp. 3129–3135, 18 Apr. 2005.

[11] T. Mudge, "Power: A first-class architectural design constraint," *IEEE Computer*, vol. 34, no. 4, pp. 52–58, 2001.

[12] R. Ramaswami and K. N. Sivarajan, *Optical Networks: A Practical Perspective*. Morgan Kaufmann, 2002.

[13] A. Shacham, K. Bergman, and L. Carloni, "The case for low-power photonic networks-on-chip," in *Proc. of the Design Automation Conf.*, June 2007, pp. 132–135.

[14] ——, "On the design of a photonic network-on-chip," in *Proc. of the The First Intl. Symp. on Networks-on-Chips (NOCS)*, May 2007.

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

[16] Q. Xu *et al.*, "Micrometer-scale silicon electro-optic modulator," *Nature*, vol. 435, pp. 325–327, 19 May 2005.

[17] ——, "12.5 Gbit/s carrier-injection-based silicon microring silicon modulators," *Optics Express*, vol. 15, no. 2, pp. 430–436, 22 Jan. 2007.