Maximizing GFLOPS-per-Watt: High-Bandwidth, Low Power Photonic On-Chip Networks

Assaf Shacham
Columbia University,
Dept. of Electrical Engineering
500 W 120th St.,
New York, NY 10027
+1 (212) 854 2768
assaf@ee.columbia.edu

Keren Bergman
Columbia University,
Dept. of Electrical Engineering
500 W 120th St.,
New York, NY 10027
+1 (212) 854 2280
bergman@ee.columbia.edu

Luca P. Carloni
Columbia University,
Dept. of Computer Science
1214 Amsterdam Ave.
New York, NY 10027
+1 (212) 939 7043
luca@cs.columbia.edu

ABSTRACT

As high-performance processors move towards multi-core architectures, packet-switched on-chip networks are gaining wide acceptance as interconnect solutions that can directly address the bandwidth and latency requirements as well as provide partial relief to the broader challenge of power dissipation. Still, studies show that the power consumed by on-chip networks will remain a major issue that has to be addressed to enable a true leap in future multi-core processors performance. Based on recent and expected technological advances in the integration of silicon photonic elements with CMOS electronics, we consider the usage of photonics to construct an on-chip network, offering unique advantages in terms of energy, bandwidth, and latency. We propose a novel architecture for a photonic on-chip network based on a hybrid approach: a network of wideband photonic switches combined with a parallel electronic control network. A high-level power analysis and comparison with electronic on-chip networks show that some of the advantages that have made photonics ubiquitous in long-haul transmission systems can be leveraged to construct photonic on-chip networks, delivering unprecedented computational capabilities, while operating at a fraction of the power of their electronic counterparts.

1. INTRODUCTION

Improvements in the performance of microprocessors have until recently been largely driven by the miniaturization of transistors, rising clock frequencies, and growing die sizes. Aggressive architectural solutions such as deep pipelines and complex cache memory organizations have accompanied these advances. The convergence of these design factors, however, has also led to the detrimental trend of exponentially increasing on-chip power dissipation [19]. To counterbalance these effects designers have sought quadratic reductions in dynamic power dissipation through aggressive supply voltage scaling. This is limited, however, by the exponential relationship between the increase in sub-threshold leakage current and the reduction of threshold voltage. Reductions in threshold voltage have made static leakage power large enough that it needs to be considered alongside dynamic transistor switching activity in the overall power budget [24] (for chips designed with 65nm technologies, for example, leakage power can account for up to 45% of the total power dissipation [6]). Consequently, in order to minimize power, the threshold voltage is now set as the result of an optimization, and not by technology scaling [19]. Meanwhile on-chip buses dissipate increasingly higher power due to the rapid growth in the number of repeaters and latches that are used to buffer and pipeline long metal lines. Specifically, up to 20% of the total power can be dissipated on long intra-chip buses, which are also typically very wide with multiple parallel lines (e.g., 32, 64 and even 128 bit wide in the recent IBM CELL[25]), each requiring substantial drivers [30].

In summary, high-performance microprocessors that keep existing architecture and circuit design techniques are expected to exceed package power limits by a factor of nearly 4X over the next decade; alternatively, microprocessor logic content and/or logic activity would need to decrease proportionally to match packaging constraints. Despite these measures, power densities are estimated to reach up to ~200W/cm^2 by 2010 [20]. This power density exceeds the air-cooling limit by a factor of 2, and is challenging for efficient liquid-cooling methods as well as for the required power supply.

1.1 The Rise of Multi-Core Architectures

Perhaps not surprisingly, the quest for both high performance and low power has brought designers to a major paradigm shift in the microprocessor architectures. The emerging trend is to replicate the computational logic in order to maintain processing

throughput while lowering clock frequencies and supply voltages. In fact, two processing cores running at half the frequency and half the supply voltage will save approximately a factor of 4 in $C \bullet V^2 \bullet f$ dynamic capacitive power, versus the "equivalent" single core [20]. This trend was marked by the arrival of the first commercial chips hosting multiple processing cores like the dual core multi-threaded IBM POWER 5 [23], the Intel ITANIUM 2 [31], and the CELL processor, a joint effort of IBM, Toshiba and Sony, which features a Power processing element and eight co-processing cores [22]. It is reasonable to assume that the number of these cores will continue to grow, leading to various generations of chip multiprocessors (CMP).

1.2 The Impact of Global Wires

Following the clear trend of CMPs, each technology generation will likely feature chips that host a larger number of smaller processing cores. To achieve the desired growth in computation power these cores, which may be physically distant, must interact in a tightly coupled manner. Thus these new CMP architectures are leading a major design shift from "computation-bound" to "communication-bound" [7]: the chip becomes a distributed system where global on-chip communication plays a dominant role during the design process.

Hence, another important technology trend, wire scaling, joins power dissipation in driving the design of high performance integrated circuits. While local interconnects scale in length approximately in accordance with transistors, global wires do not because they need to span across multiple modules to connect distant gates [17]. Consequently, long-range communication requires delays of multiple clock cycles as global wires must be heavily buffered and pipelined [34]. Traditional bus-based communication schemes, which are already "power hungry" [30], are destined to scale poorly if forced to support many IP cores [11]. In the case of general purpose microprocessors the evolution towards communication bound design implies that the amount of states reachable in a clock cycle, and not the number of transistors that can be integrated, becomes the major factor limiting the growth of instruction throughput. Furthermore, the increasing interconnect latency particularly penalizes traditional memory-oriented microprocessor architectures that strongly rely on the assumption of low-latency communication with structures such as caches, register files, and rename/reorder tables [1].

1.3 The Case for On-chip Networks

In this scenario various researchers have proposed to implement on-chip global communication with packet-switched micro-networks based on regular scalable structures such as meshes or tori [4], [8], [15], [16], [26]. These so-called *on-chip networks* are made of carefully-engineered links and represent a shared medium that is able to provide enough bandwidth to replace many traditional bus-based and/or point-to-point links. Furthermore, since they have better scaling properties, on-chip networks have the potential to mitigate the complexity of system-on-chip designs by facilitating the assembling of pre-designed and pre-validated processing cores through the emergence of new interface standards. However, as *performance-per-watt* is expected to remain the fundamental design metric for both embedded and high-performance computing for years to come, on-chip interconnection networks will have to satisfy communication bandwidth and latency requirements with minimal power dissipation.

1.4 The Photonics Opportunity

Leveraging the unique advantages of optical communication for on-chip photonic networks offers a potentially disruptive technology solution that can provide ultra-high throughput, minimal access latencies, and low power dissipation that remains independent of capacity. Recent advances in nanoscale silicon photonics have yielded improved control over device optical properties and unprecedented fabrication capabilities for the integration photonic elements in commercial CMOS chip manufacturing processes. An optical interconnection network that can capitalize on the enormous capacity, transparency, and fundamentally low power consumption of silicon photonics could deliver performance per watt that is simply not possible with all-electronic interconnects. Photonic channels could act as "on-chip super-highways" supporting large amounts of shared data traffic across longer distances in a bandwidth-oriented design of a network connecting processing cores and memories. On the other hand, electronic technology can complement the photonic network in overcoming some of the limitations inherent to photonics, namely processing and buffering.

Silicon photonic device technologies and integration have achieved unprecedented advances over the past five years. High speed optical modulators, capable of performing switching operations, have been realized using ring resonator structures [2, 37] and the free carrier plasma dispersion effect [28]. The integration of modulators, waveguides and photodetectors with CMOS integrated circuit for chip-to-chip

communication has been reported and recently became commercially available [13]. Finally, SiGe-based photodetectors and optical receivers were fabricated by numerous researchers and have become a reality [14]. These remarkable achievements, lead us to envision the integration of a fully functional photonic system on a VLSI electronic die. In particular, the photonic elements necessary to build a photonic on-chip network (dense waveguides, switches, modulators, and detectors) are now viable for integration on a single silicon chip.

We argue that combining global optical communication with electronic computation (and local communication) in a *hybrid* technology offers an unrivaled opportunity to directly address the critical latency and power challenges of next generation on-chip multiprocessors.

2. PHOTONIC ON-CHIP NETWORK

This section discusses the proposed architecture of the hybrid photonic-electronic on-chip network. The architecture is based on a two-layer structure:

(a) A photonic interconnection network, comprised of silicon broadband photonic switches interconnected by waveguides, is used to transmit high bandwidth messages;

(b) An electronic control network, mimicking the topology of the photonic network, is made of as set of electronic routers, each controlling a photonic switch.

Every photonic message transmitted is preceded by an electronic control packet (a *path-setup* packet) which is routed in the electronic network, acquiring and setting-up a photonic path for the message. Since buffering of messages is impossible in the network, as there are no photonic equivalents for storage elements (flip-flops, registers, RAM), we make use of *deflection routing* for contention resolution.

The main advantage of this approach relies on a property of the photonic medium, known as *bit rate transparency* [12]. Unlike electronic routers, which switch with every bit of the transmitted data while dissipating dynamic power scaling with the bit rate [30], photonic switches switch on and off at the message rate, 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.

In the following sections we describe the network architecture in a bottom-up fashion starting from the building blocks – the photonic switching elements.

2.1 Building Blocks

The fundamental building block of the proposed system is a broadband, waveguide-intersection photonic switching element based on internal reflection in the intersection area (Fig. 1). In the *OFF* state, when no current is injected into it, the switch functions as a passive waveguide crossover intersection (Fig. 1a). In the *ON* state, when carriers are injected into the intersection area in the form of electrical current, a change is induced in the refractive index and the light is reflected and forced to turn, thus

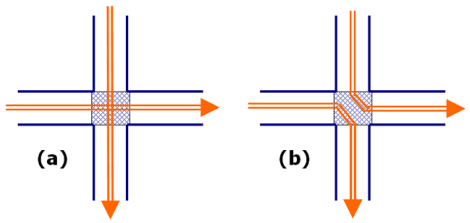

Figure 1 - Photonic switching element: (a) *OFF* state – a passive waveguide crossover. (b) *ON* state – reflection induced by carrier injection

creating a switching action (Fig. 1b).

Photonic switching elements based on the abovementioned effect have been realized in GaAs, providing the necessary wideband operation across a 35-nm wavelength band and ns-scale switching time [33]. Silicon-based switches which exhibit a lower insertion loss but larger switching times have also been reported [27]. The intersection angle in these switches, however, is substantially lower than 90°. Research efforts are now undergoing to design and fabricate a broadband, high-speed, low-loss, 90°-turn, silicon-based switching element.

Using 4×4 photonic switches enables the construction of familiar topological structures (e.g. mesh, torus etc.) and, therefore, allows us to benefit from the large body of literature on associated routing algorithms and flow control mechanisms as well as the analysis of their performance. Four photonic switching elements are, therefore, used to construct such switches (Fig. 2). Each switch is controlled by an electronic control circuit, termed an *electronic router*. Four pairs of waveguides serve as I/O links for the photonic messages and, similarly, metal lines are used as control links.

With this structure as a building-block, we can build planar 2-D topology networks that employ a photonic layer for transmission of high bandwidth messages and an electronic layer to control it. In this hybrid approach, the advantages of each technology are

leveraged to achieve optimal performance and to strengthen the weaknesses of the other technology.

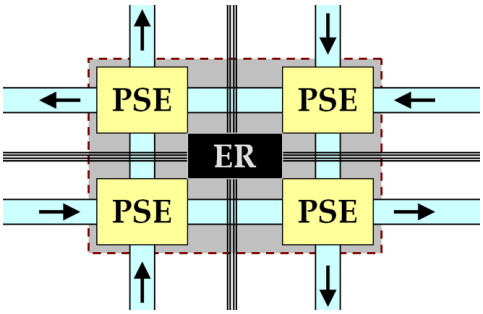

Figure 2 - 4×4 switch. Four photonic switching elements (PSE) controlled by an electronic router (ER).

2.2 Topology

The topology of choice in our design has to be compatible with the application of the entire system – a chip multiprocessor (CMP), where a number of identical processors are integrated as tiles on a single die. The communication requirements of a CMP are best served by a 2-D regular topology such as a mesh or a torus [32]. The compatibility of these topologies rises mainly from the planar, regular layout of the processor cores of the CMP and the application-based nature of the traffic – any program running on the CMP may generate a different traffic pattern [9].

The 4×4 photonic switches are suitable for integration in any planar 2-D topology. The choice of a topology (a mesh, a torus and a folded torus are a few possible options, see Fig. 3) and the tuning of the topology parameters (e.g. X and Y dimensions) depend on many factors: the number of terminals, the expected load, and which assumptions, if any, can be made on the expected traffic patterns. The following are some guidelines that will guide the topology design exploration phase.

First, additional paths in the network provide a medium for contention resolution – a function fulfilled by buffers in electronic networks. To provide and enhance this capability, and thus increase the network capacity, the path diversity in the network must be increased by adding paths and switches, or *overprovisioning* [9]. Obviously, the cost in power and area of this additional hardware should be considered but the photonic switches provide an attractive solution from this point of view – their footprint is expected to be extremely small (hundreds of microns) and they only consume power in the *ON* state. Consequently, as shown in Section 3, the power consumption doesn't increase linearly with the number of switches, but rather with the number of switches which are in *ON* state, i.e. the total number of turns that are made by messages traveling in the network. We therefore claim that a larger network, where deflections are less likely to occur, and messages experience a smaller number of turns, typically consumes *less* power than a smaller photonic network.

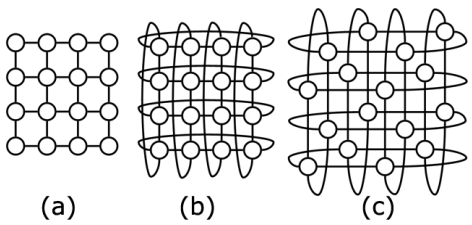

Figure 3 – 2-D planar topologies (a) mesh, (b) torus, and (c) folded torus.

Secondly, in electronic on-chip networks the power consumption of a link is also a function of links lengths [18], [35] so torus and folded torus topologies, which benefit from a lower expected number of hops per message, suffer from the drawback of having longer links (compare Fig. 3a to Fig. 3b and 3c), and thus additional power consumption. In the photonic network, the power consumption does not depend on the lengths of the links so, apparently, the tori and folded tori have clear advantages over meshes.

The best method of studying the performance of various topologies, weighing the effects of topological parameters on different systems, and verifying the abovementioned claims is via modeling and simulations. We are currently performing such a detailed study of these factors in order to optimize the photonic on-chip network design for CMPs. The results of this study will be reported in future publications.

2.3 Routing and Flow Control

A network topology is tightly coupled with a routing algorithm and a flow control technique. Any choice of each of these three elements affects the performance of the other two and the interplay between them has to be carefully studied [9].

The decision to use photonics and, more importantly, the impracticality of using photonic buffers, limits the choice of flow control mechanism and routing algorithm:

(a) Since buffering is not possible, a different method of contention resolution must be used;

(b) Since messages cannot be delayed for a long time in the network while the routing decision is made, complex routing algorithms should be avoided.

In order to cope with these constraints, we chose deflection routing as the main technique of contention

resolution in the network. In deflection routing (or "Hot Potato Routing") [3] when two or more packets contend for the same output port in a switch one of them is deflected to an undesired port and finds a different (perhaps longer) path to its destination. Deflection routing algorithms place some constraints on the network topology, since an alternative path must exist for each deflected packet. Further, it requires that the network is not heavily loaded so that over-congestion is avoided [9]. The large bandwidth offered by the photonic medium allows us to sacrifice some of the link utilization and keep the network lightly loaded to enable efficient deflection routing.

To keep the switches simple and the routing latency low, complex routing algorithms must be avoided. Since oblivious routing algorithms, (which are the simplest routing algorithms) cannot be used in conjunction with deflection routing [9], adaptive routing algorithms are necessary. Relatively simple adaptive algorithms such as adaptive XY dimension order routing or delta routing, which take into account the previous state of the switches, are an interesting option for on-chip networks that employ deflection-routing [29].

The choice of flow control method is also dictated by the medium. Firstly, "store and forward" flow control is not an option as the photonic messages cannot be stored in the switches. Secondly, the electronic control plane is decoupled from the photonic data plane, so simple wormhole routing cannot be used either. The method chosen therefore resembles wormhole routing in the sense that each switch forwards the message before it is completely received. The path, however, is not computed on a per-flit basis but rather is acquired in advance for the entire message by a *path-setup* packet traveling on the electronic layer. Similarly, after the photonic message completes its route, the path is freed up by *path-release,* another electronic packet.

2.4 Photonic Gateways

Electronic/Optic and Optic/Electronic (E-O and O-E) conversions are, obviously, necessary to transmit and receive photonic messages on the network. Small footprint silicon optical modulators and SiGe photodetectors have been reported in literature ([2], [37] and [14], respectively) and have recently become commercially available [13], to be used in photonic chip-to-chip interconnect systems.

Since the modulation rate offered by electronics will always be slow compared to the photonic transmission, a large degree of parallelism should be used. Optical time division multiplexing (OTDM) can be used for initial grouping of modulated data channels (e.g. 16 channels at 10 Gbps to a single 160 Gbps OTDM channel). Wavelength division multiplexing (WDM) can be used for additional enhancement of the bandwidth (e.g. 8 wavelengths, each modulated at 160 Gbps = 1280 Gbps). This large degree of parallelism requires a large area and a sizable power, but in the system there is only a single such gateway per processor tile – compared to 5 ports at each router in electronic equivalent on-chip networks. Since the photonic switches are 4×4, and do not have a fifth port for injection/ejection, the gateways are connected to waveguides in the network, between adjacent photonic switches, and use simple 2×2 switching elements for injection and ejection of messages.

Finally, as silicon is an indirect bandgap material, laser sources are difficult to fabricate on chip. However, external laser sources can be bonded to the chip, to provide the necessary light for modulation [13], [21].

2.5 Life of a Message

As an example of the operation of the proposed network, we describe the typical chain of events in the transmission of a message between two terminals. In this example, a "write" operation takes place from the processor in node A to a memory address located at node B (Figure 4). This procedure is designed based on the difference between the routing latency of the electronic *path-setup* packet, which has to undergo some processing in each router hop, and the photonic message that has extremely small latency because it only experiences the physical time-of-flight.

When the write address is known, even before the contents of the message are ready, a *path-setup* packet is sent on the electronic control network. The packet includes a destination address information, and perhaps additional control information such as priority, flow id, or other. The control packet is routed in the electronic network, reserving the photonic switches along the path for the photonic message which will

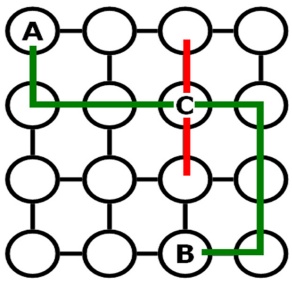

Figure 4 – A path example from A to B on a 4×4 mesh. A deflection in switch C forces the message to take slightly longer path.

follow it. The reserved path is typically not the shortest possible path, as contention is likely to occur with

other messages and cause deflections. Probabilistic methods can be used to calculate the maximum path length for a desired portion of the traffic (e.g. 99.9%). If the control packet fails to reserve a path within the maximum time, it is discarded and the last photonic switch in its path is configured to route the message to the nearest photonic message-sink where messages are discarded.

When the *path-setup* packet reaches the destination node, the photonic path is reserved and is ready to route the message. Since the photonic path is completely bidirectional a light pulse can then be transmitted onto the waveguide, in the opposite direction (from the destination to the source), signaling to the source that the path is open. The photonic message transmission then begins and the message follows the path from switch to switch until it reaches its destination.

Since special hardware and some additional complexity are required to transmit and extract the counter-directional light pulses, an alternative approach can be used: This approach is based on transmitting the message when the path is assumed to be ready according to the maximum expected path reservation latency. This approach requires less hardware but does not utilize the network resources as well the first one.

Once the photonic message has been received and checked for errors, a small acknowledgement packet may be sent on the electronic control network, to support a guaranteed delivery protocols. Lost or erroneous messages can be recovered using standard acknowledgement-based ARQ protocols such as *stop-and-wait* or *go-back-N* [5]. Faster retransmission of messages can be attained by generation of negative ack messages by the router which discards the control packet or by the destination node in the case of data errors in the message.

The layout of a 4×4 mesh with nodes A and B and a possible path between them is shown in Fig. 4.

3. HIGH LEVEL POWER ANALYSIS

As mentioned in the previous section, simulations are required to accurately analyze the network architecture and obtain true insights about its performance, namely saturation bandwidth, latency curves, power consumption, etc. However, a high-level analysis can be performed, comparing the proposed hybrid architecture to more widely-known electronic on-chip network architectures. This analysis shows the potential reduction in the power dissipation of the network for given bandwidth and port-count requirements.

We therefore present the following case study: an on-chip network for a 16-node CMP. Each processor has a terminal to the network requiring a peak bandwidth of 1024 Gb/s and an average bandwidth of 800 Gb/s. These large bandwidth figures are chosen to demonstrate the advantages of the photonic network in forwarding high-bandwidth communications. Although at first sight they may seem excessive, one should consider that state-of-the-art processors such as the CELL are already exchanging such bandwidths in their internal interconnects [25] and the trend towards higher on-chip communication bandwidth is expected to continue to grow in future systems.

Using previously published work we attempt to estimate the power dissipation of an electronic on-chip network and the proposed hybrid approach. Several assumptions are made for the simplicity of the analysis:

(1) The traffic driven by the processors is assumed to be uniform. While this is not the most accurate model of the actual expected traffic, it can still generate a preliminary comparison model for the behavior of the network elements.

(2) We only estimate the power spent by the interconnection network, on routing and forwarding the messages. The energy spent on optical modulation in the gateways, albeit large compared to the network power consumption, is assumed to be roughly equal to the driving energy of the electronic signals into the electronic network, so the two may be cancelled out.

(3) Both networks use a mesh topology and XY dimension order routing.

3.1 Reference Electronic Network

The reference network is a 4×4 mesh, where each router is integrated in one processor tile and is connected to four or fewer neighboring nodes (Figure 4). Each router has at most 5 ports: four (or fewer) for network connections and one serves as a local injection/ejection port. The routers are input-queued crossbar routers with a 4-flit buffer on every input port. This structure matches the widely accepted notion of electronic on-chip networks (see, for example [8], [32], and [36]).

As a prediction of future capabilities of electronics, we assume that the signaling rate is 3.2 GHz. The flit width, the number of parallel lines in each bus, is set to be 320 in each direction, thus providing the required peak bandwidth (1024 Gbps/3.2 GHz = 320). The injection rate is set at 0.8 to attain an average bandwidth of 820 Gbps, and for simplicity we assume that the network does not saturate under this load and

that sufficient routing resources exist to route a large number of parallel lines. The power analysis is done following the approach presented by Eisley and Peh in [10]. It is assumed that whenever a flit traverses a link, five operations are performed: (1) reading from a buffer, (2) traversing the routers' internal crossbar, (3) transmission across the inter-router link, (4) writing to a buffer in the subsequent router, and (5) triggering an arbitration decision. The link traversal energy cost ($E_{link\_traversal}$) can, therefore, be divided into the five components in Table 1.

Table 1 also lists estimates for these components, as measured for a 180-nm process [18] and extrapolated according to [36]. With process scaling the capacitance per unit length of wires is expected to remain approximately constant while the resistance per unit length is doubled for every generation, requiring the insertion of additional repeaters to maintain optimal propagation latency [18]. The energy consumption in future generations is therefore expected to be even higher thus presenting a more challenging problem to chip designers. Novel power reduction techniques, such as low swing drivers, have been proposed but they are becoming harder to implements when low supply voltages are used.

Table 1 – estimated link traversal energy components for a 320-bit flit across a 4mm link and a 5-port router

Component	Description	Estimate
$E_{bufread}$	buffer reading	1015 pJ
$E_{crossbar}$	crossbar traversal	3639 pJ
E_L	phys. link traversal	1260 pJ
$E_{bufwrite}$	buffer write	1015 pJ
$E_{arbiter}$	arbitration decision	70 pJ
$E_{link\_traversal}$	sum of the above	6997 pJ

Once $E_{link\_traversal}$ is known, we can use the network channel utilization statistics, readily available from mathematical analysis or from simulation to compute the power consumption in the network during a period of T cycles as:

$$P_N = \sum_{j=1}^{N_L} U_{L_j} \cdot E_{link\_traversal}$$

where U_{Lj} is the percentage of the T cycles that link j was used, or in other words, the utilization of link j. The total network power is the sum over all N_L links in the interconnection network.

For uniform traffic and the routing algorithm defined above (XY dimension order routing), it is straightforward to calculate the link utilization as a function of the injection rate of the nodes. For a given injection rate IR (0.8 in our case) the average link utilization across a 4×4 mesh is $IR \bullet 8/9$. Therefore, the total power consumption of the electronic on-chip network described above, which has 48 links, is:

$$P_{N,electronic} = 0.8 \cdot \frac{8}{9} \cdot 48 \cdot E_{link\_traversal} \cdot 3.2GHz = 765W$$

The main conclusion that can be drawn from this analysis is that when a truly high communication bandwidth is required on chip, even a dedicated network may not be able to provide it within reasonable power constraints. Since the electronic transmission is limited in bandwidth to a few GHz at most, the method to attain the transmission capacity is through parallelism – a very power-hungry method. Admittedly the above analysis is rather simplistic and based on a naïve implementation in a 180-nm process. Still, even an order of magnitude reduction in the calculated power consumption, obtained using aggressive power-reduction techniques or with a more accurate analysis, would still be too high to manage within reasonable packaging constraints.

3.2 Photonic Network

Unlike the electronic network, the dimensions of the photonic network do not necessarily match the number of processor tiles. Since (a) the photonic switches are integrated on a different physical layer, and (b) they do not consume static power (as shown below), it's advantageous to increase the number of photonic switches and thereby provide the path diversity necessary for deflection routing. In this design we assume an *overprovisioning* factor of 2, meaning an 8×8 photonic mesh, comprised of 256 photonic switching elements grouped into 64 4×4 switches, will serve the 4×4 CMP.

The power analysis of photonic on-chip networks is fundamentally different from the electronic network analysis. The per-flit analysis above can not be used because the power consumption of the network chiefly depends on the state of the photonic switching elements (*ON* state to force a message to turn, *OFF* state when a message proceeds undisturbed or when no message is forwarded). As mentioned above, the switching to *ON* state is achieved by injection of carriers into the switching element through a p-n junction. The amount of current required to achieve switching in a photonic switching elements depends on several physical parameters, namely structural dimensions, bandwidth, etc. As the switch area is expected to be extremely small (approximately 100 µm × 100 µm) the power required for the switching

operation is also expected to be very small – in the order of about 0.5 mW. We therefore use this figure (0.5 mW) as the power that is required to keep a photonic switching element in the *ON* state and to force a message to turn.

The total power consumption in the network depends on the number of switches in *ON* state. This number can be estimated based on network statistics as the product of the number of messages in the network at any given time and the expected number of turns each message makes. This estimation method assumes that a typical message's duration is much longer than the time necessary to traverse the network, and, consequently, that all the switches are turned *OFF* and *ON* nearly at the same time.

In non-deflection-routing meshes the dimension order routing algorithm dictates that each message makes, at most, one turn. Deflections, however, can force messages to take non-ideal paths and increase the number of turns. Preliminary simulation results show that when the network load is sufficiently low (i.e. injection rate < 0.7) the mean number of turns per message can be limited to 4. We will use this number in our analysis.

As described in the previous section, the photonic network uses OTDM and WDM to take advantage of the large bandwidth of the optical waveguides and switches. We assume each gateway is comprised of 128 modulators with a 10 Gbps tributary modulation rate, grouped into 8 wavelengths which are multiplexed together to provide a peak bandwidth of 1280 Gbps. An injection rate of 0.64 yields an average bandwidth of 820 Gbps – equal to the reference electronic network.

The average number of messages in the network at any given time can be computed as the product of the number of gateways and the injection rate to be 16•0.64 = 10.24. If we assume that each message makes, on average, 5 turns, then the number of photonic switching elements in the *ON* state can be estimated at 52 (in a 256-switch network), and the total power consumption can therefore be estimated as:

$$P_{N,photonic} = 52 \cdot 0.5mW = 26mW$$

This figure is dramatically lower than anything than can be approached by an electronic network and clearly presents the potential of bandwidth per unit power that can be provided by a photonic network.

Naturally the electronic control network in our hybrid photonic-electronic architecture does consume additional power. We approximate this power by taking advantage of the fact that the control network's topology is similar to that of the electronic reference network, except for its dimensions (4×4 mesh for the reference on-chip network, 8×8 mesh for the electronic control network of the hybrid approach). We assume that each photonic message is preceded by a 32-bit control packet and that the typical size of a message in the electronic reference network, as well as in the photonic network, is 2 Kbytes (16 kbits). We further assume that, on average, each control packet in the control network traverses a path twice as long as the path taken by an average message in the electronic reference network. Then, using some simplifying assumptions, the total power consumed by the electronic control network can be approximated as:

$$P_{N,photonic-control} = P_{N,electronic} \cdot \frac{32}{16384} \cdot 2 = \frac{P_{N,electronic}}{256}$$

Although the power analysis used here is rather simplistic and uses many assumptions to ease the calculation and work around missing data, its broader conclusion is unmistakable. The potential power difference between photonics-based on-chip networks and their electronic counterparts is immense. Even when one accounts for inaccuracies in our analysis and considers more aggressive electronic power-reduction techniques, the advantages offered by photonics represent a clear leap in terms of bandwidth-per-watt performance.

4. CONCLUSIONS

The advantages of photonic medium (high transmission bandwidth, better immunity to impairments and low power consumption, to name a few) have been known for decades. However, only recent advances in the fabrication of silicon photonic devices and the integration of those devices with CMOS electronic circuits on a silicon die have made the construction of complex structures such as integrated photonic on chip networks practical. The design of these networks requires consideration of the advantages and the shortcomings of photonics and leveraging the former to attain high performance while properly addressing the latter.

This paper presents a concept of a hybrid photonic-electronic on-chip interconnection network for a chip multiprocessor as a demonstration of the possibilities offered by photonic on-chip networks and as an example of the issues that have to be considered in their design. Obviously, the main driver and the rationale behind using photonics to provide the large communication bandwidth required by future chip multiprocessor and systems-on-chip is the potential to dramatically reduce the interconnect power consumption. As power consumption is becoming a

primary concern in the design of future integrated circuits this power reduction has been identified by many researchers as a key to future technological advances.

The on-going work following the concept presentation will include detailed simulation-based study of the performance of the network (i.e. bandwidth-latency tradeoffs) and the factors affecting it. Additionally, we intend to develop routing algorithms and flow control techniques specifically suitable for photonic on-chip networks. The goal of this research project is to ultimately provide a complete solution on which the next performance leap in processor performance can be based.

ACKNOWLEDGEMENTS

The authors would like to thank Azita Emami-Neyestanak for fruitful discussions.

REFERENCES

[1] V. Agarwal, M. S. Hrishikesh, S. W. Keckler, D. Burger, "Clock Rate versus IPC: The End of the Road for Conventional Microarchitectures," in Proc. *27th Annu. Intl. Symp. on Comp. Arch.(ISCA 2000)*, pp. 248–259, Vancouver, BC, June 2000.

[2] V. R. Almeida, C. A. Barrios, R. R. Panepucci, M. Lipson, "All-optical Control of Light on a Silicon Chip", *Nature*, vol. 431, pp. 1081-1084, Oct. 2004.

[3] P. Baran, "On distributed communications networks," *IEEE Trans. on Commun. Sys.*, vol. 12, no. 1, pp. 1-9, Mar. 1964.

[4] L. Benini, G. De Micheli, "Networks on Chips: A New SoC Paradigm', *IEEE Computer*, vol. 35, no. 1, pp. 70–80, Jan. 2002.

[5] D. Bertsekas, R. Gallager, *Data Networks*, Englewood Cliffs, NJ: Prentice Hall, 1992

[6] R. W. Brodersen, M. A. Horowitz, D. Markovic, B. Nikolic, V. Stojanovic, " Methods for True Power Minimization," In Proc. *Intl. Conf. on Computer-Aided Design (ICCAD 2002)*, pp. 35–42, San Jose, CA, Nov. 2002.

[7] L. P. Carloni, A. L. Sangiovanni-Vincentelli, "Coping with Latency in SoC Design," *IEEE Micro*, vol. 22, no. 5, pp. 24–35, Sep-Oct 2002.

[8] W. J. Dally and B. Towles, "Route Packets, Not Wires: On-Chip Interconnection Networks", Proc. *38th Design Automation Conf.* (*DAC2001*), pp. 684-689, Las Vegas, NV, Jun. 2001.

[9] W. J. Dally and B. Towles, *Principles and Practices of Interconnection Networks*, San Francisco, CA: Morgan Kaufmann, 2004.

[10] N. Eisley, L-S. Peh, "High-Level Power Analysis for On-Chip Networks," *CASES* 2004, Washington, DC, Sep. 2004.

[11] C. Grecu, P. P. Pande, A. Ivanov, R. Saleh, "Structured Interconnect Architecture: a Solution for the Non-scalability of Bus Based SoCs," In Proc. *14th ACM Great Lakes symposium on VLSI*, (*GLSVLSI '04*), pp. 192–195.

[12] C. Guillemot, "Transparent Optical Packet Switching: The European ACTS KEOPS Project Approach, " *IEEE/OSA J. Lightwave Technol.*, vol. 16, no. 12, pp. 2117-2134, Dec. 1998.

[13] C. Gunn, "CMOS Photonics for High-speed Interconnects," *IEEE Micro*, vol. 26, no. 2, pp. 58-66, Mar.-Apr. 2006.

[14] A. Gupta, S. P. Levitan, L. Selavo, D. M. Chiarulli, "High-speed Optoelectronics Receivers in SiGe,"in Proc. *17th Intl. Conf. VLSI Design*, pp. 957- 960, Mumbai, India, Jan. 2004.

[15] A. Hemani, A. Jantsch, S. Kumar, A. Postula, J. oberg, M. Millberg, D. Lindqvist, "Network on Chip: An Architecture for Billion Transistor Era," In Proc. 18th *IEEE NorChip Conference*, Turku, Finland Nov. 2000.

[16] S. Heo and K. Asanovic. "Replacing Global Wires with an On-chip Network: a Power Analysis," In Proc. *Intl. Symp. on Low Power Elect. and Design (ISLPED 2005)*, pp. 369–374, San Diego, CA, 2005.

[17] R. Ho, K.Mai, M. Horowitz, "The Future of Wires," *Proc. IEEE*, vol. 89, no. 4, pp. 490–504, Apr. 2001.

[18] R. Ho. *On-Chip Wires: Scaling and Efficiency*, Ph.D. dissertation, Dept. of Electrical Engineering, Stanford University, Aug. 2003.

[19] M. Horowitz, E. Alon, D. Patil, S. Naffziger, R. Kumar, K. Bernstein, "Scaling, Power, and the Future of CMOS," *IEEE Intl. Electron Devices Meeting (IEDM 2005)*, pp. 9–15, Washington, DC, Dec. 2005.

[20] ITRS. The international technology roadmap for semiconductors - 2005 edition. Available at http://public.itrs.net, 2005.

[21] L. A. Johansson, Zhaoyang Hu, D. J. Blumenthal, L. A. Coldren, Y. A. Akulova, and G. A. Fish, "40-GHz Dual-Mode-Locked Widely Tunable Sampled-Grating DBR Laser," *IEEE Photon. Technol. Lett.*, vol. 17, no. 2, Feb. 2005.

[22] J.A. Kahle, M. N. Day, H. P. Hofsteee, C. R. Johns, T. R Maeurer, D. Shippy, "Introduction to the CELL Multiprocessor," *IBM J. Res. Develop.*, vol. 49, no. 4-5, pp. 589–604, Sep. 2005.

[23] R. Kalla, B. Sinharoy, J. M. Tendler, "IBM Power5 Chip: A Dual-core Multithreaded Processor," *IEEE Micro*, vol. 24, no. 2, pp. 40–47, Mar. 2004.

[24] T. Kam, S. Rawat, D. Kirkpatrick, R. Roy, G. Spirakis, N. Sherwani, C. Peterson, "EDA Challenges Facing Future Microprocessor Design," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 19, no. 12, pp. 1498–1506, Dec. 2000.

[25] M. Kistler, M. Perrone, F. Petrini, "Cell Multiprocessor Communication Network: Built for Speed," *IEEE Micro*, vol. 26, no. 3, pp. 10-23, May/Jun, 2006.

[26] R. Kumar, V. Zyuban, and D. M. Tullsen. Interconnections in multi-core architectures: Understanding mechanisms, overheads and scaling. In Proc. Annual International Symposium on Computer Architecture, pages 408–419, June 2005.

[27] B. Li, S-J. Chua, "Reflection-Type Optical Waveguide Switch with Bow-Tie Electrode," *IEEE/OSA J. Lightwave Technol.*, vol. 20, no. 1, pp. 65-70, Jan. 2002.

[28] L. Liao, D. Samara-Rubio, M. Morse, A. Liu, D. Hodge, D. Rubin, U. Keil, T. Franck, "High Speed Silicon Mach-Zehnder Modulator," Opt. Express 13, 3129-3135 (2005)

[29] Z. Lu, M. Zhong, A. Jantsch, "Evaluation of On-chip Networks Using Deflection Routing," In Proc. *16th ACM Great Lakes Symposium on VLSI (GLSVLSI 2006)*, Philadelphia, PA, 2006.

[30] T. Mudge, "Power: A First-class Architectural Design Convstraint," *IEEE Computer*, vol. 34, no. 4, pp. 52–58, Apr. 2001.

[31] S. Naffziger, B. Stackhouse, T. Grutkowski, "The Implementation of a 2-Core Multi-threaded Itanium-family Processor," In *ISSCC Digest of Technical Papers*, pp. 182–183, Feb. 2005.

[32] T. M. Pinkston, J. Shin, "Trends Toward On-chip Networked Microsystems", *Int. J. High Performance Computing and Networking*, vol. 3, no. 1, pp. 3–18, 2005.

[33] C. Sato, *et al.*, "High-speed Waveguide Switches for Optical Packet-switched Routers and Networks," in Proc. *Optical Fiber Communication* (*OFC2004*), paper MF53, Los Angeles, CA, Mar. 2004.

[34] P. Saxena, N. Menezes, P. Cocchini, D.A. Kirkpatrick, "The Scaling Challenge: Can Correct-by-construction Design Help?" In Proc. *Intl. Symposium on Physical Design*, pp. 51–58, 2003.

[35] H-S. Wang, X. Zhu, L-S. Peh, S. Malik, "Orion: A Power-Performance Simulator for Interconnection Networks," in Proc. *35th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-35)*, Istanbul, Turkey, pp. 294-305, Nov. 2002.

[36] H. Wang, L-S. Peh, S. Malik, "A Technology-Aware and Energy-Oriented Topology Exploration for On-Chip Networks," in Proc. *Design, Automation and Test in Europe* (*DATE'05*), pp. 1238-1243, Munich, Germany, Mar. 2005.

[37] Q. Xu, B. Schmidt, S. Pradhan, M. Lipson, "Micrometre-scale Silicon Electro-optic Modulator", *Nature*, vol. 435, pp. 325-327, 19 May 2005.