Modeling and Evaluation of Chip-to-Chip Scale Silicon Photonic Networks

Robert Hendry, Dessislava Nikolova, Sébastien Rumley, Keren Bergman
Columbia University 116th St. and Broadway, New York, NY 10027
[rh2519, dnn2108, sr3061, kb2028]@columbia.edu

***Abstract*—Silicon photonic interconnects have been proposed as a solution to address chip I/O communication bottlenecks in multi-core architectures. In this paper, we perform comprehensive design exploration of inter-chip photonic links and networking architectures. Because the energy efficiencies of such architectures have been shown to be highly sensitive to link utilization, our design exploration covers designs where sharing occurs. By means of shared buses and silicon photonic switches, link utilizations can be improved. To conduct this exploration, we introduce a modeling methodology that captures not only the physical layer characteristics in terms of link capacity and energy efficiency but also the network utilization of silicon photonic chip-to-chip designs. Our models show that silicon photonic interconnects can sustain very high loads (over 100 Tb/s) with low energy costs (1-2 pJ/bit). On the other hand, resource-sharing architectures typically used to cope with low and sporadic loads come at a relatively high energy cost.**

Keywords—optical interconnection networks; silicon photonics; modeling; optimization

I. INTRODUCTION

As the performance of multi-core, many-core, and system-on-chip architectures continues to improve through Moore's scaling as well as through innovation in parallel processing, the demand for chip I/O communication is projected to exceed the capabilities of conventional electronic signalling and packaging. Off-chip bandwidth requirements have been projected to be tens of Tb/s [1] in order to meet future HPC requirements that strive for Teraflop/s chips, Petaflop/s racks, and Exaflop/s systems. Recent advancements in fabrication of on-chip integrated optical components, particularly in silicon photonics, have made optical communication a promising solution for realizing high off-chip bandwidth densities while remaining energy efficient.

A significant amount of research has been recently devoted to examining silicon photonic based chip-to-chip and chip-to-memory interconnect systems, and research has shown that Tb/s chip-to-chip links can be implemented with a single waveguide using very dense wavelength division multiplexing (WDM) [1-7]. Under sustained load conditions, silicon photonic links are predicted to be capable of picojoules per bit transmission efficiencies [7], [11].

Recent results have also revealed how sensitive optical networks are to deleterious effects in silicon photonics components. Though initially architectures involving many components have been proposed [21], [33], only more sober designs might be considered feasible when all these effects are taken into account, as we explored in [9]. Furthermore, in real systems, links are generally utilized in a sporadic manner. In these conditions, link utilizations around 10-20% are considered as normal while high utilizations (>70%) are often indicative of congestion and queuing delays. In such "real life" conditions, optical transmission links can exhibit substantially poorer energy efficiencies [8]. Indeed, the source laser power is consumed even while no data are transmitted on the link. Since lasers and device tuning circuits take relatively long times to stabilize (on the order of microseconds), powering down such optical links to avoid this static dissipation may not be reasonable in many circumstances. Any power waste is exacerbated by the fact that laser power represents a substantial part of the overall power envelope.

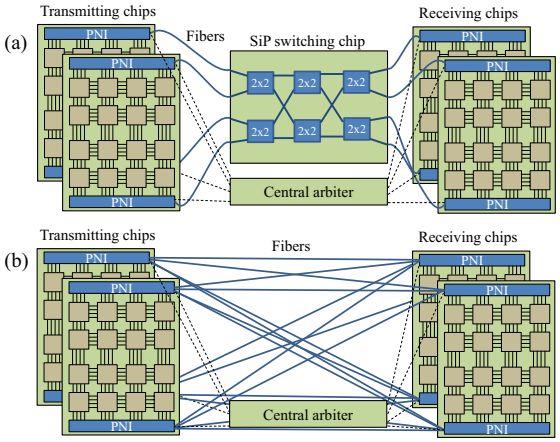

Figure 1. Two chip-to-chip interconnection architectures. (a) A switched architecture with a separate switching chip and a central arbiter. (b) A full mesh architecture with a central arbiter.

Utilizations of single links can be improved by consolidating the traffic from as many data flows as possible. But, achieving aggregation on chip also has costs in power and latency. Alternatively, consolidation can occur at the optical layer. This, however, translates into more complex optical architectures and, as is demonstrated in this paper, poorer energy efficiency. In summary, the design space for chip-to-chip interconnectivity with silicon photonics spans from simple point-to-point architectures, subject to poor energy efficiency due to low link utilization, to highly networked architectures, less affected by the link utilization but penalized by their hardware complexity. In this paper, we explore this design space and analyse how these two aspects jointly impact the resulting system performance in terms of latency and power efficiency.

We propose a modelling methodology to evaluate the power efficiency of silicon photonic interconnection networks. Our approach consists of the following steps. First we compute the power loss contributions from all devices on WDM silicon photonic link. This determines the maximum number of wavelengths and hence the total physical layer bandwidth that can be realistically supported on each link (i.e. the *link capacity*). In the following step, we derive by means of Monte-Carlo simulations the latency and bandwidth utilization as functions of the offered load. Using values reported in the literature we can realistically estimate the total power

978-1-4799-5860-3/14 $31.00 © 2014 IEEE
DOI 10.1109/HOTI.2014.14

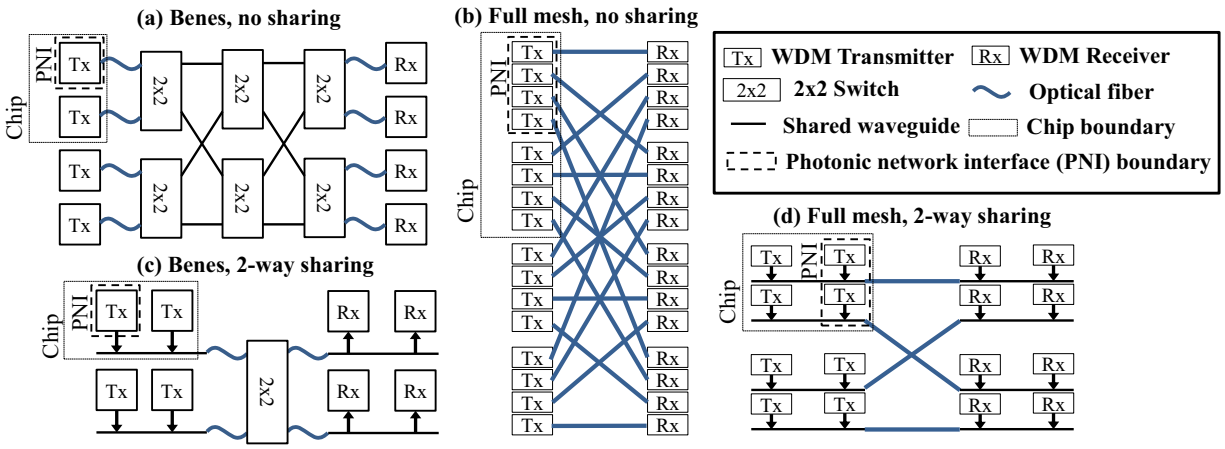

Figure 2. All of the network topologies in the design space that interconnect four photonic network interfaces residing on two separate chips. (a) A non-blocking Benes network topology. (b) A full-mesh network. (c) A Benes network topology with shared buses as inputs and outputs. (d) A full-mesh topology with shared buses as inputs and outputs.

consumption of the system. Finally, the energy-per-bit figures are derived for each explored architecture and traffic load.

Using the aforementioned methodology we obtain the complete latency/energy efficiency trade-offs that can be achieved with silicon photonic inter-chip interconnection networks as applied to the Benes and full mesh architectures, illustrated in Figure 1.

With these results at hand, we can for the first time conclude on the realm of applicability of silicon photonics chip-to-chip links. First, we show that with the latest silicon photonic devices reported in literature, a bisectional bandwidth of up to 100 Tb/s can be realized, with excellent energy efficiency if highly utilized. Second, we show that silicon photonics perform best in the situations where the offered bisectional bandwidth is at least 1 Tb/s. Finally, considering today's parameters, we show that the advantages of complex networked architectures are probably not worth their associated energy cost. That is, with complex silicon photonic switch fabrics, latency *and* energy suffer, contrary to the motivation for sharing network resources.

The paper is organized as follows. Section II describes the specific chip-to-chip design space explored. Section III provides details on the physical models used to evaluate link capacity. Section IV describes the simulations used to evaluate the network performance of the target architectures. Section V summarizes the models for the static and dynamic power consumption in these networks. Following, Section VI presents an evaluation of the architectures in terms of their overall performance and energy efficiency. In the last section the conclusions are drawn.

II. Silicon Photonic Chip-to-Chip Architectures

A. Network topologies

Silicon photonic technology offers the possibility of co-packaging computing electronic logic and photonic network interfaces (PNIs). The PNIs convert electronic data packets to dense WDM optical signals and carry the data off-chip. Once in an optical fiber, data can be carried relatively long distances (tens to hundreds of meters) at virtually no extra cost. We can imagine augmenting systems consisting of multiple, many-core processing chips with an optical network using such PNIs as end points. The relative distance independence and high bandwidth density of the optical interconnect can blur the boundary of what is considered local and non-local to the processing chip, potentially benefiting the ever-increasing parallelism of modern applications.

In this work we consider multicore chips residing locally on one or a few boards within an HPC system. A silicon photonic transceiver chip is integrated with each compute chip (for example via flip-chip integration [10]). These compute chips may have any number of compute logic cores. The analysis in this paper is generally agnostic of the total number of chips in the system or even the number of PNIs per chip, except in the cases when two PNIs share a waveguide and laser source for which we assume they are collocated on the same chip. The best dimensioning of PNIs to processing cores for a particular system, however, depends on many factors (e.g. application mapping, processor performance, network-on-chip architecture) that fall outside the scope of this paper. A PNI can serve a single, powerful computing core or the aggregate traffic from multiple cores. The number of PNIs in the system and their bandwidth capacity determines the maximum possible bisectional network bandwidth.

The PNIs can be interconnected through a switching fabric or through many point-to-point links. In this study, we consider two network topologies: a Benes made up of 2x2 optical switches and a full mesh. Although there are certainly many more topologies possible we wish to represent two distinct classes of networks: highly shared networks and non-shared networks. We chose the Benes network specifically to offer rearrangeably non-blocking operation while incurring only minimal complexity on the optical path through the switch: signals need only propagate through $2\log_2(R) - 1$ switching stages, where R is the radix of the switch.

Throughout this work, we assume that the Benes switching fabric resides on an independent chip, though in practice it could be integrated within one of the compute chips. In the presented evaluation, we assume for both topologies the presence of a central arbiter to manage the synchronization and reservation of network resources. The arbiter sets up circuits on demand on a first-come, first-served basis, and

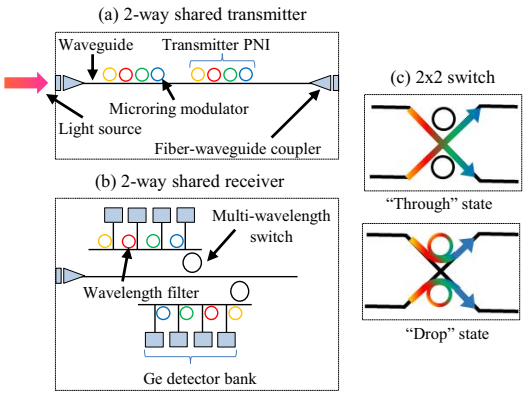

Figure 3. Overview of micro-ring based silicon photonic WDM links. (a) Two modulation sites coupled to a shared waveguide, each with four WDM channels. (b) Two receiver sites coupled to a shared waveguide. (c) A multi-wavelength switch in its two possible states.

communicates with the PNIs using out-of-band signaling channels. Again, there are many possibilities to accomplish arbitration and scheduling. However, in this work we choose a very simple mechanism in order to focus on the interaction between the physical layer design and first-order network performance.

Figures 2 (a) and 2 (b) illustrate the Benes and full-mesh network topologies, respectively. When using a Benes network, circuits through the switch fabric are established before the PNIs begin transmitting data. When a PNI receives a grant from the central arbiter, it begins serializing packets, and these packets propagate directly to their destination PNIs, as there is no buffering in the optical switch fabric. A Benes with four interconnected PNIs, as in Figure 2 (a), has four transmitters (i.e. sets of modulators) and four waveguides which carry the data off chip, and four fibers which carry the data to the switch fabric. Conversely, a full mesh has point-to-point links available from every source PNI to every destination PNI, as in Figure 2 (b). The PNI simply serializes packets on the transmitter that is associated with the packet's destination. Interconnecting four PNIs with a full mesh, as in Figure 2(b), requires sixteen transmitters with sixteen waveguides to carry data off chip, and sixteen fibers connecting the two chips.

Each PNI can have multiple waveguides connecting it to other chips, as in the full mesh in Figure 2(b). In order to introduce more opportunity for network resource sharing in our design space, each PNI can *also* share one or more on-chip silicon photonic waveguides with other PNIs located on the same chip. In this case, we say that two PNI are sharing the same *bus*. Figures 2(c) and 2(d) depict the Benes and full-mesh topologies interconnecting four PNIs with 2-way sharing, i.e. two PNIs share each waveguide and fiber. These cases introduce more possibility for contention, since two PNIs sharing the same waveguide cannot use that waveguide simultaneously. Therefore, in both the Benes and full mesh, the central arbiter must keep track of this blocking condition as it issues grants to the transmitting PNIs. In the Benes topology, if the total number of PNIs is maintained, sharing reduces the number of required waveguides and fiber by half, and also reduces the radix (thus complexity) of the switching fabric. In the full mesh topology, the reduction in the number of waveguides and fibers is the square of the amount of sharing; in the case depicted in Figure 2(d), the number of fibers is reduced by a factor of four. The potential energy savings from sharing is apparent considering that each path must be constantly powered by a multi-wavelength laser source.

Introducing sharing into these topologies, however, adds complexity and loss to the optical paths through each network. This added loss reduces the number of WDM channels that can feasibly be employed in each network. Therefore, the exact number of channels (i.e. the link capacity) must be calculated for each particular architecture; sharing provides a means for trading off link capacity for fewer, more utilized links. The link capacity calculation is detailed in Section III.

B. Silicon photonic implementation

The chip-to-chip network architectures in consideration are facilitated by the recent advances in silicon photonics devices. Silicon photonic microring and microdisk resonators have been demonstrated as energy efficient, high-data-rate optical modulators [11], [12]. For simplicity, we will refer to both types of devices simply as microrings, since their functionality is very similar. Microrings can be cascaded on the same waveguide and tuned to operate at different frequencies to form dense WDM modulation sites [13]. Microrings are especially suitable for integration with processing chips because of their low area and power footprint. Figure 3 (a) illustrates how waveguide sharing can be implemented in an example with two transmitting PNIs using four WDM channels. While one PNI is permitted to modulate the input optical power, the modulators in the other PNI remain off resonance, allowing most of the light to pass by.

WDM demultiplexing can also be implemented with microrings [14], and individual channels can be detected with CMOS-compatible germanium detector arrays [15]. Entire WDM signals can be "comb" switched by sizing microrings such that their repetitive resonant frequencies match the channel spacing in the WDM signal [16]. Figure 3 (b) illustrates how receiver sites can share a waveguide using comb switches to first switch data to the appropriate receiver before demultiplexing and detection. This configuration is desirable over directly cascaded filter banks, like the modulators in Figure 3(a), because it results in less optical loss and eliminates the need to have each filter actively controlled. Figure 3 (c) depicts two comb switches located at opposite corners of a waveguide crossing, forming a 2x2 non-blocking switch [17]. These switches can be cascaded to form a wide variety of switch fabric topologies. Optical fibers can be bi-directionally coupled to waveguides through vertical grating couplers [18] or taper couplers [19] to allow the networks to span multiple chips. These links are powered by multi-wavelength laser sources [20].

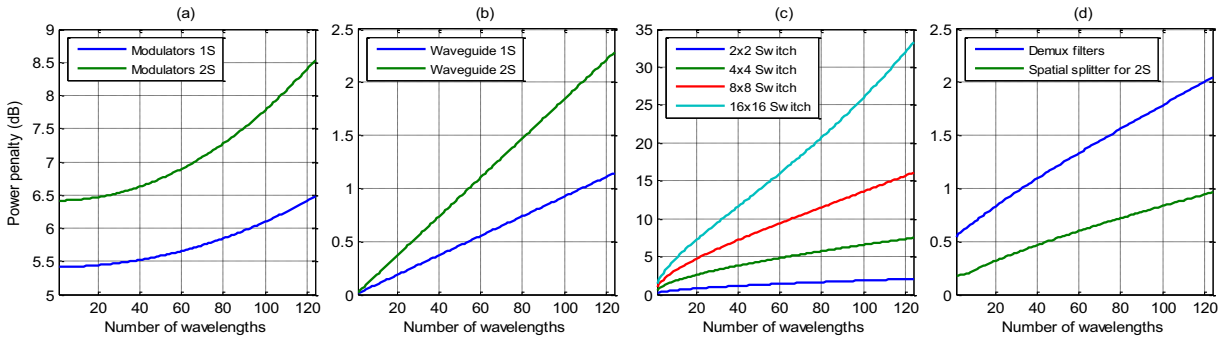

Figure 4. Loss contribution from the wavelength dependent components in the photonic link. Here "1S" and "2S" indicate no waveguide sharing and two-way waveguide sharing, respectively.

The collection of these silicon photonic devices creates a platform for a diverse set of optical network implementations. In the following sections we explore the tradeoffs between different silicon photonic networks built from these devices.

III. Physical Layer Modeling

As light propagates through silicon photonic devices, optical power is lost. In order to design a silicon photonic chip-to-chip data-link we must ensure that the optical power reaching the photodetector for all wavelengths is above its sensitivity P_d. Suppose that P_{in} is the optical power that is injected in the system by the laser, A is the attenuation of the system, and N_λ is the number of wavelength channels used in each link. To ensure the all wavelengths reach the photodetectors with sufficient power, P_{in}/A must be greater than the total power required at the receivers $P_d N_\lambda$, assuming that the input optical power and losses are divided equally amongst all the wavelength channels. It is easy to see that

$$P_d N_\lambda \leq \frac{P_{in}}{A} \equiv N_\lambda \leq \frac{P_{in}}{A P_d} \quad (1)$$

In order to maximize the link bandwidth, N_λ must be as large as possible while falling within this constraint. Therefore, P_{in} should be maximized, while A and P_d should be minimized. In practice, P_{in} is limited by the power above which non-linear effects cause too significant signal distortions in waveguide (generally assumed to be of 100mW [31]), and P_d is a characteristic of the photodetector (6.3uW in this study or -22dBm [15], considering a non-return-to-zero modulation at 10Gb/s). Therefore, the number of wavelengths supported by a link is determined by the attenuation A. This relationship illustrates the aforementioned trade-off between network complexity and link capacity.

In fact, other perturbations to the optical signals, caused by effects like imperfect filtering and crosstalk, can be converted to power penalty and added to the attenuation. Assuming the total attenuation and power penalties on the worst-case optical path through the system is PP_{tot}^{dB}, Eq. (1) expressed in dB becomes:

$$P_{in}^{dBm} - PP_{tot}^{dB} \geq P_d^{dBm} + 10\log_{10} N_\lambda \quad (2)$$

In addition, as in shown in the following section, the attenuation and power penalties incurred by silicon photonic devices, i.e. PP_{tot}^{dB}, are often a function of the spacing between wavelength channels. This spacing depends on N_λ because the total spectral range available is limited by the transparency of the waveguide(typically 50nm is assumed), thus adding more wavelength channels reduces the spacing between them. Therefore, Eq. (1) and Eq. (2) cannot be solved analytically in a straight-forward manner; an optimization algorithm is used to maximize N_λ for each architecture in our design space, given its cost function $PP_{tot}^{dB}(N_\lambda)$. Lastly, we assume a minimum channel spacing of 50GHz to avoid excessive overlap of modulating channels which corresponds to a maximum of 125 wavelength channels per link given a 50nm total spectrum.

A. Device loss models

To estimate the total power penalty, the contributions of all the components present along a link contributing to signal degradation must be evaluated. The components included in this study include waveguides, switches, modulators, waveguide couplers, passive filters, fiber-waveguide couplers, and photodetectors. Losses in fibers are assumed negligible for this study due to the short distances involved (less than 1m). The losses considered are briefly summarized here but a more in-depth discussion can be found in [9].

Modulator microrings incur a 1dB signal attenuation to the wavelength they modulate when modulating a "one" bit. In practice, imperfect extinction causes power penalty we assume to be 2dB. Since the data is modulated through on-off keying, the average optical power lost by using both low amplitude ("zero" bits) and high amplitude ("one" bits) signals is 2.4 dB, assuming a 50% probability of modulating each. In addition, modulators that operate on other waveguides contribute a small amount of loss to each wavelength channel. Figure 4 (a) plots the sum of all these effects over a varying number of wavelengths for both the basic and two-way shared configurations.

The silicon photonic waveguides that have been demonstrated to interact with the other devices assumed in this work incur a constant loss of about 1dB/cm. Figure 4 (b) shows the loss in the waveguide for different sharing factors. This loss changes as a function of the number of wavelength channels for two reasons: first, because the number of modulating rings and filters increases with the number of wavelength channels, requiring more waveguide length to interconnect them all; and second, the size of the microring switches in the Benes network grow with the number of wavelengths that need to be switched due to their fundamental physical properties, thus more waveguide length is required in the switching fabric.

The network switch loss, figure 4 (c) accounts for the loss and cross talk penalty for passing through several stages of 2-by-2 microring switches. The calculations of these exact

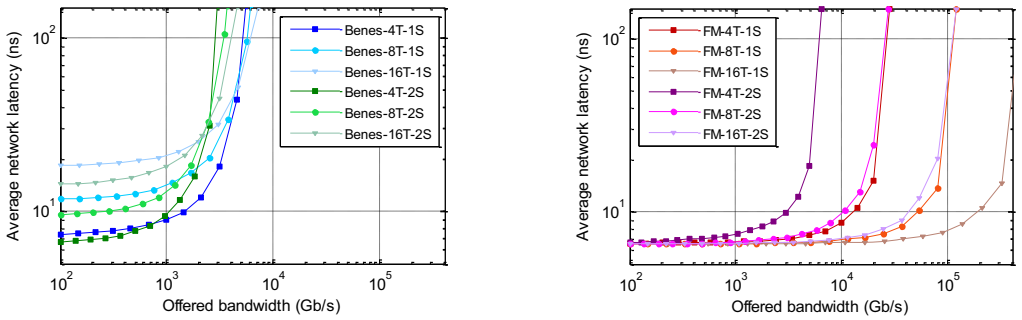

Figure 5. Bandwidth vs. latency for all network configurations. The legends indicate the topology, the number of transmitters, and the amount of waveguide sharing.

values are taken from [23], where optimized parameters [22] are assumed. As mentioned earlier, if R is the radix of the switch in a configuration, $2\log_2(R) - 1$ switches must be traversed by the optical signals.

We account a 0.028 dB loss [32] per waveguide crossing. Since we assume the photonic devices are etched onto a single silicon plane, the particular embedding of Benes network used will impact the number of waveguide crossings in the Benes configurations. We chose an embedding equivalent to two back-to-back butterflies because of its relatively low number of crossings (compared to other embeddings, such as using an omega-like layout, though no formal proof is given here).

Before detection, the WDM signals are filtered by an array of microring filters. These losses include power penalties due to signal truncation and constant insertion loss. In the case of two-way sharing, a single 1x2 switch is required to move the data to the correct array of filters and photodetectors. This loss is calculated similarly to the network switches. Both of these losses are shown in Figure 4 (d).

Finally, each waveguide-fiber coupler, required to move light on and off chips, incurs a 1dB coupling loss. There are four of these couplers along the path of each signal in the Benes configurations and only two in the full mesh configurations. We also assume a 2dB power penalty due to jitter in the clocking mechanisms.

B. Physical layer evaluation

At this point, we are able to estimate the total contribution of each component to the signal attenuation and deterioration, PP_{tot}^{dB} of Eq. (2), for each network configuration and therefore determine the total number of wavelengths supported by each. Figure 6 summarizes the results of this analysis in terms of the peak bisectional bandwidth, which multiplies the total number of wavelengths per link by the number of links in the bisection and again by the data rate (10 Gb/s). It is clear that by sharing network resources, via the Benes network and waveguide sharing, the available link capacities and therefore peak bisectional bandwidths are reduced.

In the following section, we explore how the combination of this link capacity degradation and traffic aggregation due to sharing affect the overall performance of the networks in comparison with the non-sharing full mesh.

IV. Network Traffic Simulation

The sharing factor and choice of network topology effect the network performance in three ways: (1) as seen from the previous section, these architectural choices affect the link capacity and, in effect, serialization latency and bisectional

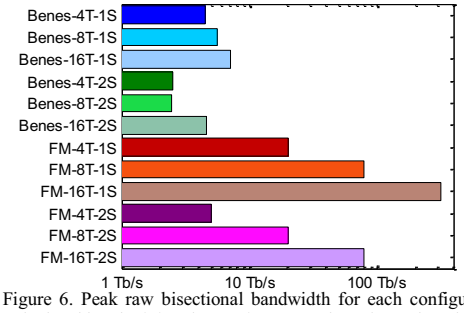

Figure 6. Peak raw bisectional bandwidth for each configuration. These are calculated by obtaining the maximum number of wavelengths, multiplying by the number of links in the interconnect, and multiplying by the assumed data rate, 10 Gb/s.

bandwidth; (2) the total number of links available in each configuration, along with the capacity of each link, defines the peak bisectional bandwidth; and (3) the choice of topology determines the contention characteristics of the network, which inevitably impacts latency due to queuing (which happens when two source-destination pairs sharing a link are active at the same time). We use simulation to determine the average message latency and average link utilization of each possible configuration in the design space.

The performance of each network architecture is obtained though Monte Carlo simulation using random Poisson traffic generators and infinite queues. The centralized arbiter allocates network paths on a first-come, first-served basis with a zero latency path reservation time. In other words, the arbiter issues a grant as soon as any transmitter requests a path through the network, allowing the transmitter to begin serializing data immediately, provided the requested path is available. Though this assumption is optimistic, especially considering small messages sizes use in our simulations, the goal of the present work is not to evaluate the quality an arbitration scheme, but rather to find bounds in terms of performance assuming a quasi-ideal arbiter. As is illustrated in later sections, this model is sufficient to confidently make conclusions about the utility of switching and network sharing in silicon photonic networks at this scale.

The combination of the physical layer analysis and this network simulation is crucial for fully understanding the power and performance impact of architectural changes in silicon photonic network architectures. Figure 5 plots the average latency versus the total offered bandwidth from the transmitting chip in the non-blocking Benes (a) and the full-mesh topologies (b). Each line represents an architectural configuration under a varying aggregate input load. The figure

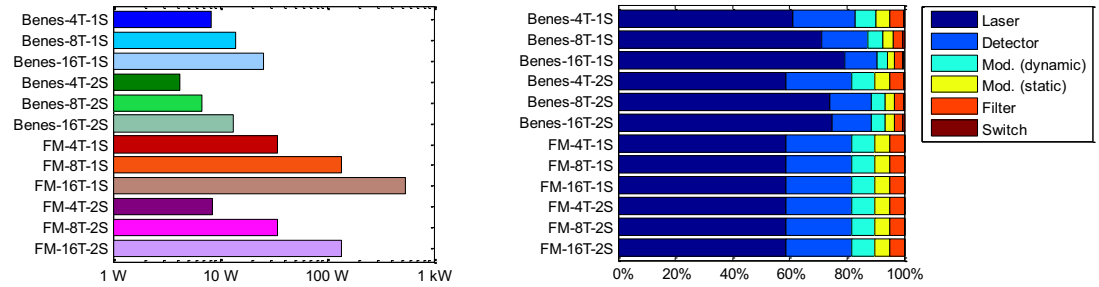

Figure 7. Total power of each configuration and breakdown of relative contributions from each power component model.

labels indicate the type of topology (e.g. "Benes"), the number of PNI **t**ransceivers (e.g. "4T"), and amount of waveguide **s**haring (e.g. "2S"). "Benes-4T-1S" represents the architecture with four transmitters and four receivers interconnected by a Benes network, with no waveguide sharing (i.e. one PNI per waveguide), while "FM-16T-4S" represents a full mesh connecting 16 sites with four transmitters per input waveguide and four receivers per output waveguide.

The minimum latency for each architecture, which is shown by the data points before saturation, is dominated by the packet serialization time. This serialization time is determined by the link capacity calculated in Section III and the message size, which we assume is 1kB. However, as previously mentioned, in this work we do not consider the effects of scheduling and reservation in the central arbiter or the processing that occurs in the network interfaces. Therefore, these latencies should not be taken as absolute predictions of the performance of these networks, but rather as a means to characterize an upper bound on their performance.

The maximum offered bandwidth from each configuration varies over several orders of magnitude for two reasons: (1) the link capacity increases for less complex links, and (2) the number of total available links (i.e., paths through the network) increases for less shared network topologies. From the combination of these two effects, given the same amount of waveguide sharing, the full mesh topology always outperforms the Benes topology in terms of bisectional bandwidth and latency. However, the full mesh network also requires many more links to be simultaneously powered. In the following section, we present the power models which are used to measure the impact these architectural changes have on the energy efficiency of the network.

V. Device Power Modeling

The power of the different architecture configurations is calculated using measurements from demonstrated devices where available and projected energy efficiencies for future silicon photonic device. We consider power consumers that are directly involved with optical transmission, including laser sources, modulators, switches, filters, and detectors. Other network power contributors, such as the electronic network interface and arbiter logic, are excluded in order to maintain a tractable set of design parameters.

The largest contributor to power consumption is the laser source. Because lasers cannot be turned on and off according the link activity, due to slow (millisecond) stabilization times, underutilized links are the major inhibitor to realizing very good power efficiency. Currently, the best laser source reported in literature in the CMOS-compatible silicon photonic platform operates with 4.5% power efficiency [20]. Considering the large research effort in the area we use a common [7] projected multi- wavelength laser source model with 10% efficiency. Still, this power efficiency remains low, and only exacerbates the aforementioned utilization problem. Since the upper-bound on the off-chip input optical power is 125mW, all of the target architectures require roughly 1.25W of wall-plug power *per link* to provide optical power to the network. Thus, the power efficiency of each architecture depends heavily on the number of powered links and the utilization of these links.

In most configurations, the static power dissipated by the photodetectors is the second-largest contributor to the overall dissipation. These require 3.95mW per detector [26], and one detector is required for every wavelength channel in every receiver bank.

Resonant frequencies of ring-based devices (i.e. modulators, switches, and filters) tend to shift under thermal fluctuations, which are normal and expected to occur in computing systems as the computational logic dissipates heat. In addition, the resonant frequency of a microring is highly sensitive to small structural changes that inevitably occur during fabrication. Therefore, mechanisms must be in place to stabilize the resonant frequency of each device around the desired center frequency. Local heaters are employed to accomplish this thermal tuning [27], and feedback systems have been demonstrated for continuous stabilization [28], [29]. In this work, we assume 0.875 mW for this thermal tuning and trimming for each modulator and filter, and 3.5 mW for each switch. Athermal solutions have also been proposed [30], but require a very large area footprint that prohibits their usefulness in this work.

The dynamic power dissipated to drive modulators during modulation is also modeled. Modulators have been demonstrated to consume 1.35 mW with a 10Gb/s data rate [26]. This power is added to the overall power consumption for each modulator, but only while they are in use according to the results from network simulation. The dynamic power consumption of the switch is estimated to be negligible in comparison with the static tuning and trimming power consumption.

Figure 7 shows the total power consumption of each configuration at peak power (100% utilization) and the relative contribution of each modeled component. The peak power of each configuration corresponds roughly to its peak bandwidth mainly because both the bisectional bandwidth and power are largely dependent on network radix (i.e. the number of transmitters). However, in the cases where the bandwidth is limited by loss—as with the Benes networks—the power continues to grow with the number of transmitters because the loss induces a higher input power per channel. The laser

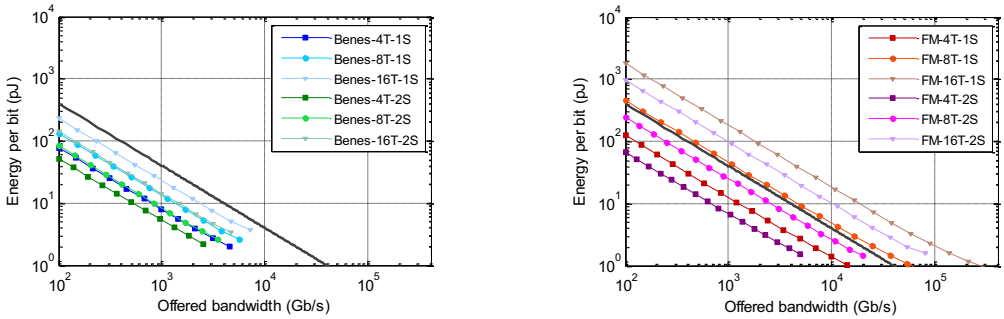

Figure 8. Energy per bit vs. offered bandwidth for all configurations.

power dominates the total power consumption, especially when additional laser power is required to overcome lossy optical paths. The power breakdown for the full mesh configurations are equivalent because these configurations are not limited by loss (i.e. they are limited only by the minimum channel spacing) and thus each link has the utilizes the same number of wavelengths. Since we assume for simplicity that the laser is always operating at 1.25W, the power per channel in this case is always the same.

VI. Evaluation

Provided the link capacity, average latency, link utilization, and a means to calculate static and dynamic power, we evaluate each architecture configuration under varying loads in terms of the energy required per bit transmitted. Figure 8 (a) and Figure 8 (b) plot the energy per bit for all of the target network configurations against the offered load. A line is drawn for reference where the total transmission power is 40W, which is often cited as the maximum power designated for chip I/O [1]. Since we exclude some of the power contributions from the on-chip network interfaces, the reference line at the very least tells us which architectures are infeasible given this power budget.

As the offered load increases, the utilization of each architecture improves. As a result, the energy per bit improves inversely proportional to the offered load, visible in the slope of the lines. These appear linear because dynamic consumption never accounts for more than 12% of the total power. For a given bandwidth requirement, the most energy efficient architecture is the one associated with the lowest line directly vertical to the desired offered bandwidth. Five different network architectures dominate in terms of energy efficiency across the bandwidth range from 100 Gb/s to 320 Tb/s.

The smaller bandwidths, up to ~2.5 Tb/s, are dominated by a Benes with 2-way waveguide sharing. In this case, bandwidth is sacrificed for more network resource sharing, therefore higher laser power utilization, and therefore better energy efficiency at these loads. However, when the bandwidth exceeds 2.5 Tb/s, the full mesh architectures are more energy efficient, even when under-utilized. Provided our optimistic network simulation, this indicates that the complexity added to gain network resource sharing is not worthwhile for these loads.

As the offered load approaches the peak bandwidths of each configuration, the average latency inevitably rises due to network saturation and queuing. Thus, even where one architecture dominates in terms of energy efficiency it may be advantageous to choose a different architecture that has a higher peak bandwidth, but is less energy efficient, to reduce the average latency.

Figure 9 illustrates this latency and energy efficiency tradeoff for various offered loads. There is indeed Pareto optimality for each load; the architectures which fall on the optimal front are circled. With a relatively low load (400 Gb/s), the energy-per-bit dominating architecture is the Benes-4T-2S. However, if some additional energy per bit is allowed, the FM-4T-1S or FM-4T-2S architectures can be used to improve average latency. The Pareto-optimality of these three architectures indicates that, for the given load, no other of the explored architectures out-perform in *both* latency and energy-per-bit. The Pareto-optimality extends through higher loads, which we show for 1 Tb/s, 4 Tb/s, and 40 Tb/s. Thus, the bandwidth requirement, tolerance to latency, and power budget are all determining factors for the best physical layout and network topology to choose.

The final conclusion we can draw from our results is evident by observing the minimum energy per bit for each load in Figure 9. Under lower loads, no configuration falls below 10 pJ/bit, which may be a prohibitive figure given alternative electronic links. This shows that, even with our optimistic network arbitration assumptions, the cost of complexity in the shared networks is greater than the improvement in utilization due to sharing. On the other hand, we show that a full mesh network, which requires virtually no arbitration, can sustain very high loads with only a few pJ/bit.

Although there may be opportunity to further decrease the energy per bit by using fewer wavelengths (note in Section III we always maximize wavelengths), this will inevitably come with increased latency. Additionally, dynamic techniques, such as putting the links in “sleep” states, could be used to further improve the energy efficiency under low loads, though this operation would likely only be worthwhile under sufficiently “bursty” traffic loads. We leave these analyses for future work.

VII. Conclusions

We show through a multi-layered analysis the expected performance of a range of silicon photonic inter-chip network architectures. Using physical layer models and network simulations we are able to characterize an upper bound on the energy efficiency of these network architectures and determine the best architecture for a given set of load, energy, and latency requirements. This first-order analysis—in the sense that we do not distinguish between data and bits used in network protocols, or model network scheduling in detail—is

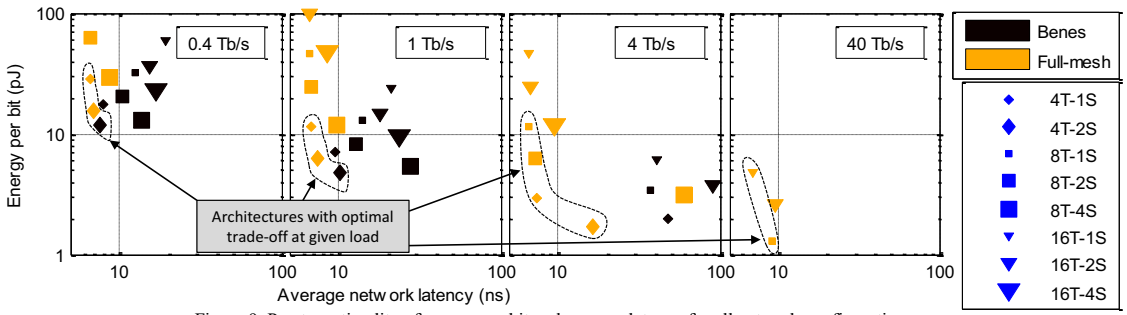

Figure 9. Pareto-optimality of energy per bit and average latency for all network configurations.

sufficient to highlight important trends in silicon photonic network design. Specifically, when optical resource sharing is advantageous (i.e. under low loads) it comes at the cost of increased energy-per-bit. Second-order effects in arbitration scheduling algorithms, network protocols, and other system facets will likely emphasize this trend even more. We leave such comprehensive analysis as future work.

We conclude by observing that while complex silicon photonic networks may not offer competitive energy efficiency for chip-to-chip scale networks under low loads, networks loaded with tens of Terabits per second are possible with this technology. It is these systems that will yield the lowest energy per bit and get the most benefit from silicon photonic links.

ACKNOWLEDGEMENTS

We gratefully acknowledge support for this work under MIT Lincoln Laboratory PO MIT-7000135026 and the U.S. Department of Energy Sandia National Laboratories PO 1426332.

REFERENCES

[1] M. Haney, R. Nair, T. Gu. "Chip-scale integrated optical interconnects—a key enabler for future high performance computing." Proc. of SPIE, vol. 8267, 2012, invited.

[2] Y. Arakawa, T. Nakamura, Y. Urino, T. Fujita. "Silicon photonics for next generation system integration platform." Communications Magazine, IEEE, vol 51., no. 3, March 2013.

[3] A. V. Krishnamoorthy, et al. "Computer systems based on silicon photonic interconnects." Proc. of IEEE, vol. 97, no. 7, July 2009.

[4] A. V. Laer, M. R. Madarbux, P. M. Watts, T. M. Jones. "Towards zero latency photonic switching in shared memory networks." SiPhotonics workshop in HiPEAC, January 2014.

[5] S. Beamer, et al. "Re-architecting DRAM with monolithically integrated silicon photonics." ISCA, June 2010.

[6] H. Wang, et al. "Nanophotonic optical interconnection network architecture for on-chip and off-chip communications." OFC, 2008.

[7] N. Ophir, C. Mineo, D. Mountain, K. Bergman. "Silicon photonic micro-ring links for high-bandwidth-density, low-power chip I/O." IEEE Micro, 2013.

[8] S. Bokhar. "Role of interconnects in the future of computing." Journal of Lightwave Technology, Vol. 31, No 24, 2013.

[9] R. Hendry, D. Nikolova, S. Rumley, N. Ophir, K. Bergman. "Physical layer analysis and modeling of silicon photonic WDM bus architectures." HiPEAC 2014 workshop: *Exploiting Silicon Photonics for energy-efficient heterogeneous parallel architectures*, January 2014. Invited.

[10] H. Thacker, et al. "Flip-chip integrated silicon photonic bridge chips for sub-picojoule per bit optical links." Electronic Components and Technology Conference, 2010.

[11] M. R. Watts, D. C. Trotter, R. W. Young, A. L. Lentine. "Ultralow power silicon microdisk modulators and switches." 5th IEEE International Conference on Group IV Photonics, September 2008.

[12] P. Dong, et al. "Low V_{pp} ultralow-energy, compact, high-speed silicon electro-optic modulator." Optics Express, vol. 17, no. 25. November 2009.

[13] Manipatruni, S., Chen, L., and Lipson, M. 2010. Ultra high bandwidth WDM using silicon microring modulators. Optics Express, vol. 18, no. 16, 2010.

[14] Xiao, S., Shen, H., Khan, M. H., Qi, M.. 2008. Silicon microring filters. *Conference on Lasers and Electro-Optics/Quantum Electronics and Laser Science Conference and Photonic Applications Systems Technologies*, OSA Technical Digest (CD) (Optical Society of America, 2008), paper JWA84.

[15] Masini, G., et al. 2012. CMOS photonics for optical engines and interconnects. *OFC/NFOEC Technical Digest* (2012).

[16] Lee, B., Biberman, A., Dong, P., Lipson, M., Bergman, K. 2008. All-optical comb switch for multiwavelength message routing in silicon photonic networks. *IEEE Photonics Technology Letters*, vol. 20, no. 10 (2008).

[17] B. Lee, A. Biberman, N. Sherwood-Droz, C. Poitras, M. Lipson, K. Bergman. "High-speed 2x2 switch for multi-wavelength silicon – photonic networks-on-chip." Journal of Lightwave Technology, vol. 27, no. 14. July 2009.

[18] Selvaraja, S. K., et al. 2009. Highly efficient grating coupler between optical fiber and silicon photonic circuit. *Conference on Lasers and Electro-Optics/International Quantum Electronics Conference,* OSA Technical Digest, (2009).

[19] Pu, M., Liu, L., Ou, H., Yvind, K., Hvam, J. 2010. Ultra-low-loss inverted taper coupler for silicon-on-insulator ridge waveguide. *Optics Communications*, vol. 283, no. 19 (October 2010).

[20] Zheng, X., et al. 2013. Efficient WDM Laser Sources Towards Terabyte/s Silicon Photonic Interconnects. *Journal of Lightwave Technology*, vol. 31, no. 15 (December 2013).

[21] J. Chan, G. Hendry, K. Bergman, L. Carloni. "Physical-layer modeling and system-level design of chip-scale photonic interconnection networks." IEEE Transactions on computer-aided design of integrated circuits and systems, vol. 30, no. 10. October 2011.

[22] A. Yariv. "Critical coupling and its control in optical waveguide-ring resonator systems." IEEE Photonic Technology Letters, Vol. 14, No. 4, (2002).

[23] D. Nikolova, R. Hendry, S. Rumley, K. Bergman. "Scalability of silicon photonic microring based switch", ICTON'2014, Graz, Austria, (2014).

[24] R. Ramaswami, K. Sivarajan, G. Sasakim, *Optical Networks,* Elsevier (2010)

[25] Bogaerts et al., 2012.Silicon microring resonators*, Laser Photonics Rev.,* vol. 6, no. 1 (2012), pp. 47-73.

[26] Zheng, X., et al. 2011. Ultra-efficiency hybrid integrated silicon photonic transmitter and receiver. *Optics Express*, vol. 19, no. 6, pp. 5172-5186 (2011).

[27] D. H. Geuzebroek, E. J. Klein, H. Kelderman, A. Driessen. "Wavelength tuning and switching of a thermo-optic microring resonator." ECIO, 2003.

[28] C. Qiu, J. Shu, Z. Li, X. Zhang, Q. Xu. "Wavelength tracking with thermally controlled silicon resonators." Optics Express, vol. 19, no. 6. March 2011.

[29] K. Padmaraju, D. Logan, X. Zhu, J. J. Ackert, A. P. Knights, K. Bergman. "Integrated thermal stablization of a microring modulator." Optics Express, vol. 20, no. 27. 2012.

[30] B. Guha, K. Preston, M. Lipson. "Athermal silicon microring electro-optic modulator." Optics Letters, vol. 37, no. 12. June 2012.

[31] Preston, K., et al. 2011. "Performance guidelines for WDM interconnects based on silicon microring resonators." CLEO (2011).

[32] Y. Ma et al., "Ultralow loss single layer submicron silicon waveguide crossing for SOI optical interconnect." *Optics Express*, vol. 21, no. 24. November (2013).

[33] D. Vantrase et al. "Corona: System Implications of Emerging Nanophotonic Technology" , ISCA'08, Washington, DC, USA (2008).