DCAF - A Directly Connected Arbitration-Free Photonic Crossbar For Energy-Efficient High Performance Computing

Christopher Nitta, Matthew Farrens, and Venkatesh Akella
University of California, Davis
Davis, CA USA
Email: cjnitta@ucdavis.edu, farrens@cs.ucdavis.edu, akella@ucdavis.edu

***Abstract*—DCAF is a directly connected arbitration free photonic crossbar that is realized by taking advantage of multiple photonic layers connected with photonic vias. In order to evaluate DCAF we developed a detailed implementation model for the network and analyzed the power and performance on a variety of benchmarks, including SPLASH-2 and synthetic traces. Our results demonstrate that the overhead required by arbitration is non-trivial, especially at high loads. Eliminating the need for arbitration, sizing the buffers carefully and retransmitting lost packets when there is contention results in a 44% reduction in average packet latency without additional power overhead. We also use an analytical model for ScaLAPACK QR decomposition and find that a 64 processor DCAF could outperform a 1024 node cluster connected with 40Gbps links on matrices up to ∼500MB in size.**

***Keywords*-on-chip networks; nanophotonics; arbitration free; directly connected;**

I. Introduction

The promise of higher bandwidth and lower energy use per bit has lead to extensive recent research into on-chip photonic networks. Numerous photonic on-chip network topologies have been proposed, and their power and performance benefits demonstrated. However, the networks proposed thus far do not appear to take full advantage of the capabilities of optical interconnects, which led us to ask the following question – *Can the unique properties and physics of on-chip photonics be exploited in a way that will allow the creation of highly desirable networks which are impractical (if not outright impossible) to build using only electronics?*

For example, the advantages of directly (fully) connected topologies are well known. They offer the highest bisection bandwidth and are far more resilient to failures on links, since packets can be routed through unaffected nodes. In a large scale on-chip multiprocessor, a fully connected network is particularly useful and desirable at the higher level of a hierarchical topology.

Can the unique properties of photonics be exploited to enable fully connected topologies with a wide data path? Optical waveguides can carry multiple signals simultaneously without interference, using a mechanism called wavelength division multiplexing (WDM). This fact has been widely exploited in the literature to realize high performance, low power networks. However, the existing research assumes that all the photonics exist on a single layer, despite the fact that it is possible to create photonic vias (grating couplers [10], [21], for example). In the same way that more complex electrical networks can be realized when multiple layers of metal are available, using photonic vias to support multilayer photonics can allow the realization of highly scalable fully connected topologies. In fact, we will show that the number of waveguides is not the major impediment to a fully connected photonic network - the limiting factor is the external laser power required to feed all the transmitters and receivers. We will present ways to circumvent this problem later in this paper.

Most (if not all) of the on-chip photonic networks proposed in the literature so far require arbitration, which is done either electrically [20] or optically [24]. Unfortunately, arbitration is a problem for two reasons - first, it is a cost that must always be paid (the arbitration overhead, in terms of power and performance, is incurred whether or not communication occurs), and second, arbitration is a possible point of failure (if any part of the arbitration network fails, the entire system is rendered useless, making the network less resilient). Therefore, the use of an arbitration-free network could result in better performance and lower power consumption, while at the same time providing more resilience. And resilience is an important topic to keep in mind when designing with new technologies (such as on-chip photonic devices), whose fabrication process is relatively immature.

Eliminating arbitration does introduce some new challenges. First, some form of flow control is required to deal with the limitation of finite buffer sizes. Second, packets may need to be buffered and re-transmitted, which could offset the power/performance benefit of not having arbitration. However, a key point to keep in mind is that flow control kicks in only when the buffers are full, making it a relatively rare event. The overhead of detecting buffer overflow and requiring retransmission is only paid *when necessary*, as opposed to arbitration which is a cost that is *always* incurred.

In this paper we introduce a directly connected topology that does not require arbitration. We call this Directly Connected Arbitration Free Topology DCAF, and it has at its core a fully connected topology (i.e. a direct optical link

between every pair of nodes).

The paper is organized as follows. In Section II we will present a fairly detailed introduction to the basic building blocks of nanophotonic interconnects that enable the topology evaluated in this paper. Next we present an overview of related work in Section III. In Section IV we describe the DCAF network and in Section V we describe the power model that was assumed. In Section VI we evaluate the performance and power consumption of DCAF and compare it with another well-known network topology. We discuss the implications of the results in Section VII and conclude in Section VIII.

II. Background

Ring Resonators: Microring resonators are designed to resonate when presented with specific individual wavelengths and remain quiescent at all other times. The ability to respond to specific wavelengths enables the removal (filtering) of specific wavelengths from a waveguide, and these resonators are the primary technology used to bundle the high quantity of wavelengths per waveguide needed for Dense Wavelength Division Multiplexing (DWDM). This filtering can be achieved using either passive or active microrings. Figure 1(a) shows a high-level view of a passive microring that is biased during fabrication to extract only λ_1 from the incoming waveguide and steer it down a perpendicular waveguide.

Since the passive microrings are biased during fabrication to always respond to a single wavelength, they cannot be used for modulation. Modulating requires an active microring resonator, which is designed to change its resonance frequency based on the amount of current present in the n^+ base. Figures 1(b) and 1(c) illustrate an active microring resonator in the "On" and "Off" states, respectively. If the electrical current is present ("On" state), λ_1 is extracted from the *input/through* waveguide and sent down the *drop* waveguide – if there is no current applied ("Off" state), λ_1 will continue down the *input/through* waveguide unaffected.

Generally, we assume that the presence of a wavelength represents a logic 1 and the absence represents a logic 0, and the method by which an active microring modulates depends upon the configuration of the incoming and outgoing waveguides. For example, if the incoming waveguide is also the outgoing waveguide, then a zero can be created by using the microring to remove the wavelength by bending it onto a dead-end drop waveguide, and a one is created by allowing the wavelength to pass unaffected (this is shown in Figure 1(b)). If the incoming and outgoing waveguides are not the same, then ones are created by bending the wavelength onto the outgoing waveguide, and zeros by allowing the wavelength to continue unperturbed along the incoming waveguide. (This is shown in Figure 1(b) if waveguide II is the outgoing waveguide, and not a dead-end drop.)

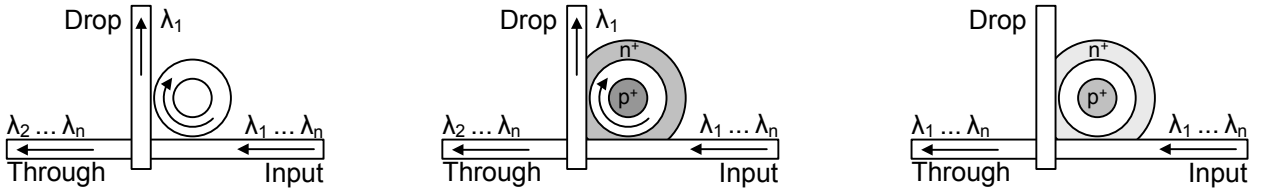
Photonic Vias: Waveguides carrying different signals can intersect on the same layer without complete signal interference, unlike wires carrying electronic signals. Intersections of waveguides at 90 degrees allow for signals traveling down each waveguide to continue on intact, although each signal will suffer a small attenuation (often modeled as $\sim$0.1dB). This characteristic of photonics has allowed on-chip optical networks to be laid out on a single layer without a need to transition to waveguides on other layers. However, the cumulative effect of a large number of intersections may make a single layer waveguide layout infeasible – therefore, waveguides may need to be routed on different layers[1] to avoid excessive intersections.

In the electronic domain signals can easily move from layer to layer using vias - transitioning photonic signals to different layers is done in a similar manner. Grating couplers are used to couple optical fibers and waveguides [10], [21], and we propose using a vertical grating coupler to connect waveguides on different layers. For this work we assumed that the signal attenuation of such a coupling is 1dB, a conservative estimate considering optical fiber and waveguide couplings of less than 1dB loss have already been demonstrated.

Grating couplers are not the only possible structure for use as a photonic via. Plasmonics have the capability to drastically change the direction of light, which could be useful when changing layers; however, plasmonics suffer from high path attenuation (typically $\sim$0.2dB/μm [2]). Over the relatively short distances required for an inter-layer via (assumed less than 10μm), the loss experienced by a plasmonic based photonic via may be acceptable; we do not investigate the possibility of using plasmonics as a photonic via, but only mention it as an example of a possible alternative to grating couplers.

Trimming: The wavelengths that individual microrings respond to are set during fabrication - however, variations in fabrication tolerances may require that certain microrings have their resonance frequency moved up or down slightly. Furthermore, the refractive index (n) of silicon changes with temperature (ΔT), which can be modeled as - $\Delta n \approx 1.84 \times 10^{-6} \times \Delta T$. As a result, microring resonators are very sensitive to temperature and drift spectrally approximately 0.09nm/°C. The resonance frequency can be "trimmed" to account for both fabrication imperfections and thermal drift, which can be accomplished dynamically by electrically injecting current (to shift the resonance towards the blue) or by heating the ring (to shift towards the red) [1]. However, these active trimming techniques can result in a dramatic increase in the overall power requirements and even thermal runaway [12]. Trimming through heating has been shown to have a non-linear relationship to microring count [12],

[1] A detailed description of multi-layer fabrication is available in Appendix A of [14], and a brief overview in the appendix of this paper.



(a) Passive Microring Resonator, tuned to resonate to λ_1

(b) Active Microring Modulator, which is "On" for λ_1

(c) Modulator Active Microring Modulator, turned "Off" for λ_1

Figure 1: Microring Resonators. (a) shows a passive microring, which at fabrication time was set to resonate only to λ_1. (b) and (c) show active microrings, which use the presence or absence of charge in the n^+ base to change the wavelength they will resonate to (λ_1, here.)

and recently researchers [3], [18] have shown a significant reduction in thermal sensitivity from that of PMMA upper cladding [25] investigated in [12]. Therefore, in this study we assume only current injection-based active trimming of microrings with a thermal sensitivity of 1pm/°C and a Temperature Control Window of 20°C.[2]

III. Related Work

Within the research community there has been a growing interest in harnessing the benefits of optics in order to address the shortcomings of electrical interconnects. In [24] HP researchers describe a 64x64 WDM based crossbar (called Corona) for a 256-core CMP. Corona uses a multiple-writer single reader crossbar architecture, which requires arbitration (realized using a distributed scheme and additional optical channels). Cornell researchers described a bus-based scheme to connect clusters of processors in [6], and more recently propose a hybrid opto-electronic on-chip network called Phastlane that uses a low complexity nanophotonic crossbar supported by an electrical network for buffering and arbitration. Phastlane uses packets with a single flit and an Automatic Repeat-reQuest (ARQ) based flow control scheme, where packets are allowed to be dropped. DCAF uses a similar flow control scheme, with the exception that it is ACK instead of NAK based.

MIT and Berkley researchers [5] propose a multistage Clos network using a mixture of electronic routers that are connected by WDM based photonic links. Clearly, this network has less flexibility and a higher average hop-count than a crossbar. Furthermore, the CMXBar described in the paper requires arbitration, which DCAF does not. The authors in [20] propose a photonic 2D torus network that employs an electrical network for arbitration and flow control. The network is evaluated on a variety of synthetic and scientific benchmarks [4] to show that the hybrid photonic torus network can achieve a factor of 37x improvement in performance per energy spent. This paper also points out that many scientific workloads exhibit communication patterns that change over time, which is another reason the directly connected nature of DCAF is so attractive.

Firefly [17] is another hybrid opto-electronic network proposal that uses an electrical network for intra-cluster communication and a nanophotonic crossbar for inter-cluster communication. The Single Writer Multiple Reader (SWMR) network discussed in [17] requires a broadcast network in order to send the head flit, and this broadcast network will require arbitration - the timing between the sending of the head flit and transmitting the data flits will also require precise delay. In addition, the broadcast network will require power, which is likely to be nearly equal to that of the SWMR crossbar itself.

The FexiShare network is a flexible photonic crossbar [16] that is a combination of a Multiple Write Single Read (MWSR) and a SWMR design. The FlexiShare network decouples the number of communication channels from the number of number of nodes, in an attempt to reduce the required photonic power. FlexiShare implements a token stream for arbitration and credit sharing, adopting the reservation assisted scheme from Firefly.

Sun Labs/Oracle researchers [7] recently investigated using silicon photonics for the interconnection network of a multi-chip system or "Macrochip". They analyzed three different photonic networks in the multi-die system that used mirrors to couple light between dies, and concluded that a statically routed point-to-point network outperformed the other networks analyzed. The point-to-point networks analyzed in [7] were limited to 2-bit site-to-site connections, which the authors admit "is a potential performance limiter". The inter-layer coupler assumed in [7] differs from our photonic vias in that the inter-layer coupler connects signals between two dies, where our photonic via couples between layers of the same die.

IV. Description of Experimental Topologies

In order to analyze and evaluate DCAF, we needed a representative network to compare it to. We wanted to compare DCAF to a flat topology which had identical total, bi-sectional, and link bandwidth, so we created the Crossbar Optical Network (CrON). CrON is modeled closely after

[2]Temperature Control Window is defined in [12] as the range of temperatures within which the network must be kept.

Table I: Corona/CrON Network Parameters

Network	Tech	WGs	Microrings		Bandwidth		
			Active	Passive	Total	Bisection	Link
Corona	17nm	257	∼1M	∼16K	20TB/s	20TB/s	320GB/s
CrON	16nm	75	∼292K	∼4K	5TB/s	5TB/s	80GB/s

the Corona design, primarily because Corona has been very carefully scrutinized over the years and there are enough details publicly available to allow it to be modeled relatively accurately. In the following section we describe CrON and how it differs from Corona.

A. Crossbar Optical Network (CrON)

The Corona design is a 64 x 64 256 bit crossbar operating at 10GHz (double clocked 5GHz). Therefore, CrON also assumes 64 nodes and a similar serpentine layout to bring the waveguides to each crossbar node, although CrON uses a bus width of 64 bits instead of 256. The decision to model a 64 instead of 256 bit data path was driven by the fact that we were modeling a 64 "core" instead of a 256 "core" system. Table I highlights the structural differences between Corona and CrON.

Arbitration in CrON is handled in a manner similar to the Token Channel with Fast Forward described in [23]. Due to the nature of the protocol, a processor can wait up to 8 clock cycles (at 5GHz) to receive an uncontested token. Increases in die area and node count will increase the serpentine waveguide length and therefore increase propagation delay, meaning that the delay for uncontested tokens will grow with increased clocking speeds, die area, and node count. (The CrON design, however, does have the capability of a simultaneous one-to-many transmission if a single node were by chance to acquire arbitration tokens for multiple receivers.) The Token Channel with Fast Forward protocol was chosen over the Token Slot since Token Slot can lead to node starvation [23]. Token Channel with Fast Forward was chosen over the Fair Slot protocol since a broadcast waveguide is required in order to support Fair Slot [23], which our detailed simulations show would require a increase in the required arbitration photonic power of a factor of 6.2.

B. Directly-Connected Arbitration-Free (DCAF)

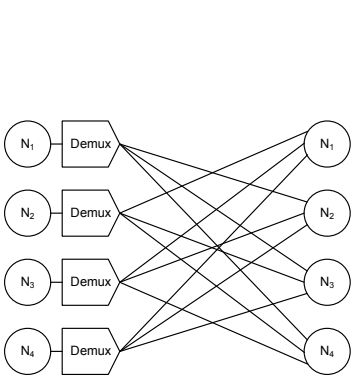
As mentioned earlier, the DCAF design features waveguides which directly connect each source/destination pair, creating a fully-connected backbone; however, DCAF incorporates additional microring resonators in the transmitter section of each node which are used to limit the number of destination nodes that can simultaneously have information sent to them to one. DCAF is in essence a many-to-one crossbar - a single node can simultaneously receive from multiple sources, but can send to only one.

Figure 2(a) shows the equivalent network connectivity for a four node DCAF. Since the dedicated links make it possible for each node to receive messages from all other nodes simultaneously, no arbitration is required. DCAF essentially has a locally controlled demultiplexer in its transmit section, while CrON has the equivalent of a receive multiplexer which must be globally arbitrated. Figure 2(b) is an example of how a 1:4 optical demultiplexer can be constructed using microring resonators. Figure 2(c) illustrates the DCAF transmitter section - in this figure λ_1 and λ_2 are being transmitted to node 2, while λ_3 is not (in other words, node 4 is transmitting a binary 011 to node 2).

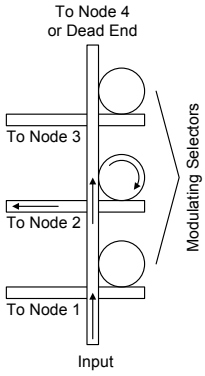
DCAF does not require arbitration in order to transmit a flit, and therefore it will not be subject to the limitations imposed by systems which require global clock synchronization. However, even though DCAF is arbitration-free, it does require flow control. This is accomplished in DCAF using an ARQ scheme. If a flit arrives at a reception node and there is no available space in the buffer, the flit is dropped and the ACK is not sent back. A Go-Back-N ARQ scheme was chosen over a conventional credit based flow control approach since multiple flits can be in flight simultaneously on a single waveguide - or, to put it another way, the round trip of a single link can be much greater than 2 cycles. The ARQ scheme allows for efficient flow control without the need for excessive buffering. Reliable communication is another benefit of using an ARQ scheme for flow control, since lost flits or potentially corrupted flits can be retransmitted.

The size of the ARQ ACK token was chosen to be 5 bits since it allows for worst case round trip propagation delay and therefore will support uninterrupted flow. It should be noted that the 5 bit sequence number per flit is not additional overhead that DCAF will incur when compared to CrON - CrON will require 6 bits to designate the flit source, which DCAF does not need to provide since DCAF has a dedicated receiver for each source.

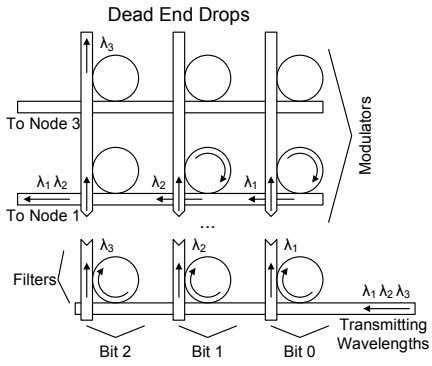
Considering the number of node connections (and hence the number of required waveguide crossings) and an assumed 0.1dB loss per intersection, a single layer implementation of DCAF would not be realizable (the creation of a very low loss intersection could make a single layer DCAF feasible, however). The use of photonic vias and multiple photonic layers, though, do enable the creation of directly connected networks like DCAF (a brief description of the multi-layer fabrication process is presented in the Appendix). It is important to do a more detailed evaluation of how DCAF might actually be laid out, of course, since the number of waveguides needed in DCAF grows quadratically



(a) Arbitration Free Crossbar (Demux ensures that a node can send to only one other node at a given time.)



(b) 1:4 Optical Demux (Signal is being output to Node 2.)



(c) TX Section for Node 4 (λ_1 and λ_2 are being transmitted to Node 2, while λ_3 is not. A binary 011 is being transmitted to Node 2.)

Figure 2: DCAF: 4 Node Network Equivalent (a), 1:4 Optical Demultiplexer (b), and Transmit Section (c)

Table II: CrON/DCAF Network Parameters

Network	Tech	WGs	Microrings		Bandwidth		
			Active	Passive	Total	Bisection	Link
CrON	16nm	75	∼292K	∼4K	5TB/s	5TB/s	80GB/s
DCAF	16nm	∼4K	∼276K	∼280K	5TB/s	5TB/s	80GB/s

with node count. This is shown in Figure 3, which presents the entire layout for a 16 node DCAF using a 16-bit bus. Assuming an 8μm ring pitch (3μm ring and 5μm ring spacing), and a 1.5μm waveguide pitch (0.5μm waveguide and 1μm waveguide spacing), the network as illustrated occupies an area of $\sim$1.15mm^2. Each color of waveguide indicates a different layer; green waveguides connect node groups in the vertical direction, while aqua waveguides connect node groups in the horizontal.

A 64 node DCAF could be constructed by clustering four groups of 16 nodes and interconnecting them in the same way 4 node clusters are interconnected in the 16 node case. Laying out a DCAF network in this fashion requires that the number of layers grow as $log_2(N)$, though fewer layers could be used at a cost of more complicated waveguide routing. Given our assumed layout technique (which routes waveguides around the microring area) a 64 node DCAF will require $\sim$58.1mm^2.

Table II illustrates the structural differences between CrON and DCAF. Note that the number of waveguides in CrON is somewhat misleading - if one considers a single loop around the chip as just one waveguide, then the number is 75; however, if one considers each segment between nodes to be a separate waveguide then there are actually $\sim$4.6K, which is more than is used by DCAF. DCAF also requires $\sim$88% more microrings than CrON, although there are in fact fewer *active* (power-consuming) microrings required in DCAF than in CrON. As stated earlier, the total, bi-sectional, and link bandwidth of the two networks are identical.

V. MINTAKA

Performing a thorough analysis of DCAF requires a detailed simulation infrastructure for photonic networks. We used the Mintaka simulator to conduct these experiments. We will briefly describe the simulator here – for more details please see [12] and [14].

The photonic power estimates in Mintaka are derived using a link loss approach similar to that used in [1] for Corona, and power levels for each possible path through a link are maintained (all photonic energy is tracked inside Mintaka). Mintaka also performs a thorough thermal analysis, which is essential to understanding the true power consumption in on-chip optical networks, since items such as microring "trimming" power and buffer leakage are functions of temperature.

Mintaka was validated by comparing the optical and electrical components separately.[3] We found that the worst case path attenuation for DCAF is 9.3dB, which is significantly lower than the 17.3dB for CrON. There are several reasons for this; primarily it is because the number of off resonance rings that photons must pass through in CrON (4095) is much higher than in DCAF (200), although another contributing factor is the fact that in CrON the worst case light path must make two passes around the serpentine in

[3]In fact, the simulator is so thorough and accurate that we discovered (with the help of several of the Corona authors [22]) that if power flows counter to that of the tokens in Corona, a gap in photonic power can occur when a token needs to be injected. This discovery in no way diminishes or negates the previous findings regarding photonic tokens, but does change the structures that must be assumed for token injection.

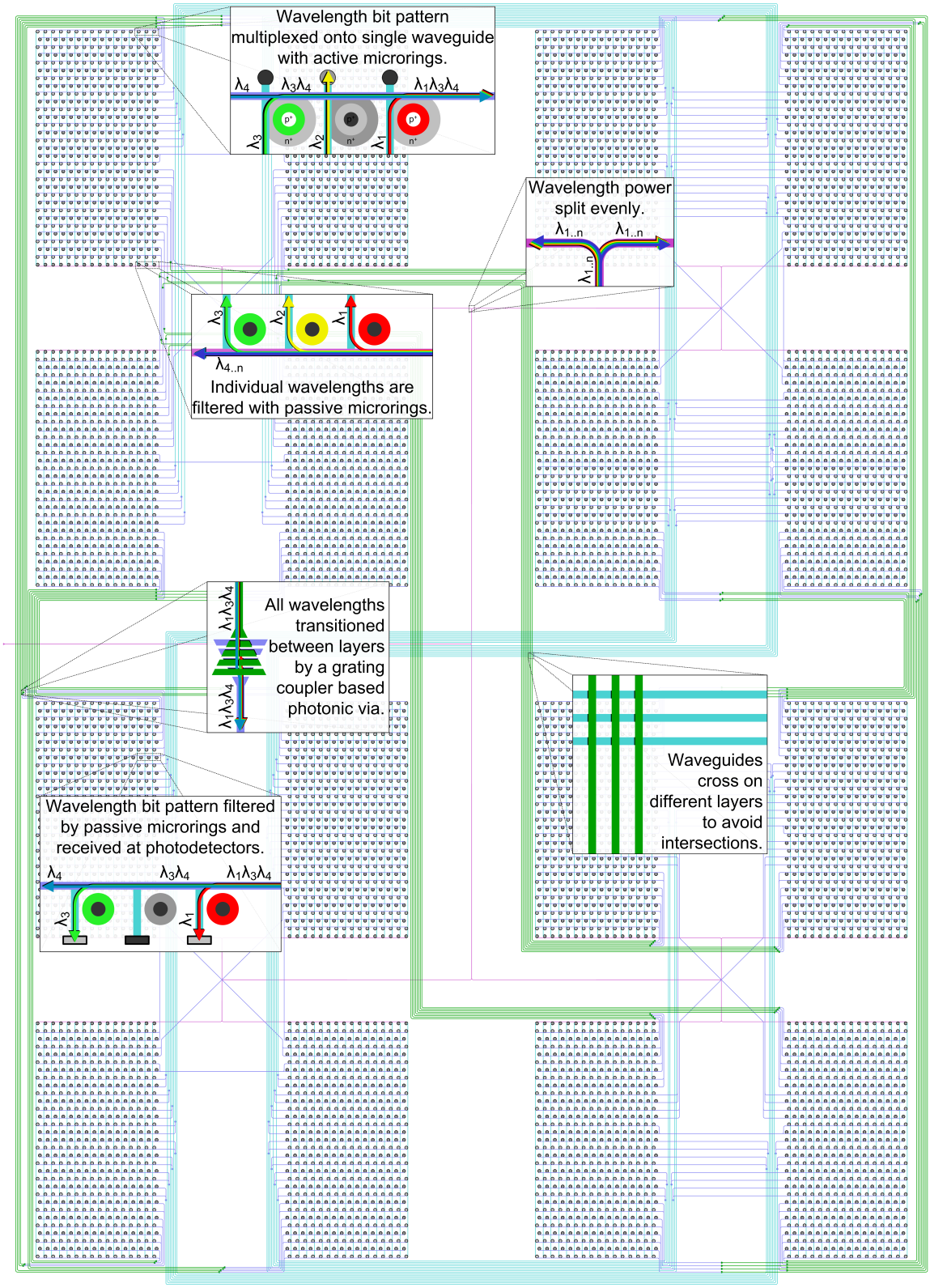


Figure 3: Entire layout for a 16 node DCAF using a 16-bit bus. Assuming an 8μm ring pitch and a 1.5μm waveguide pitch, the network as illustrated occupies an area of $\sim$1.15mm^2. The network is on its own layer, with the processing nodes directly beneath each cluster of rings. Each color of waveguide designates a different layer; green waveguides connect node groups in the vertical direction, while aqua waveguides connect node groups in the horizontal.

order to reach the receiver, while in DCAF the worst case path is much more direct.

VI. Experimental Setup

In order to evaluate the performance of DCAF we created a trace-driven network performance simulator to determine the latency, average and maximum queue depths, average and peak bandwidth, and total execution time. In [13] we showed that not including packet dependencies can yield misleading performance results, so we used the same dependency tracking simulator created in [13] and added the CrON and DCAF networks to it in order to more accurately ascertain network performance. The base architecture we modeled was a 64 node network with a 64-bit data path between nodes, built using 16nm technology. The "cores" were assumed to operate at 5GHz and be capable of generating and consuming one 128-bit flit per cycle. The on-chip network occupies an entire level of a 3D stacked processor design, with an area of 484mm^2.

The "traces" (or more correctly Packet Dependency Graphs (PDG)s) used in the performance simulations were a combination of synthetic traffic patterns and SPLASH-2 benchmarks. The synthetic traffic patterns chosen were *uniform random*, *negative exponential distribution* (NED) [19], *hotspot*, and *tornado*. All synthetic traces were run with a standard range of offered load (no dependencies) in order to determine maximum network throughput and average packet/flit latency. The SPLASH-2 benchmark PDGs used were a 16 million point FFT, Water SP, LU, Radix, and Raytrace. The PDGs were obtained from multiple 64 node full system simulations using the GEMS framework that includes the Garnet network simulator; packet dependencies were then inferred using the algorithm outlined in [13].

A. Buffering Analysis

The amount and configuration of network buffering is an important factor in analyzing the performance and power consumption of on-chip networks. The amount of transmit and receive buffering (in the form of FIFOs) at a given node alone is insufficient to determine the power/performance of the network, however - for example, one cannot assume shared buffering for all transmitters at a node in CrON, since multiple flits can be simultaneously transmitted. For the buffers to be shared, one must also include an electrical crossbar to connect the buffers to the transmitters. The same is true on the receive side in DCAF - sharing the receive buffer requires a crossbar to connect the receivers to the shared buffer. These local crossbars require N-1 input and output ports, and including the power consumed by these crossbars diminishes the power advantages of using photonics.

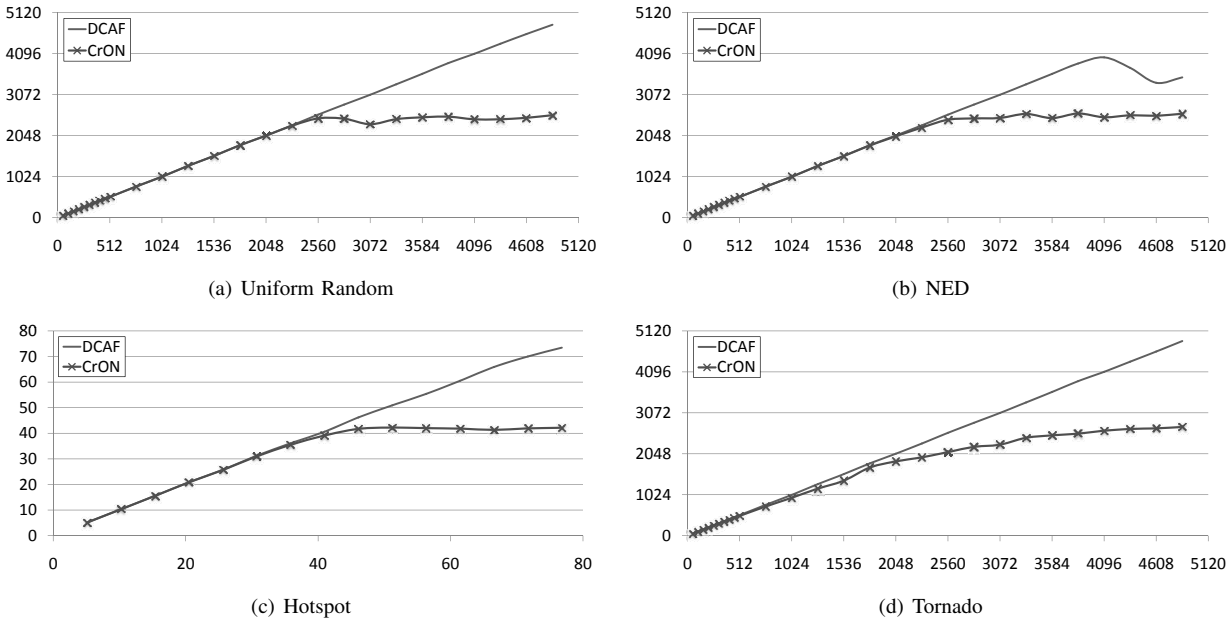
It is possible for DCAF to have a smaller local crossbar, with N-1 input ports and less than N-1 output ports; this would allow the same number of flits as output ports to be simultaneously transferred from the private buffers to a shared buffer. The same is not true of the transmit side of CrON, though, since flits must be sent sequentially once arbitration has been obtained. (DCAF can drop an incoming flit if the private buffers are full.) In our analysis we assume DCAF uses a small shared receive buffer, connected to the N-1 private receive buffers.

In CrON we assume each node has a shared receive buffer, since there is only one receiver per node. The amount of buffering must match the token size, so in order to avoid wasting photonic power the receive buffer size was chosen to be 16 flits since it evenly divides into the 64 wavelengths (this was also the assumption in [23]). DCAF does not require a private buffer for each transmitter, since only k simultaneous transmissions are possible. We assume a single shared 32 flit transmit buffer for DCAF, since it corresponds well with the ARQ scheme chosen. The small shared receive buffer also stores 32 flits, to match the size of the transmit buffer.

In order to determine the optimal amount of buffering for CrON and DCAF, the throughput of the networks with various buffering configurations was compared to that of an equivalent network with infinitely large buffers. The NED traffic pattern was used because its behavior closely approximates a real FFT application. The results of the buffering analysis showed that CrON had degraded throughput when only 4 flit buffers were employed, and had no loss in throughput when 8 flit buffers per transmitter were available. The performance of DCAF was diminished when only 2 flit buffers were used (even assuming a 2-output port local crossbar), but using a 4 flit buffer per receiver resulted in maximal throughput for the topology. Thus, the performance and power results presented in the remainder of this paper assume 8 flit buffers per transmitter and 16 flit buffers per receiver for CrON, and 32 flit transmit buffers, 4 flit receive buffers and a 32 flit shared receive buffer for DCAF. This results in a total of 520 and 316 flit buffers per node for CrON and DCAF, respectively.

B. Performance Results

The synthetic traffic "traces" provided an average offered load with an average packet size of 4 flits per packet, using a burst/lull distribution. The burst/lull injection distribution was chosen over a Bernoulli distribution since real traffic tends to be more "bursty" in nature. The throughput in GB/s is shown as a function of offered load in GB/s for DCAF and CrON in Figure 4. DCAF outperforms CrON on every one of the synthetic traffic patterns. Note that for the *hotspot* traffic pattern the offered load is limited to 80GB/s, since the maximum throughput of a single node is 80GB/s and any offered load above that is guaranteed to overwhelm any network, regardless of topology. Note also that the throughput for DCAF with the NED traffic pattern does not maintain a maximum level, but actually tapers off as



(a) Uniform Random

(b) NED

(c) Hotspot

(d) Tornado

Figure 4: Throughput (GB/s) vs. Offered Load (GB/s)

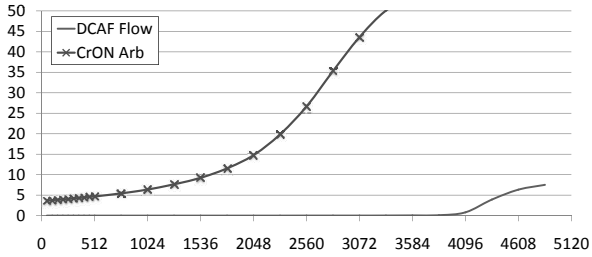


Figure 5: Latency (cycles) vs. Offered Load (GB/s) for NED Traffic Pattern

a higher load is offered. This is due to the ARQ flow control - as the offered load increases, more flits are dropped and must be retransmitted.

From the graphs it appears that DCAF performs ideally on all traffic patterns except for NED. In reality, the performance of DCAF is slightly lower than the ideal starting at 56GB/s for *hotspot* and 4096GB/s for *uniform random*. The performance of DCAF does match the ideal for *tornado*, and this would also be true for *nearest neighbor*, *transpose*, *bit inverse*, and any other synthetic traffic pattern where each destination can only receive from a single source. This holds because DCAF does not require arbitration in order to send a flit, so it is not possible for a single source to trigger the need to drop a flit.

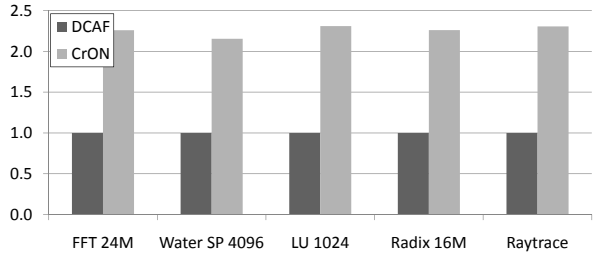
The average flit or packet latency is another common metric which is used to compare networks. We decided to look in more detail at the components of the average flit latency. Figure 5 shows the average flit latency component due to arbitration in CrON and flow control in DCAF when using the NED traffic pattern.[4] Notice that arbitration in CrON adds latency to each flit even under low loads, but the ARQ flow control in DCAF only adds latency when the network has become overwhelmed. As was stated earlier, arbitration is an overhead that must be paid for all communication, while the ARQ flow control is an "on-demand" type of penalty that is only paid when the network is overwhelmed.

The performance results of the SPLASH-2 runs are shown in Figure 6. Figures 6(a) and 6(b) show the average flit and packet latencies for DCAF and CrON, normalized to the network with the lowest latency (in all cases DCAF). The figures show that DCAF has dramatically lower average latencies across all the benchmarks; however, the lower latency does not result in as dramatic a difference in the overall execution time.

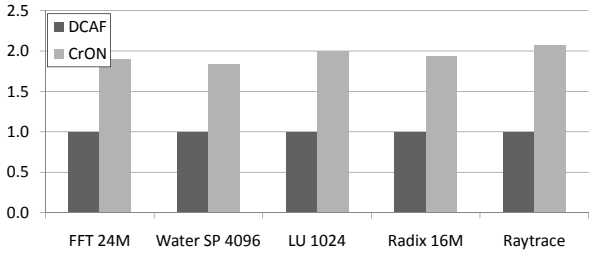
Figure 6(c) shows the execution time of each benchmark normalized to the shortest execution time, and the figure shows that DCAF executed the benchmarks from 1% to 4.6% faster than CrON. The reader may be left wondering why reducing the latency by a factor of 2 would result in such a small decrease in execution time; the answer is that the average required network throughput for the benchmarks is quite low when compared to the networks capabilities.

Figure 6(d) shows the average throughput in GB/s for the various benchmarks. The average throughput of the SPLASH-2 benchmarks equates to $\sim$0.4% of the total

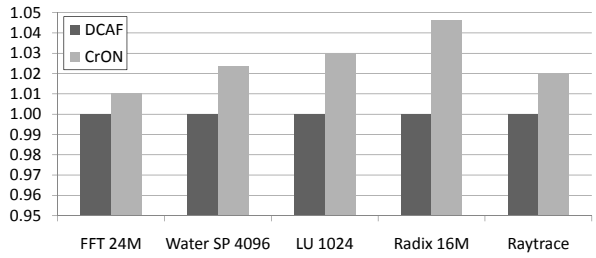
[4]NED was chosen because the flow control component in DCAF is by far the highest in NED - it is negligible in the other traffic patterns.



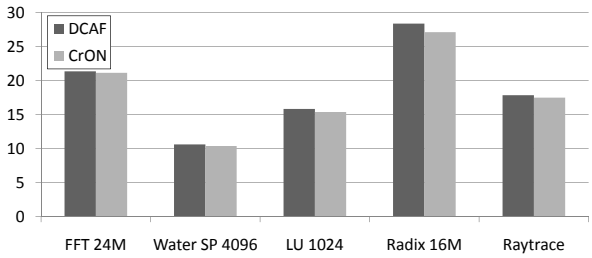
(a) Normalized Flit Latency



(b) Normalized Packet Latency



(c) Normalized Execution Time



(d) Average Throughput (GB/s)

Figure 6: SPLASH-2 Performance Results

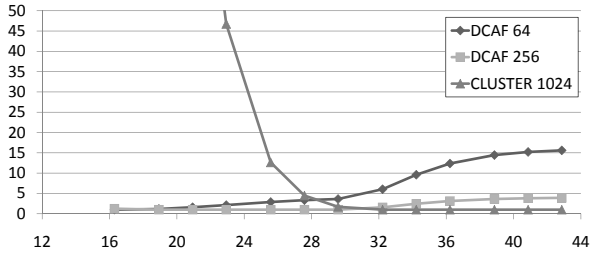


Figure 7: Normalized Execution Time vs. log_2(Matrix Size)

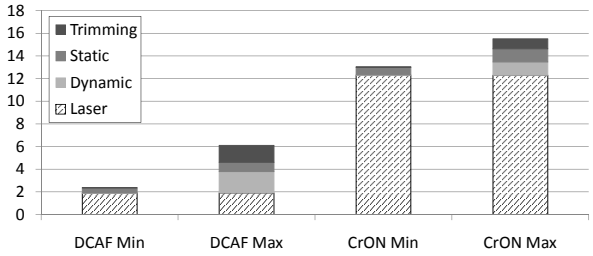


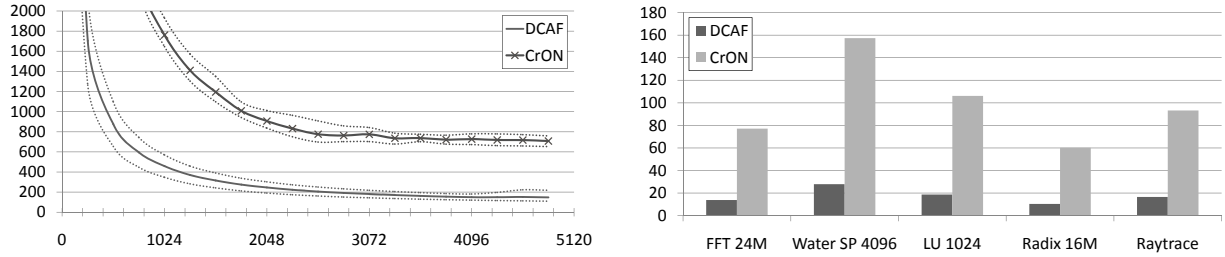
Figure 8: Power (W) vs. Network (Min/Max Load)

network bandwidth. While it may at first appear that the networks are over-designed, it is important to note that the average of the *peak* throughputs attained on the benchmarks was ∼25.3% of the total network bandwidth for CrON and ∼99.7% for DCAF. At some point during the execution on DCAF the maximum network throughput was obtained on every benchmark except for Radix, meaning that there are critical points at which all the network bandwidth is utilized.

In addition, there are many programs in the scientific realm that can easily benefit from lowering the communication cost. For example, we decided to look at how QR decomposition might perform on DCAF. Figure 7 shows the estimated normalized execution time for a QR decomposition using ScaLAPACK as a function of matrix size. The results were generated assuming a single level 64 node DCAF, a two level 256 node DCOF and a 1024 node cluster of processors connected with 5GB/s links. The results show that the improved performance of DCOF can significantly decrease the execution time for QR decomposition, even when fewer computational nodes are used. This is an example of why we believe strongly that one must be careful not to unwisely restrict the flexibility of tomorrow's on-chip processor network based on the results of running yesterday's parallel processing benchmarks.

C. Power Results

The minimum and maximum power consumption for DCAF and CrON is shown in Figure 8. The minimum power consumption is the minimum power that must be consumed even when the network is idle and at its lowest ambient temperature, while the maximum power is the maximum observed across all the simulations. The dominant factor for both networks is the laser power, which is consumed regardless of activity. The reader may notice that CrON also consumes dynamic electrical power even when idle; this is due to the fact that arbitration tokens must be replenished every loop, requiring modulation of the arbitration microrings.



(a) Offered Load (GB/s)

(b) SPLASH-2 Benchmark

Figure 9: Energy Efficiency in (fJ/b) vs. Offered Load (GB/s) (a) and in (pJ/b) vs. SPLASH-2 Benchmark (b)

As one might expect, the overall maximum trimming power required for DCAF is higher than for CrON, since DCAF has ∼88% more microrings. However, the average trimming power *per microring* is actually 18% higher for CrON. In [12] we observed that the heating power required for trimming has a non-linear relationship with microring count, and our findings show that current injection has a non-linear relationship as well. CrON requires more trimming power per microring since the network operates at a higher temperature due to the greater power consumption when compared to DCAF.

The maximum amount of dynamic power consumed by DCAF is much higher than that of CrON, but DCAF also greatly outperforms CrON in the maximal case. Figure 9(a) shows the energy efficiency (in fJ/b) as a function of offered load (in GB/s). The energy efficiency shown in Figure 9(a) is calculated by taking the power consumed divided by the actual network throughput (not the theoretical maximum throughput). The solid lines for DCAF and CrON are the average energy efficiencies (the average power consumed divided by average throughput). The dotted lines show the minimum and maximum energy efficiencies for the two networks; the efficiency varies with achieved throughput and ambient temperature. DCAF is clearly more energy efficient than CrON, and the result is most apparent under high offered load (since CrON is unable to actually achieve higher throughputs). In the best case DCAF and CrON approach 109 and 652 fJ/b respectively, although this only occurs under high load.

The energy efficiencies that can be obtained by DCAF and CrON under high load are not observed when the networks execute the SPLASH-2 benchmarks, which can be seen in Figure 9(b). The average energy efficiency for DCAF and CrON on the SPLASH-2 benchmarks was 24.1 and 104 pJ/b. The lower energy efficiency observed in these photonic networks under low load is a problem that will likely be shared with future on-chip electrical networks; while electric networks will not have the static laser overhead, the static electrical leakage is of greater and greater concern as we move from deep submicron into nanoscale technologies.

A network with lower performance may have the potential for higher energy efficiency, but a lower performing network will also impact the energy efficiency of the cores and caches due to the increased number of stalled cycles. Examining the impact of network performance on the energy efficiency of the cores is beyond the scope of this work.

VII. Discussion

Average energy efficiency is a common concern among computer architects. As was shown in the previous section, the average throughput of the SPLASH-2 benchmarks is very low compared to the total network bandwidth, and this low average throughput leads to low average energy efficiency. However, reducing the capabilities of the network is not necessarily desirable, since the entire network bandwidth *is* utilized at certain points in the benchmarks. The main reason for the energy inefficiency at low load is the large amount of static power overhead (the static leakage and fixed laser power). Reducing the static leakage power is a well-studied area, but the approach of reducing the fixed laser power or adjusting it to match the workload has not yet been examined.

At this point scaling the laser power is not a viable option, since lowering the incoming laser energy uniformly drops the power on all links. However, it is possible the unused energy could be recaptured – the photons not used to communicate could be captured and turned into electricity. Converting the unused photons to electrons would be relatively straightforward, requiring only the modification of existing photodiode structures. The number of photons available for recapture is a function of the activity occurring on each wavelength, which is related to the workload and the distribution of ones and zeros. We are currently examining the costs and benefits of taking such an approach.

Another common concern of architects is the scalability of network topologies. A 64-bit DCAF with 128 nodes will require an area of $\sim$293mm^2, but a 256 node DCAF would require $\sim$1,650mm^2. The photonic power of DCAF does not scale linearly either, although there is a less than 5% increase in required channel power scaling from 64 to 128 nodes. A 64-bit CrON with 256 nodes will require a smaller area ($\sim$323mm^2), but the photonic power requirements of

Table III: 16x16 All-Optical Hierarchical DCAF Network Parameters

Component	WGs	Microrings Active	Microrings Passive	Area (mm^2)	Bandwidth Total	Photonic Power (W)
Local Node	N/A	1,120	1,190	0.177	80GB/s	0.016
Local Network	272	∼20K	∼19K	3.01	∼1.3TB/s	0.277
Global Node	N/A	1,050	1,120	0.165	80GB/s	0.017
Global Network	240	∼16K	∼18K	2.65	1.25TB/s	0.277
Entire Network	∼4.5K	∼314K	∼334K	55.2	20TB/s	4.71

CrON will likely prevent it from scaling to even 128 nodes. The number of off-resonance rings which light must pass through will roughly double when scaling CrON from 64 to 128 nodes, and this fact alone will increase the path attenuation by over 6dB. Our estimates show that a 128 node CrON would require over 100W of photonic power. While the scalability of DCAF is limited to 128 nodes, CrON is limited to half that.

The bandwidth capability of DCAF is likely sufficient to support multiple cores per network node. As was shown by the SPLASH-2 benchmark performance results, the average network utilization of modern benchmarks is quite low. It is probable that an architect would choose to electrically cluster multiple cores per node, as was done in [24], and then use DCAF to connect those clusters.

If creating a hierarchical network is the chosen method for scaling, then connecting multiple smaller DCAF networks in a hierarchy may be a better solution. If the goal is to support 256 nodes, for example, a designer could either use the existing DCAF network and electrically cluster 4 cores at each node, or DCAF could be used to connect 16 cores, and then these 16-core nodes could be connected using another level DCAF network. The local networks would have 17 nodes (16 cores plus one connection to the global network).

Table III presents more detailed information about the all-optical 16x16 DCAF hierarchy. Notice that the required photonic power is less than 4x that of the 64 node DCAF – this is due to the reduction of off-resonance rings though which the light must travel in the smaller networks, as well as the shortening of the worst case paths because of the network hierarchy. Another counter intuitive result is that the required area is reduced while the microring count increases – this is due to the fact that the area calculation takes into account the waveguides surrounding the perimeter of each node, and the number of waveguides that must surround each node in the hierarchical is much smaller than in the 64 node case.

When comparing the average hop count and the energy efficiency of the two configurations, the all-optical DCAF network appears to have a slight edge over the hybrid network. The average hop count is 2.88 and 2.99 for the 16x16 node hierarchical DCAF and four core electronically clustered 64 node DCAF, respectively. The energy efficiency for the 16x16 will approach 259fJ/b, while the 4x64 would approach 264fJ/b; these numbers are very close, but it is important to note that the electrically clustered network value does not take into account the energy needed by the repeaters (and repeaters will be required to get the signal to the optical interface, since according to the equations in [11] the furthest a 10GHz signal can be sent in 16nm is $\sim 600\mu$m). In fact, the need to get the electrical signals to the optics is a significant challenge, one that has not been addressed thus far in the literature – we are currently investigating this issue.

VIII. Conclusions

In this paper we have shown that by using multiple photonic layers, it is possible to build directly connected arbitration free photonic crossbars (DCAF). Since DCAF is realized by restricting the number of transmitters in a fully-connected network, it offers reliability and the opportunity to scale its bandwidth for future workloads by increasing the number of transmitters per node. We have presented a power and performance analysis of DCAF on a variety of workloads, including synthetic traces, SPLASH-2 benchmarks and a QR decomposition (an important linear algebra kernel). We have shown the value of using flow control instead of arbitration, exploiting the fact that arbitration is an overhead that is incurred whether or not it is needed, while flow control is a penalty only when the network is overwhelmed. We observed that even though DCAF and a comparison network (CrON) have identical link, bi-sectional and total bandwidth in theory, DCAF performs better than CrON while simultaneously consuming less power.

We found that the energy efficiency of both networks under low load is dramatically lower than it is under high load; however, DCAF reached maximum total throughput on all but one of the SPLASH-2 benchmarks, indicating that there are certain points at which all the available network bandwidth is utilized. In addition, it is important to remember that the SPLASH-2 benchmarks are old, and one has to be very careful not to make the critical error of designing tomorrow's machine using yesterday's programs.

References

[1] J. Ahn, M. Fiorentino, et al. Devices and architectures for photonic chip-scale integration. *Applied Physics A: Materials Science & Processing*, 95:989–997, June 2009.

[2] J. Dionne, L. Sweatlock, et al. Silicon-based plasmonics for on-chip photonics. *Selected Topics in Quantum Electronics, IEEE Journal of*, 16(1):295 –306, jan.-feb. 2010.
[3] B. Guha, B. B. C. Kyotoku, and M. Lipson. Cmos-compatible athermal silicon microring resonators. *Opt. Express*, 18(4):3487–3493, Feb 2010.
[4] G. Hendry, S. Kamil, et al. Analysis of photonic networks for a chip multiprocessor using scientific applications. *Networks-on-Chip, International Symposium on*, 0:104–113, 2009.
[5] A. Joshi, C. Batten, et al. Silicon-photonic clos networks for global on-chip communication. In *NOCS '09: Proceedings of the 2009 3rd ACM/IEEE International Symposium on Networks-on-Chip*, pages 124–133, Washington, DC, USA, 2009. IEEE Computer Society.
[6] N. Kirman, M. Kirman, et al. Leveraging optical technology in future bus-based chip multiprocessors. In *MICRO 39: Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture*, pages 492–503, Washington, DC, USA, 2006. IEEE Computer Society.
[7] P. Koka, M. O. McCracken, et al. Silicon-photonic network architectures for scalable, power-efficient multi-chip systems. In *Proceedings of the 37th annual international symposium on Computer architecture*, ISCA '10, pages 117–128, New York, NY, USA, 2010. ACM.
[8] Y. Kokubun. Vertically coupled microring resonator filter for integrated add/drop node. *IEICE TRANSACTIONS on Electronics*, E88-C(3):349–362, Mar 2006.
[9] B. Little, S. Chu, W. Pan, D. Ripin, T. Kaneko, Y. Kokubun, and E. Ippen. Vertically coupled glass microring resonator channel dropping filters. *Photonics Technology Letters, IEEE*, 11(2):215–217, Feb 1999.
[10] G. Maire, L. Vivien, et al. High efficiency silicon nitride surface grating couplers. *Opt. Express*, 16(1):328–333, 2008.
[11] A. Naeemi, J. Xu, et al. Optical and electrical interconnect partition length based on chip-to-chip bandwidth maximization. *Photonics Technology Letters, IEEE*, 16(4):1221 –1223, 2004.
[12] C. Nitta, M. Farrens, and V. Akella. Addressing system-level trimming issues in on-chip nanophotonic networks. In *High Performance Computer Architecture, 2011. HPCA 2011. IEEE 17th International Symposium on*, Feb. 2011.
[13] C. Nitta, K. Macdonald, M. Farrens, and V. Akella. Inferring packet dependencies to improve trace based simulation of on-chip networks. In *Networks-on-Chip (NOCS), 2011 Fifth ACM/IEEE International Symposium on (to appear)*, May 2011.
[14] C. J. Nitta. *Design and Analysis of Large Scale Nanophotonic On-Chip Networks*. PhD thesis, University of California, Davis, 2011.
[15] S. Pae, T. Su, J. Denton, and G. Neudeck. Multiple layers of silicon-on-insulator islands fabrication by selective epitaxial growth. *Electron Device Letters, IEEE*, 20(5):194–196, May 1999.
[16] Y. Pan, J. Kim, and G. Memik. Flexishare: Channel sharing for an energy-efficient nanophotonic crossbar. In *High Performance Computer Architecture, 2010. HPCA 2010. IEEE 16th International Symposium on*, Jan. 2010.
[17] Y. Pan, P. Kumar, et al. Firefly: illuminating future network-on-chip with nanophotonics. *SIGARCH Comput. Archit. News*, 37(3):429–440, 2009.
[18] V. Raghunathan, W. N. Ye, et al. Athermal operation of silicon waveguides: spectral, second order and footprint dependencies. *Opt. Express*, 18(17):17631–17639, Aug 2010.
[19] A.-M. Rahmani, I. Kamali, et al. Negative exponential distribution traffic pattern for power/performance analysis of network on chips. In *VLSID '09: Proceedings of the 2009 22nd International Conference on VLSI Design*, pages 157–162, Washington, DC, USA, 2009. IEEE Computer Society.
[20] A. Shacham, K. Bergman, and L. P. Carloni. On the design of a photonic network-on-chip. In *NOCS '07: Proceedings of the First International Symposium on Networks-on-Chip*, pages 53–64, Washington, DC, USA, 2007. IEEE Computer Society.
[21] D. Taillaert, P. Bienstman, and R. Baets. Compact efficient broadband grating coupler for silicon-on-insulator waveguides. *Opt. Lett.*, 29(23):2749–2751, 2004.
[22] D. Vantrease and N. Binkert. personal communication about photonic token design in Corona, 2011.
[23] D. Vantrease, N. Binkert, et al. Light speed arbitration and flow control for nanophotonic interconnects. In *Micro-42: Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*, pages 304–315, New York, NY, USA, 2009. ACM.
[24] D. Vantrease, R. Schreiber, et al. Corona: System implications of emerging nanophotonic technology. In *ISCA '08: Proceedings of the 35th International Symposium on Computer Architecture*, pages 153–164, Washington, DC, USA, 2008. IEEE Computer Society.
[25] L. Zhou, K. Okamoto, and S. Yoo. Athermalizing and trimming of slotted silicon microring resonators with uv-sensitive pmma upper-cladding. *Photonics Technology Letters, IEEE*, 21(17):1175–1177, Sept.1, 2009.

APPENDIX

This appendix includes a brief discussion of how vertically coupled microrings and multiple layers of photonics can be fabricated - a much more detailed description is available in Appendix A of [14]. Vertically coupled microring resonators can be built on silica material with controlled coupling efficiency and signal routing flexibility [8], [9] – however, it is more difficult to realize them in silicon material because of its high index contrast and the lack of a deposition method for crystalline silicon. Therefore, it is assumed in this discussion that epitaxial growth is used to stack several layer of crystalline silicon as material platform for our microring resonator-based optical network [15]. Photonic layers are created by patterning the waveguides using photolithography and Reactive Ion Etching; this is followed by a Plasma Enhanced Chemical Vapor Deposition (PECVD) to cover the whole wafer with SiO_2. Each additional layer of silicon starts with a Chemical Mechanical Polishing (CMP) to eliminate the surface fluctuation. An oxide etch is then used to expose a silicon island which provides the seed needed for the silicon epitaxial growth. After crystalline silicon is grown to cover the whole wafer, the surface is again planarized by CMP, readying the wafer for waveguide patterning again.