Reuse Distance Based Circuit Replacement in Silicon Photonic Interconnection Networks for HPC

Ke Wen, David Calhoun, Sébastien Rumley, Xiaoliang Zhu, Keren Bergman
Electrical Engineering
Columbia University

Lian-Wee Luo, Michal Lipson
Electrical and Computer Engineering
Cornell University

Yang Liu, Ran Ding, Tom Baehr-Jones, Michael Hochberg
Electrical and Computer Engineering
University of Delaware

***Abstract*—Optical interconnects, which support the transport of large bandwidths over warehouse-scale distance, can help to further scale data-movement capabilities in high performance computing (HPC) platforms. However, due to the circuit switching nature of optical systems and additional peculiarities, such as sensitivity to temperature and the need for wavelength channel locking, optical links generally show longer link initialization delays. These delays are a major obstacle in exploiting the high bandwidth of optics for application speedups, especially when low-latency remote direct memory access (RDMA) is required or small messages are used.**

These limitations can be overcome by maintaining a set of frequently used optical circuits based on the temporal locality of the application and by maximizing the number of reuses to amortize initialization overheads. However, since circuits cannot be simultaneously maintained between all source-destination pairs, the set of selected circuits must be carefully managed. This paper applies techniques inspired by cache optimizations to intelligently manage circuit resources with the goal of maximizing the circuit successful 'hit' rate. We propose the concept of "circuit reuse distance" and design circuit replacement policies based on this metric. We profile the reuse distance based on a group of representative HPC applications with different communications patterns and show the potential to amortize circuit setup delay over multiple circuit requests. We then develop a Markov transition matrix based reuse distance predictor and two circuit replacement policies. The proposed predictor provides significantly higher accuracy than traditional maximum likelihood prediction and the two replacement policies are shown to effectively increase the hit rate compared to the *Least Recently Used* policy. We further investigate the tradeoffs between the realized hit rate and energy consumption. Finally, the feasibility of the proposed concept is experimentally demonstrated using silicon photonic devices in an FPGA-controlled network testbed.

Keywords— circuit switched; reuse distance; silicon photonics; initialization; replacement; cache; RDMA

I. INTRODUCTION

Increased parallelism and data intensity have added to the communication demands within high performance computing (HPC) systems. Silicon photonic (SiP) interconnects [1-3], which have been shown to provide large bandwidth densities at high energy efficiencies, have the potential to significantly reduce the latency of data transmission. Moreover, due to the low loss of optical media, high bandwidth data movement can scale over warehouse distances and provide end-to-end connectivity across HPC platforms [4].

Despite these advantages, SiP interconnects also have a set of peculiarities and special operation requirements. For example, resonance based devices such as microring resonators will require wavelength tuning to reach the designed operating wavelengths [5]. Resonator devices are also sensitive to surrounding environmental temperature due to the high thermal-optic constant of silicon, and thus require thermal stabilization or re-initialization [6]. These requirements add up to longer link initialization delays compared to electronic links. The link initialization delays further increase with cascaded switching components.

The optical interconnect system delays directly add to the execution time of HPC applications and are a major obstacle in exploiting the high bandwidth of optics for application speedup. In particular, the latency penalty could be especially detrimental in scenarios when remote direct memory access (RDMA) [7-9] is enabled or when small messages are used.

Compared with traditional two-sided communications, RDMA mitigates synchronization overheads such as tag matching between the sending and receiving processes [10]. Instead, communication can be initiated by only one side using tools such as MPI One-sided [11-13], OpenSHMEM [14], PGAS [15], etc. RDMA enables a process to directly access remote memory space of another without involvement of the latter. To maximize this advantage, physical layer communication with small initialization overhead is desirable. Increasing setup delays can easily negate the advantages of RDMA, making the link similar in performance to having two sided synchronization delay. Instead, latencies for remote and local memory access latencies should be unified as in a flattened memory architecture [14]. Thus it is critical for SiP circuit-switched networks to provide circuits in such way that an RDMA request can immediately find a corresponding circuit upon arrival (we call this a *circuit hit*). Otherwise, the request sees a *circuit miss* and has to suffer from the setup penalty. The role of circuits in a circuit-supported RDMA system resembles that of caches in microarchitectures.

Although the initialization delay of SiP devices is hard to minimize presently due to the limitation of the silicon thermal constant [16], such penalty can be overcome through careful architectural design. One such method is to explore the temporal locality in an application's communication pattern,

978-1-4799-5860-3/14 $31.00 © 2014 IEEE
DOI 10.1109/HOTI.2014.20

where a node could reference a remote memory space multiple times within a short period. If the circuit corresponding to the requested end point already exists, messages can be immediately transmitted, avoiding the circuit setup penalties. Taking advantage of temporal locality, a set of optical circuits can be maintained for the frequently accessed neighbors, significantly increasing the circuit hit rate and hence the application performance. Maximizing the number of reuses of these circuits also helps amortize their initialization overheads.

A requested circuit, of course, will not always exist to be reused. This is because optical connections cannot be maintained for all source-destination pairs, and because application communication patterns can change over time. The challenge then becomes to carefully select and update the set of circuits to maintain.

This paper applies techniques inspired by cache optimizations to intelligently manage circuit resources with the goal of maximizing the *circuit hit rate*. We propose the concept of "*circuit reuse distance*" and design circuit replacement policies based on this metric in order to avoid circuit setup penalty. Since our work focuses on optimizing replacement performance at runtime, an estimation of the next reuse distance of a circuit is needed. We propose two predictors for predicting the circuit reuse distance and show that a novel *Transition Matrix Based Predictor* (TMBP) can provide up to 40% accuracy gain compared to the traditional *Maximum Likelihood Based Predictor* (MLBP). Two reuse distance-based replacement policies are also studied: the *Farthest Next Use* (FNU) policy and the *Minimum Reuse Score* (MRS) policy. Simulations based on scientific benchmarks show that both policies have the potential to achieve much higher hit rates than the *Least Recently Used* policy. Considering the distinction between circuits and caches, we also investigate the tradeoff between the hit rate and energy consumption. Finally, we collect data on the circuit setup delay using an FPGA-controlled network testbed containing the latest SiP devices.

The remaining content is organized as follows: Section II reviews related works and introduces the definition of circuit reuse distance. Section III profiles the distribution of reuse distances based on a set of scientific HPC benchmarks and shows the evidence of temporal locality in circuit uses. This profiled information leads us to the design of online reuse distance predictors (Section IV), which are key mechanism in our circuit replacement policies detailed in Section V. In Section VI, we investigate the tradeoff between the hit rate and energy consumption of circuits. In Section VII, we demonstrate the viability of our circuit management techniques and estimate circuit setup times with a physical layer SiP network testbed. Section VIII concludes the paper.

II. Related Work And Definitions

A. Initialization Delays on Silicon Photonic Links

Photonic elements suitable for next generation optical interconnects--such as waveguides, low-crosstalk crossings, modulators, and detectors--have been demonstrated on the silicon photonic platform. Of particular interest are microring resonator based devices that leverage the high index contrast between silicon and silicon-on-insulator to push the limits in terms of footprint and power efficiency. Microring resonators also exhibit intrinsic capability for wavelength division multiplexing, which is necessary for high bandwidth density.

Conversely, the reliance on resonance makes microring devices highly sensitive to environmental temperature and fabrication variation. The high thermal optic coefficient of silicon means that device resonance varies strongly with temperature, which must be maintained within sub-kelvin accuracy for normal system operation. To overcome this thermal dependence a variety of methods have been demonstrated [6], with very promising results using active control systems to drive a local integrated heater to maintain temperature stabilization. Active control systems increase the circuit setup latency because they require time to stabilize [16].

Fabrication variation causes an inherent offset between realized microring resonances and the laser wavelengths in the system. The overall silicon photonic control system must be also capable of initializing microring elements to their operating wavelengths in an operation called wavelength locking. The time required to do so is on the order of tens to hundreds of microseconds and is limited by the thermal time constant of the device [16]. It should be noted that most currently demonstrated control systems require optical power going into the circuit to maintain stability. When the path is turned off device temperature will no longer be stable and can require re-initialization. This is another motivation to selectively maintain active circuits and maximize reuse.

B. Reuse Distance for Circuits and Cachelines

Efforts have been made to explore the use of optical circuits in HPC environments [17, 21] or for memory accesses [18-20]. However, few of these works consider minimizing setup penalties based on temporal circuit reuse patterns. Authors in [21] attempt to hide the setup overhead by proposing an "asynchronous circuit programming" model, which requires programmers to explicitly insert circuit setup commands into the code before communication. While this method can help to overlap circuit setup delays with computations, it adds to the burden of programmers and the circuit management is not optimized. In this work, we propose investigating temporal reuse patterns of circuit communications and use the metric of "reuse distance" to optimize circuit management.

While reuse distance has been an important metric in cache performance optimizations [22], its concept can also apply to circuit-switched networks, especially when each node is allowed to maintain multiple transceivers. Such resource redundancy is enabled by the large-scale integration achievable via SiP technologies.

In this work we consider circuit reuses from the perspective of the source node. A *reuse distance* of a circuit *C* is defined as the number of circuit requests made by its source node *S* between two consecutive uses of *C* by *S*. For example, if the sequence of circuit requests made by *S* is *C, A, B, F, E, C* (labeled by destination nodes), then the reuse distance of circuit *C* is 4. We also call the outgoing circuits simultaneously maintained by a source node its *circuit set*. Despite this specific perspective, the proposed techniques can also apply in view of destination nodes or the entire network.

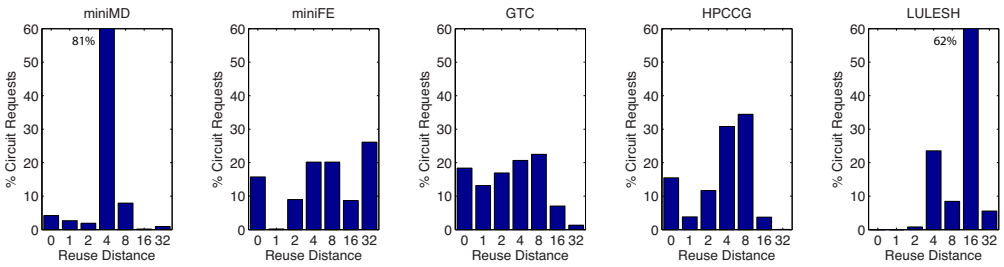

Fig. 1. Distribution of reuse distances for HPC benchmarks (64 nodes). Each bin corresponds to a range between its own label (included) and the next label (excluded), same for Fig. 2 and 3.

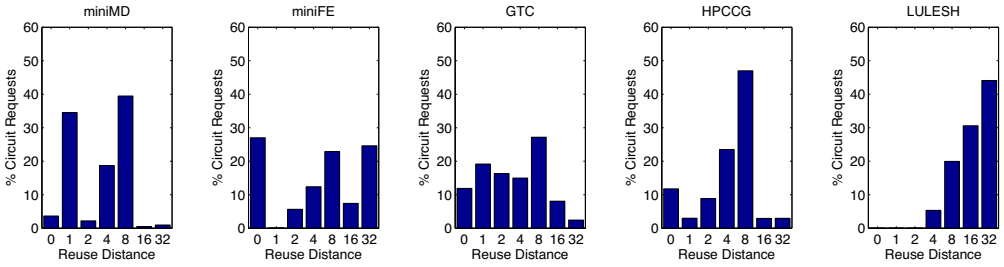

Fig. 2. Distribution of reuse distances for HPC benchmarks (256 nodes); 512 nodes for LULESH.

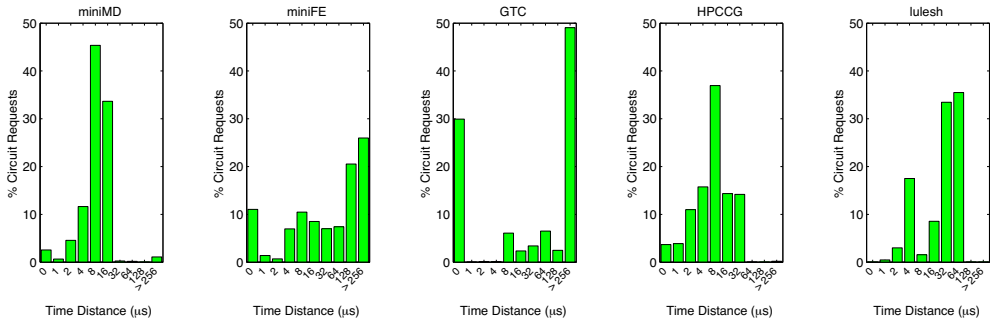

Fig. 3. Distribution of time-based reuse distances for HPC benchmarks (64 nodes). For miniMD, GTC and HPCCG, a high percentage of circuit reuses are within 16 μs.

The reuse distance of circuit *C* can be also measured in time, i.e. the time elapsed between two consecutive uses of *C*. However, the count-based reuse distance (short as *reuse distance* hereafter) and time-based reuse distance (short as *time distance* hereafter) play different roles in different optimization problems. Similar to cache optimization, we use reuse distance for circuit replacement design. However, a difference between circuits and cachelines is that circuits consume static power due to use of lasers. While flushing a cacheline before a miss makes little sense, turning off circuits that are not likely to be used in near future could potentially save energy. Here we use the time distance for optimizing the tradeoff between circuit hit rate and energy consumption, because the time distance is directly related to the energy cost for circuit maintenance.

III. PROFILING CIRCUIT REUSE DISTANCE

For measuring the temporal locality in scientific applications, we start by analyzing the distribution of reuse distances based on a group of representative HPC applications (whose description is in Appendix A). This will guide us in designing prediction and replacement policies.

As a node progresses through its workload, it issues communication requests. The node counts the number of circuit requests it makes and maintains a table to keep track of the last request index for each of its circuits. Upon a new circuit request, the entry corresponding to the circuit is consulted, and the difference between the current and the last request index of the circuit is a *sample* of the reuse distance of the circuit. By collecting such samples along program execution, an estimation of the reuse distance distribution is obtained. A fine-grained estimation of the distribution is not necessary. Instead, we cover a wide distance range and use power-of-two based bin divisions, i.e. [0], [2^0], [2^1, 2^2), [2^2, 2^3), etc. The time distance can be similarly collected based real time elapsed on the node.

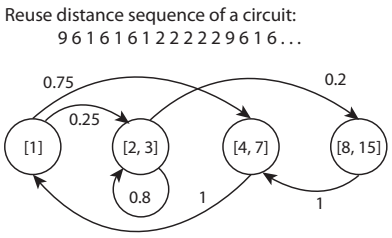

Fig. 4. Example for Transition Matrix Based Preditor. Upper: reuse distance sequence of a circuit. Lower: modeling of the sequence transition using a Markov chain. Each state of the Markov chain corresponds to a bin in the distribution histogram.

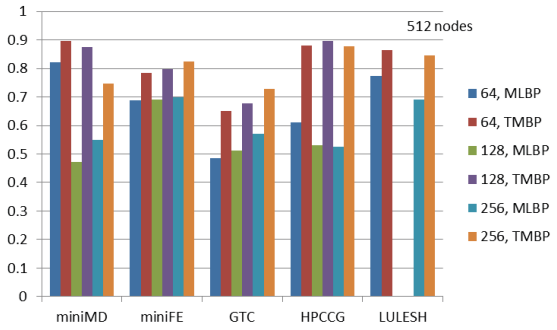

Fig. 5. Reuse distance prediction accuracy of Transition Matrix Based Preditor (TMBP) versus Maximum Likelihood Based Predictor (MLBP), across different benchmarks and different numbers of nodes. TMBP shows as much as 40% and 36% higher accuracy than MLBP in cases of miniMD and HPCCG, respectively.

A. Profiling Results

The resulting distance histograms are shown in Fig. 1, 2 and 3. Each application leads to a different reuse pattern. Applications such as miniMD show very nonuniform distributions, while some others (e.g. GTC) are more uniform. Such difference is related to the application's communication degree (i.e. the number of nodes toward which a given node issues most of its traffic), as well as irregularity of the communication pattern. The results show a high probability that a source node will reuse its circuits within a small distance. For instance, reuse distances in miniMD with a value smaller than 8 comprise 90% of the samples, while this percentage is of 43%, 70% and 60% for miniFE, GTC and HPCCG, respectively. Applications such as miniMD, miniFE and GTC even show a high percentage for reuse distances from 0 to 2. Only LULESH shows relatively longer distances, which is due to its higher communication degree. Fig. 3 presents the time distance distribution and show that a large portion of the circuits are reused within tens of microseconds. Applications such as miniMD, GTC and HPCCG, even show a high percentage for time distances less than 16 µs.

These results provide evidence that there is a high potential to reuse a circuit for multiple near requests, thereby amortizing setup delays. While analyzing *a posteriori* communication patterns is helpful, using such information for better *runtime* optimization is yet another thing. In particular, we need to determine which circuit should be maintained to seize the reuse opportunities, given that the size of the circuit set is limited. We hence explore *online* techniques that utilize the observed circuit-use history at runtime to optimize circuit replacements.

IV. Predicting Reuse Distance

One key step of utilizing observed reuse history for circuit replacement is to predict the reuse distance for replacement candidates when a circuit miss occurs. In this way, the circuit that is the least likely used in the near future can be removed. In this section, we describe two techniques for predicting the reuse distance. Our *online* methods presented here differ from previous works that considered offline cases. We also compare the prediction accuracy of the two predictors.

A. Maximum Likelihood Based Predictor (MLBP)

MLBP looks at the currently-collected reuse distance distribution of *a circuit* and selects the bin with the highest frequency as the prediction. Although the prediction has the maximum likelihood, MLBP suffers from two major drawbacks: 1) its prediction accuracy largely depends on distribution pattern: if the distribution has one or more bins with comparable frequency to the highest bin, the prediction accuracy is hindered; 2) MLBP neglects the temporal pattern of the reuse distance sequence collected.

B. Transition Matrix Based Predictor (TMBP)

TMBP avoids the drawbacks of MLBP. It explores the temporal aspect of the reuse distance sequence observed for a circuit and offers prediction based on transition patterns in the sequence. To extract the pattern, TMBP models the transition of reuse distance using a Markov chain (Fig. 4). The states of the Markov chain correspond to the histogram bins, while the transition matrix represents the probability of the reuse distance transiting from one bin to another. Each time a reuse distance sample is collected, the matrix element corresponding to the transition from the last bin to the current bin increments by 1. Upon predicting the next reuse distance, TMBP finds the bin to which the current bin has the greatest transition probability. Such Markov chain is maintained per circuit.

C. Prediction Performance

Fig. 5 shows how our two prediction techniques lead to different prediction accuracies across the applications and problem sizes. Each time a circuit is used, its distance until the next use is predicted. If this prediction has the same *log2* value as the reuse distance observed (at the next use), then the prediction is considered accurate; otherwise, it is considered not accurate. In the case of miniMD, Fig. 5 shows that MLBP sees a severe accuracy drop when the number of nodes increases from 64 to 128 and 256. The reason lies in Fig. 1 and 2, where the distribution of miniMD transforms from a single-tower shape into a two-tower one. In comparison, the accuracy of TMBP remains at high, with a gain of 40% and 36% over MLBP observed in the cases of miniMD and HPCCG, respectively.

V. Managing Circuit Replacement

Prediction of circuit reuse distances allows us to approximate an optimal replacement algorithm because we can

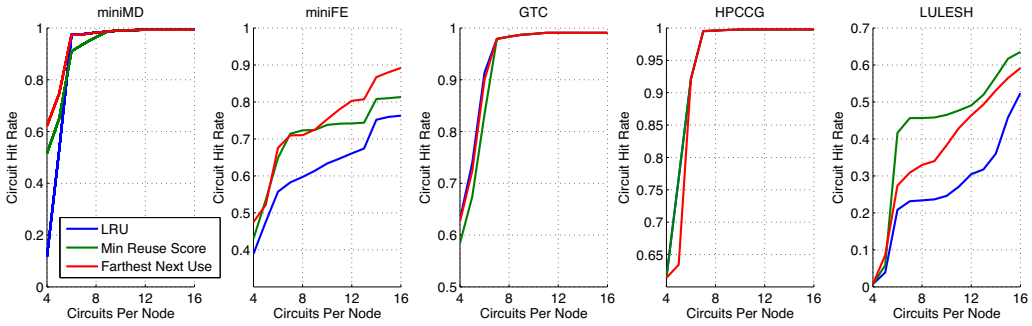

Fig. 6. Circuit hit rates (64 nodes) for replacement policies: LRU, Farthest Next Use and Minimum Reuse Score, across different benchmarks.

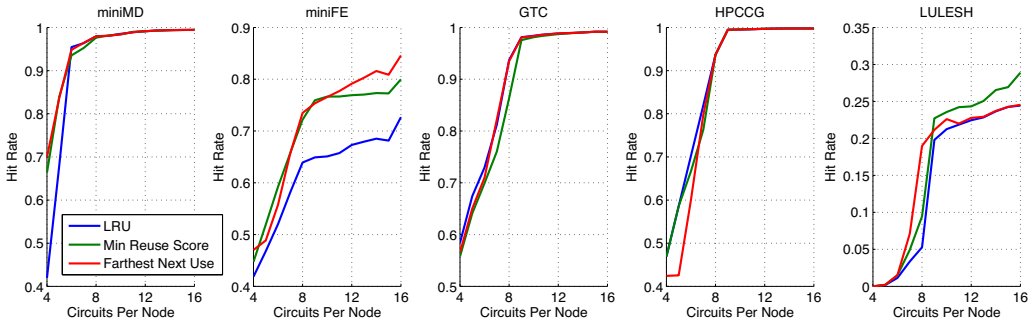

Fig. 7. Circuit hit rate (256 nodes) for replacement policies: LRU, Farthest Next Use and Minimum Reuse Score. 512 nodes for LULESH.

attempt to preempt future communication patterns with appropriate circuit configurations. The circuit that is the least likely used in the near future is "sacrificed" when a circuit miss occurs. In this section, we describe two ways of using the reuse distance information for circuit replacement.

A. Farthest Next Use (FNU)

The FNU policy selects for replacement the circuit that is going to be reused in the farthest future. Each time a circuit miss occurs, circuits that are currently maintained but not in data transmission become replacement candidates. Similar to [22], the estimated time to access (ETA) a circuit can be calculated by adding the predicted reuse distance to the circuit's last use time minus the current time. However, not every circuit has a positive ETA, some may have a negative value due to the passing of its expected access. In this case, the *decay time* is used, i.e. how much time a circuit has not been used. Different from [22], we also use the *decay time* if credibility of the ETA prediction is not high. The circuit with the largest value for ETA or decay time will be replaced.

B. Minimum Reuse Score (MRS)

In MRS, each circuit is associated with a score regarding its frequency of reuse. Instead of granting every reuse with equal weight, reuses within smaller distances retain higher "values". Each time a circuit is used, its score increases by ($2^{max\_bin} - reuse\_distance$). Each time a replacement is needed, the vacant circuit with the lowest score is replaced.

C. Replacement Performance

Performance of the two aforementioned replacement policies is compared with the *Least Recently Used* (LRU) policy via simulation. Our simulation assumes a fully-connected network topology and that the destination node has adequate receivers (slightly greater than its communication degree) to receive incoming circuits. These assumptions make sure that network contention and receiver contention will not affect the state and replacement of the circuit set at source nodes. Global network-based or destination-based replacement can be also investigated with our proposed techniques and will be included in our future work.

As Fig. 6 and 7 show, in most cases FNU (based on the prediction result of TMBP) and MRS lead to much better or comparable hit rate than the LRU policy, and hence the setup penalty due to circuit misses is minimized. It is worth noting that FNU and MRS perform better than the other in different cases. The reason is that the two policies account circuit history differently. MRS collects scores from the beginning of an application; a circuit's score acquired during an early phase could still secure its position in the circuit set in a later phase even if the circuit is not frequently used in the latter. Such effect could keep dead circuits that have long been vacant from exiting the circuit set. In the case of FNU, if a circuit has long passed its expected access time, the increased decay time will flush it out of the circuit set. Hence, the performance of FNU is better than MRS in many cases, except for LULESH. From the distribution, we know that LULESH shows more

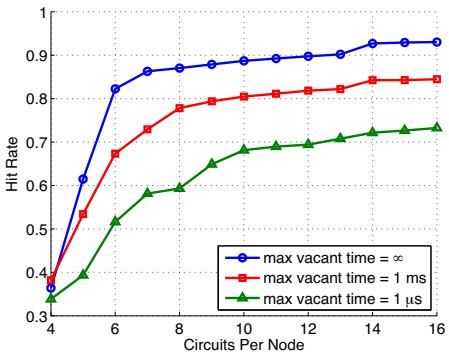

Fig. 8. Circuit hit rate (geometric mean of all benchmarks except LULESH) when maximum vacant time is set to ∞, 1ms and 1 µs.

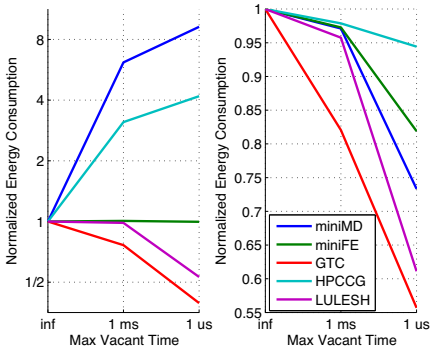

Fig. 9. Energy consumption of circuits versus maximum vacant time. Left: max circuits per node = 6, Right: max circuits per node = 16. All energy values are normalized to the infinite-vacant-time case.

likelihood towards long reuse distances. FNU replaces these long-distance circuits, which however, contribute most reuse opportunities.

VI. Energy Consumption Tradeoff

Although circuit set and cache share many similarities in shaping the hit/miss characteristics of data accesses, several notable differences persist. An obvious difference is that maintaining a circuit explicitly costs time-proportional energy consumption (e.g. laser power), while maintaining a cacheline costs little. Maintaining circuits as long as possible can help further reduce the miss rate; however, such reduction comes at a price of excessive energy consumption and the reduction might not be proportional to the price paid. A long-time-no-use circuit can still remain in the circuit set if no replacement occurs. In this case, a mechanism is needed to actively turn off the circuits without the help of replacement. One such method is to predict the time distance of a circuit—if the circuit is not going to be used again until far in the future, it will be turned off. Note that such proactive turn-off (PTO) will not lead to additional penalty if the circuit is to be replaced by a miss before a reuse. However, PTO could indeed lead to the drop of hit rate if the time-distance prediction is not accurate and a circuit reuse does arrive. Instead, if the time-distance prediction is trustworthy, a circuit could be turned off right after its use if its predicted next distance is larger than an allowed *maximum vacant time* (MVT). Such a method can provide more energy savings than waiting until a circuit's decay time reaches MVT.

The change in hit rate with respect to MVT is presented in Fig. 8, where the MVT is set to infinity, 1 ms and 1 µs, respectively. The energy consumption of the circuit set, however, may not necessarily drop as MVT becomes tighter. If the circuit set size is small (e.g. 6 per node, left of Fig. 9), PTO could lead to counter effects on energy consumption. For example, the energy consumption of miniMD and HPCCG increases as MVT shrinks. The reason is that too eager PTO creates more misses, and more energy is consumed during miss penalty periods. However, if the circuit set size is relatively large (e.g. 16 per node, right of Fig. 9)—in which case the circuit resource might be overprovisioned—PTO adaptively shuts down the excessive resources and the energy consumption is reduced.

VII. Experimental Demonstration

We demonstrate dynamic reconfiguration of optical circuits using state-of-the-art silicon photonic devices interfaced to high-speed FPGAs. We construct a 20 Gbps wavelength division multiplexed (WDM) optical network that can be rapidly reconfigured using a silicon photonic switch driven by the FPGA, and then wavelength filtered using a silicon photonic demultiplexing device. We characterize all latencies involved in network reconfiguration.The motivation is to provide parameters for the circuit reuse design optimization.

A. Experimental Setup

The experimental setup is shown in Fig. 10a. An Altera Stratix V GT Signal Integrity Kit FPGA is used to generate PRBS $2^{31} - 1$ data at 10 Gbps. The data is modulated on two DFB laser outputs (1550 nm and 1552 nm) using commercial $LiNbO_3$ modulators and combined using a 50:50 passive optical splitter. The data is then amplified and launched onto a silicon photonic Mach-Zehnder interferometer (MZI)-based 2x2 switch. A second Altera FPGA drives a 40 mVpp, 12 ns-period square wave having a DC offset of 92.2 mV to change the switch between the cross and bar states. The output of the switch is sent to a microring based demultiplexer (demux) for wavelength filtering. The filtered wavelengths are then amplified for data reception through PIN/TIA optical-to-electrical converters and 12.5 GHz limiting amplifiers interfaced directly to another Stratix V FPGA. The 2x2 MZI switch was fabricated through the OpSIS multi-project-wafer foundry service [23] and features both thermal and fast P-I-N electrical switching functionality. The switch is capable of 15 dB cross-bar port extinction ratio and has a fast switch speed of 2 µs [24]. The demux device was fabricated at the Cornell Nanofabrication facility on a standard silicon-on-insulator (SOI) platform and contains localized heaters for thermal tuning. The device has a measured extinction ratio of 15 dB and a thermal time constant of 4 µs.

B. Experimental Results

Fig. 10(b-c) show the temporal response of the MZI switch simultaneously switching two wavelengths. We show optical eye patterns for λ_1 and λ_2 passing through each output state and after filtering by the demultiplexing filter. This demonstration

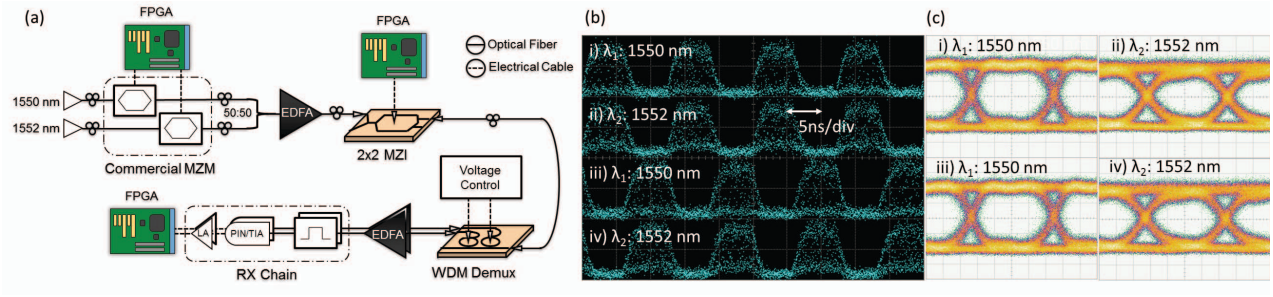

Fig. 10. (a) Experimental setup for dynamic WDM circuit reconfiguration (*Only one switch to demultiplexer path is shown).* (b) Optically-switched WDM data: (i) 1550 nm and (ii) 1552 nm through one path of the 2x2 MZI switch; (iii) 1550 nm and (iv) 1552 nm through the other path of the switch. (c) Optical eye patterns of modulated data.

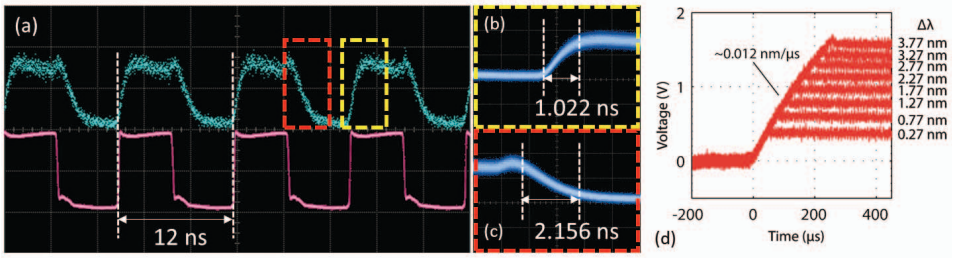

Fig. 11. (a) Optical circuit switching latencies detected using a high speed digital communications analyser. The bottom waveform is the electrical driving signal, and the top waveform is the optical output of the switch. (b-c) Rise and fall times measrued from 10-90%, the fall time is slower because of free-carrier lifetime. (d) demux thermal wavelength locking latencies. Time is on the x-axis and a set of wavelength shifts is located on the right. The ramp waveform shows the time it takes for the output heater voltage to stabilize to the wavelength offsets.

shows a 1.0 ns rise time and 2.2 ns fall time, which is the fastest time achieved yet for OpSIS MZI switch devices. This optical response shown in the figure is the result of the aforementioned 12 ns-period digital square wave used for switching. The electrical rise and fall times are 144 ps and 256 ps, respectively. Fig. 11(a-c) shows the optical switching properties of the OpSIS 2x2 MZI according to the aforementioned electrical control signal.

Fig. 11d shows the thermal tuning and stabilization time for the demux ring filter for a variety of wavelength offsets. The time is on the x-axis and shows stabilization time on the order of 200 µs for a large wavelength offset. This initialization time is added each time the device is not locked to its operation wavelengths and each time wavelength assignments change. For the temporal figures shown previously the thermal tuning is assumed to be complete and unchanged.

Hardware description language (HDL) was used to implement state-based logic, which counted execution times of essential steps in PHY initialization and synchronization logic. PHY initialization requires time to setup electrical transmit and receive component, such as phase-locked-loops and shift registers. We measure an average of 2.635 ms for the PHYs driving each optical datapath. Optical errors are not reflected in the ultimate data delivery due to adaptive equalization of receiver components in the PHY; however, optical errors are reflected in the word alignment process of PHY initialization. We implement a syncword-based word alignment that relies on successful delivery of 5 successive syncwords before the link is available for data transmission. We measure an average synchronization time of 1.2 µs—after initialization—for data delivery over each optical data path. We demonstrate successful delivery of $5x10^{12}$ bits (5 Tb) of PRBS data consecutively on each optical circuit without error. The experimental results show combined latency characteristics of the link initialization process. Faster PHY initialization times are possible using commercial ASICs. Considering the SiP circuit setup latencies, however, we still see the need for circuit reuse design optimization.

VIII. Conclusion and Future Work

In this work, we study architectural solutions for avoiding the setup penalty of silicon photonic circuits. The investigation of circuit reuse distances based on HPC benchmarks provides evidence for the temporal locality of circuit requests and the opportunity to amortize setup overheads. Inspired by previous cache optimization techniques, we investigate the performance of reuse distance based circuit replacement techniques. The proposed *Transtion Matrix Based Predictor* is shown to provide much higher prediction accuracy than previous maximum likelihood prediction for HPC communications. Based on the reuse distance prediction, the two replacement policies--*Farthest Next Distance* and *Minimum Reuse Score*--also effectively increase the circuit hit rate compared to the LRU policy and hence avoid the setup penalty.

Our future work will include a comprehensive evaluation of application performance improvement and energy savings when using the proposed approach. Specifically, we will focus on how the improvement of circuit hit rates could translate into application runtime speedup. We will also study methods for

determining the optimal maximum vacant time (MVT) in order to optimize the performance-energy trade-off.

APPENDIX A

miniMD, miniFE and HPCCG are developed by the Mantevo project [25]. miniMD is a proxy application for molecular dynamics simulations. It has a single kernal with a few AllReduce collectives [26]. HPCCG is a simple conjugate gradient benchmark and "intended to be the best approximation to an unstructured implicit finite element or finite volume application in 800 lines or fewer" [25]. miniFE is also an proxy application for unstructured implicit finite element codes, but with complete computation steps. LULESH (*Livermore Unstructured Lagrangian Explicit Shock Hydrodynamics*) intends to mimic computation of hydrodynamic simulations [27]. LULESH allows perfect weak scaling over distributed architectures [28]. GTC (*Gyrokinetic Toroidal Code*) solves a set of non-linear partial differential equations and is extensively used for fusion energy research. It has a "toroidal" communication pattern and is further described in [21].

ACKNOWLEDGMENT

This work was partly supported by the U.S. Department of Energy (DoE) National Nuclear Security Administration (NNSA) Advanced Simulation and Computing (ASC) program through contract PO 1319001 with Sandia National Laboratories. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

The authors gratefully acknowledge support from Portage Bay Photonics, and from Gernot Pomrenke of the Air Force Office of Scientific Research, and thank Jeremiah Wilke, Cy Chan, Muhammad Madarbux and Philip Watts for discussion.

REFERENCES

[1] A. Shacham, K. Bergman, and L. P. Carloni, "Photonic networks-on-chip for future generations of chip multiprocessors," *Computers, IEEE Transactions on*, vol. 57, no. 9, pp. 1246–1260, 2008.

[2] A. V. Krishnamoorthy, *et al.*, "Computer systems based on silicon photonic interconnects," *Proceedings of the IEEE*, vol. 97, no. 7, pp. 1337–1361, 2009.

[3] A. Bianco, D. Cuda, M. Garrich, G. Castillo, R. Gaudino, and P. Giaccone, "Optical interconnection networks based on microring resonators," *Optical Communications and Networking, IEEE/OSA Journal of*, vol. 4, no. 7, pp. 546–556, July 2012.

[4] M. Glick, "Optical interconnects in next generation data centers: An end to end view," in *Optical Interconnects for Future Data Center Networks*, ser. Optical Networks, C. Kachris, K. Bergman, and I. Tomkos, Eds. Springer New York, 2013, pp. 31–46.

[5] K. Padmaraju, *et al.*, "Wavelength locking of a wdm silicon microring demultiplexer using dithering signals," in *Optical Fiber Communication Conference*. Optical Society of America, 2014, p. Tu2E.4.

[6] K. Padmaraju and K. Bergman, "Resolving the thermal challenges for silicon microring resonator devices," *Lateral*, vol. 60, no. 1554.7, pp. 1554–8, 2013.

[7] T. S. Woodall, G. M. Shipman, G. Bosilca, R. L. Graham, and A. B. Maccabe, "High performance rdma protocols in hpc," in *Recent Advances in Parallel Virtual Machine and Message Passing Interface*. Springer, 2006, pp. 76–85.

[8] M. Nussle, M. Scherer, and U. Bruning, "A resource optimized remote-memory-access architecture for low-latency communication," in *Parallel Processing, 2009. ICPP '09. International Conference on*, Sept 2009, pp. 220–227.

[9] H. W. Jin, S. Narravula, G. Brown, K.Vaidyanathan, P. Balaji, and D. K. Panda, "Performance evaluation of rdma over ip: A case study with the ammasso gigabit ethernet nic," in *Workshop on High Perf. Interconnects for Distributed Computing; In conjunction with HPDC-14*, 2005.

[10] P. Balaji, A. Chan, W. Gropp, R. Thakur, and E. Lusk, "The importance of non-data-communication overheads in mpi," *Int'l Journal of High Performance Computing Applications*, vol. 24, no. 1, pp. 5–15, 2010.

[11] B. W. Barrett, G. M. Shipman, and A. Lumsdaine, "Analysis of implementation options for mpi-2 one-sided," in *Recent Advances in Parallel Virtual Machine and Message Passing Interface*. Springer, 2007, pp. 242–250.

[12] T.Hoefler, J.Dinan, D.Buntinas, P.Balaji, B.W.Barrett, R.Brightwell, W. Gropp, V. Kale, and R. Thakur, "Leveraging mpis one-sided communication interface for shared-memory programming," in *Recent advances in the message passing interface*. Springer, 2012, pp. 132–141.

[13] T. Hoefler, *et al.*, "Remote memory access programming in mpi-3," *Argonne National Laboratory, Tech. Rep*, 2013.

[14] S.W.Poole, O.Hernandez, J.A.Kuehn, G.M.Shipman, A.Curtis, and K. Feind, "Openshmem-toward a unified rma model," in *Encyclopedia of Parallel Computing*. Springer, 2011, pp. 1379–1391.

[15] K. D. Underwood, M. J. Levenhagen, and R. Brightwell, "Evaluating nic hardware requirements to achieve high message rate pgas support on multi-core processors," in *Proc. of the 2007 ACM/IEEE Conf. on Supercomputing*, SC '07. New York, NY: ACM, 2007, pp. 36:1–36:10.

[16] X. Zhu, *et al.,* "Fast Wavelength Locking of a Microring Resonator," *IEEE Optical Interconnects Conference* 2014 MB4 (May 2014).

[17] K. J. Barker, *et al.*, "On the feasibility of optical circuit switching for high performance computing systems," in *Proceedings of the 2005 ACM/IEEE Conference on Supercomputing*. Washington, DC, USA.

[18] A. Shacham, B. Lee, A. Biberman, K. Bergman, and L. Carloni, "Photonic noc for dma communications in chip multiprocessors," in *High-Performance Interconnects, 2007. HOTI 2007. 15th Annual IEEE Symposium on*, Aug 2007, pp. 29–38.

[19] M. Madarbux, A. Van Laer, and P. Watts, "Low latency scheduling algorithm for shared memory communications over optical networks," in *High-Performance Interconnects (HOTI), 2013 IEEE 21st Annual Symposium on*, Aug 2013, pp. 83–86.

[20] G. Hendry, *et al.*, "Circuit-switched memory access in photonic intercon- nection networks for high-performance embedded computing," in *High Performance Computing, Networking, Storage and Analysis (SC), 2010 International Conference for*, Nov 2010, pp. 1–12.

[21] G. Hendry, "Decreasing network power with on-off links informed by scientific applications," in *Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2013 IEEE 27th International*. IEEE, 2013, pp. 868–875.

[22] G.Keramidas, P.Petoumenos, and S.Kaxiras, "Cache replacement based on reuse-distance prediction," in *Computer Design, 2007. ICCD 2007. 25th International Conference on*, Oct 2007, pp. 245–250.

[23] http://www.opsisfoundry.org.

[24] T.Shiraishi, *et al.*, "A reconfigurable and redundant optically-connected memory system using a silicon photonic switch," in *Optical Fiber Communication Conference*. OSA, 2014, pp. Th2A–10.

[25] M. A. Heroux, *et al.*, "Improving performance via mini-applications," *Sandia National Laboratories, Tech. Rep. SAND2009-5574*, 2009.

[26] R. Numrich and M. Heroux, "A performance model with a fixed point for a molecular dynamics kernel," *Computer Science - Research and Development*, vol. 23, no. 3-4, pp. 195–201, 2009.

[27] I. Karlin, *et al.*, "Lulesh programming model and performance ports overview," Technical Report LLNL-TR-608824, Lawrence Livermore National Laboratory, Tech. Rep., 2012.

[28] I.Karlin, *et al.*, "Exploring traditional and emerging parallel programming models using a proxy application," in *Parallel & Distributed Processing (IPDPS), 2013 IEEE 27th Interna- tional Symposium on*. IEEE, 2013, pp. 919–932.