Dynamic Reconfiguration of Silicon Photonic Circuit Switched Interconnection Networks

David Calhoun, Ke Wen, Xiaoliang Zhu, Sébastien Rumley, 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*—Silicon photonic interconnection networks can provide significant advantages in bandwidth densities and data communication energy efficiencies over electronic-interconnected systems. However, due to the circuit switched nature of photonics, the latencies associated with circuit setup and reconfiguration must be managed to avoid added delays in application execution time. This work characterizes circuit setup latencies and explores architectural solutions and methodologies for amortizing the setup latencies. We first collect data on the initialization delay of an end-to-end silicon photonic link using the latest generation of Field Programmable Gate Array (FPGA) hardware to control a Mach-Zehnder interferometer (MZI)-based 2x2 switch and a microring-based demultiplexing filter. We demonstrate nano-second scale broadband switching for wavelength division multiplexing (WDM), and a scalable FPGA-based method for wavelength locking and thermal stabilization of the microring demultiplexer. The burst-mode receiver synchronization time of the FPGA is also measured for a 20 Gbps datapath. Using the collected data we then estimate the overall delay to set up an optical circuit, and investigate the effects of a circuit management technique to amortize this delay. The technique is adapted from cache optimization and relies on a metric called circuit reuse distance (CRD). Reuse distance profiles are generated based on a group of representative HPC applications with different communications patterns, and the potential to amortize circuit setup delay over multiple circuit requests is shown. Simulations show that a Markov transition matrix based reuse distance prediction method has significantly higher accuracy than traditional maximum likelihood prediction. Additionally, we show that two replacement policies using Markov chain prediction effectively increase the hit rate compared to the least recently used policy.**

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

I. Introduction

The communication demands for high performance computing (HPC) systems continue to increase as we advance towards exascale systems. The requisite high-bandwidth connectivity for future HPC systems necessitates low cost, energy efficient, and high bandwidth density networks. Silicon photonic (SiP) interconnects [1-3] have been demonstrated with large bandwidth densities at high energy efficiency, and show the potential to significantly reduce the latency of data transmission at warehouse scale distances. SiP interconnects can provide high bandwidth *end-to-end* connectivity across HPC platforms [4].

To achieve high bandwidth density and energy efficiency requirements, interconnects will likely be composed of wavelength division multiplexed (WDM)-capable components such as broadband MZI switches and microring resonators that facilitate wavelength selective filters for demultiplexing. The microring-based filter is particularly suitable because it has advantages of low-power operation, small footprint, and inherent WDM capability [5]. However, the root of resonator based devices' advantages—high index contrast and compacted light confinement—also impose special operation requirements to deal with thermal fluctuation and fabrication variation. A variety of methods have been demonstrated that use active control systems to drive a local integrated heater to maintain temperature stabilization [7].

Using resonance-based devices in an interconnect system requires the system to perform wavelength locking and thermal stabilization [11], in effect adding a link setup delay before the link can be used. Because current initialization and control systems require optical power to function, each time light is switched away from an established circuit in an interconnect, that circuit will suffer the initialization delay upon reuse. These optical interconnect initialization delays add to the execution time of HPC applications on the microsecond scale due to SiP thermal time constants [11]. Such a latency penalty could be especially detrimental in scenarios when remote direct memory access (RDMA) [8-10] is enabled or when small messages are used; however, the advantages of optics can still be exploited for application speedups with careful architectural design.

This study expands on previous work [18] with extended simulation results and further demonstration of dynamic optical circuit reconfiguration using state-of-the-art SiP devices interfaced to high-speed FPGAs. This work constructs a 20 Gbps wavelength division multiplexed (WDM) optical network that can be rapidly reconfigured using a SiP switch driven by the FPGA, and then wavelength filtered using a SiP demultiplexing device. We expand on the experimental demonstration to include a scalable FPGA-implemented microring resonator initialization and thermal stabilization platform, thus showing a more accurate study of what is likely to be encountered in a commercial SiP interconnect system during circuit reconfiguration. All latencies involved in network reconfiguration are characterized and SiP circuit initialization delays are confirmed experimentally, showing the need for intelligent circuit management techniques.

978-1-4799-6233-4/14/$31.00 ©2014 IEEE

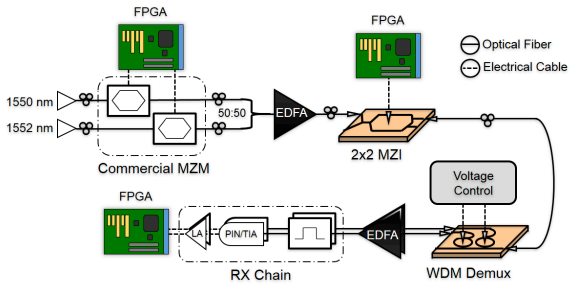
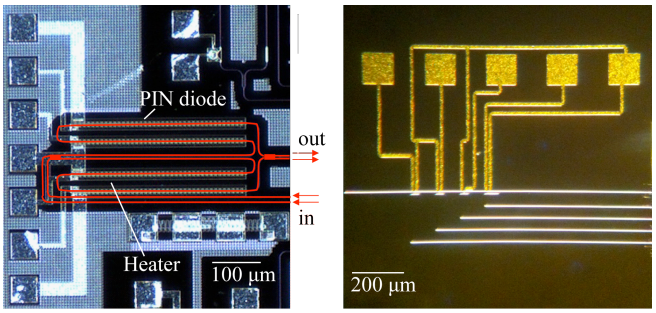


Fig. 1. Experimental setup for dynamic WDM circuit reconfiguration. (*Only one switch to demultiplexer path is shown.)*



(a) (b)

Fig. 2. (a) Image of the MZM swtich, the waveguides are highlighted in red and the aluminum electrical contacts are on the left side. The device contains both thermal and PIN control. (b) Microring based demultiplexer device, with WDM input on the left and 4 outputs plus through port waveguides on the right. Thermal tuning is accomplished using nickel-chromium heaters whose gold contact pads are shown near the top.

Techniques inspired by cache optimizations are applied to intelligently manage optical circuit resources with the goal of maximizing the *circuit hit rate*. This work shows circuit replacement policies based on a metric known as circuit reuse distance (CRD). We show that the CRD metric, as a basis for a *transition matrix based predictor* results in 40% hit rate (accuracy) gain compared to the traditional *maximum likelihood based predictor* for previously untested 128 system nodes. Two reuse distance-based replacement policies are also studied: the *farthest next use* policy and the *minimum reuse score* policy. Simulation based on scientific benchmarks shows that both policies have the potential to achieve much higher hit rates than the *least recently used* policy.

The remaining content is organized as follows: Section II describes the physical layer FPGA-SiP testbed and motivates the burden of circuit initialization delays with experimental latency data. Section III defines circuit reuse distance and profiles its distribution based on a set of scientific HPC benchmarks. This profiled information leads us to the design of online predictors (Section IV), which are key mechanisms in our circuit replacement policies detailed in Section V.

II. EXPERIMENTAL DEMONSTRATION

A. Experimental Setup: Circuit Switching

Phase 1 of the experimental setup is shown in Fig. 1. An Altera Stratix V GT Signal Integrity Kit FPGA is used to generate two streams of 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 a third Stratix V FPGA. The 2x2 MZI switch was fabricated through the OpSIS multi-project-wafer foundry service [13] and features both thermal and fast P-I-N electrical switching functionality. The switch is capable of 20 dB cross-bar port extinction ratio and has a fast switch speed of 2 µs [14]. 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. The switch and demux are shown in Fig. 2 (a) and (b), respectively.

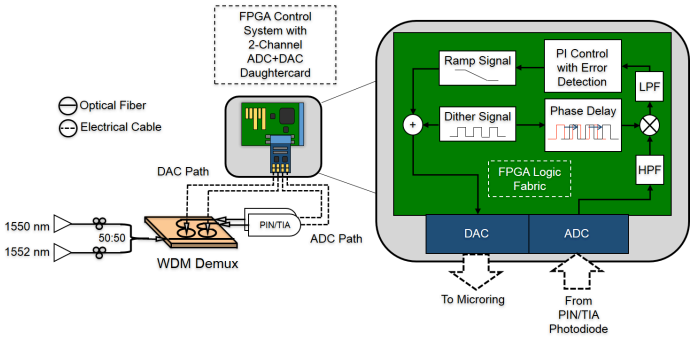


Fig. 3. Experimental setup for FPGA-enabled wavelength locking. The FPGA and ADC+DAC daughtercard components are expanded to show connections and control (for one channel connected to one microring) within the system.

B. Experimental Setup: Wavelength Locking and Stabilization

Phase 2 of the experimental setup is shown in Fig. 3. The microring-based demultiplexer device was used in conjunction with an Altera Stratix V GX Development Kit FPGA and HSMC analog-to-digital (ADC) and digital-to-analog (DAC) daughter card to create an actively controlled, wavelength-stable and -selective control system. The ADC operates with 14 bits of resolution at 65 megasamples per second (MSPS) and the DAC operates with 14 bits of resolution at 125 MSPS. Two DFB laser outputs (1550 nm and 1552 nm) were filtered and simultaneously locked and stabilized using a closed-loop proportional-integral (PI) feedback system. According to the technique successfully demonstrated in [6] and [11], a dithering signal—in the form of a 50 mVpp square wave at 100 kHz—is mixed (additive) with a voltage ramp and applied to each microring to generate an anti-symmetric error signal. The control system was designed to first initiate the locking sequence for the shortest wavelength (1550 nm); upon generating and locking to the first error signal, the second longest wavelength (1552 nm) locking sequence is initiated and its error signal is generated and locked. It should be noted that the FPGA-enabled microring locking system is a turn-key approach to the "voltage control" block depicted in Fig. 1. The

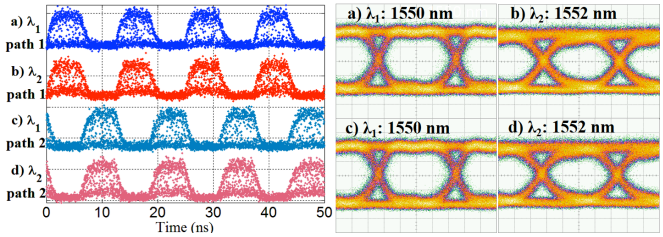


Fig. 4. (*LEFT*) Optically-switched WDM data with time (x-axis) in nanoseconds: (a) 1550 nm and (b) 1552 nm through one path of the 2x2 MZI switch; (c) 1550 nm and (d) 1552 nm through the other path of the switch. (*RIGHT*) Optical eye patterns of modulated data.

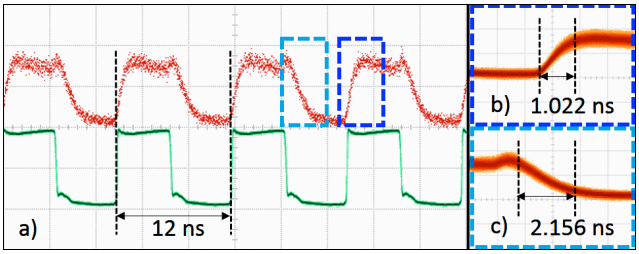


Fig. 5. (a) Optical circuit switching latencies detected using a high speed digital communications analyzer (DCA). The bottom waveform is the electrical driving signal, and the top waveform is the optical output of the switch. (b-c) Rise and fall times measured from 10-90%, the fall time is slower because of free-carrier lifetime.

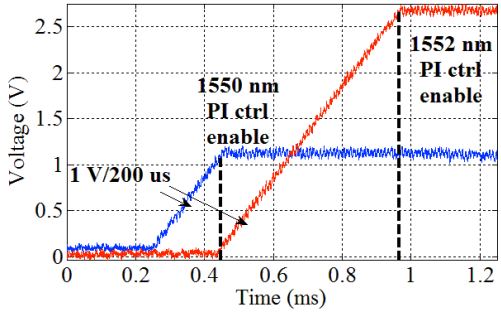


Fig. 6. FPGA-enabled demux wavelength locking control voltages. Time is on the x-axis and heater output voltage is on the y-axis. A locking time of approximately 200 µs is shown in blue for ring 1 locking to 1550 nm with a control voltage of approximately 1.1 V. A locking time of approximately 500 µs is shown in red for ring 2 locking to 1552 nm with a control voltage of approximately 2.7 V.

FPGA locking system is automatic but disjoint for phase 2 of the experiment, whereas wavelength locking is performed manually in phase 1.

C. *Experimental Results: Phase 1*

Fig. 4 shows 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 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. 5 shows the optical switching properties of the OpSIS 2x2 MZI according to the aforementioned electrical control signal. This portion of the experiment uses a hand-tuned and static stabilization method for the demultiplexing filter to ensure that each microring is tuned to the appropriate wavelength.

Another important metric in circuit setup is the burst mode receiver-related physical layer (PHY) initialization. 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 components, 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. A syncword-based word alignment method is implemented 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 circuit-based link initialization process. Faster PHY initialization times are possible using commercial ASICs.

D. *Experimental Results: Phase 2*

Fig. 6 shows the thermal tuning control voltage and stabilization time for the FPGA-enabled demux ring filter-locking system, controlling 1550 nm and 1552 nm. With a ramp speed of 1 V/200 µs, the system locks to 1550 nm in approximately 200 µs and 1552 nm in approximately 500 µs. This initialization time is added each time the device is not locked to its operation wavelengths and each time wavelength assignments change.

The completed control loop was implemented on the FPGA using HDL and consists of a forward path from the DAC and a feedback path from the ADC. These paths are coupled together using state-based logic in a closed loop fashion. The feedback loop is designed to detect the error signal on the ADC path—the threshold for detecting the error signal and the stabilization reference level were determined experimentally and hard-coded in software for each microring separately. As depicted in Fig. 6, the control system responds to the error signal by enabling PI control and immediately stabilizes with less than 150 mVpp variation in the control signal.

E. *Implications for Commercial SiP Links*

SiP device peculiarities impose penalties on optical circuit setup and maintenance, and these peculiarities scale directly with the number of devices included in an interconnect system. A commercial SiP interconnect will include many devices to realize transmission, switching, filtering, and receiving functionalities, and will therefore require many control systems. The control system shown here is easily replicated on the FPGA and can scale to a large number of devices. The control system can also function concurrently with the HDL-coded data transmission hardware and PHY implementations. Integrating HDL state-based logic from each experimental phase of this study can provide complete end-to-end link

functionality, thus enabling emulation and study of real commercial SiP links. PHY and stabilization/locking control logic are easily transferred to an ASIC platform and optimized for size, weight, area, and power (SWAP) – which are important metrics in SiP-enabled HPC systems. The data collected here shows the limitations of physical layer dynamic reconfiguration and motivates architectural methods to optimize temporal circuit reuse, presented in the next section.

III. CIRCUIT REUSE DISTANCE

A. Preface: Circuit Hit and Circuit Miss

Compared with traditional two-sided communications, RDMA mitigates synchronization overheads between sending and receiving processes by allowing one process to directly access the remote memory space of another [12]. If a RDMA request can immediately use a corresponding circuit without waiting for initialization, we call this a *circuit hit*, otherwise the request sees a *circuit miss* and suffers from the initialization penalty. In this sense, the role of circuits in a SiP circuit supported RDMA system is analogous to caches in microarchitectures. A way to increase the hit rate and avoid the miss penalty is to maintain a set of frequently accessed circuits and carefully update them as needed. Considering temporal locality in an application's communication pattern—i.e., a node referencing a remote memory space multiple times within a short period—we can improve the application performance by reducing initialization overheads.

It is impossible to achieve 100% successful circuit hit rate indefinitely because optical connections cannot be maintained for all source-destination pairs and application communication patterns can change over time. The challenge then becomes to carefully select and update the set of maintained circuits.

B. Definition

Reuse distance is a metric for the frequency of successive circuit resource usages. This important metric not only applies to guiding cache optimization [15], but also to circuit-switched networks, especially when each node is allowed to maintain multiple transceivers due to the large-scale integration capability of SiP technologies.

In this work we consider circuit reuses from the source-node perspective. A *reuse distance* of a circuit *C* is the number of circuit requests made by its source node *S* between two consecutive uses of *C*. For example, if the sequence of circuit requests made by *S* is *C, A, B, F, E, C* (labeled by destinations), then the reuse distance of circuit *C* is 4. We also call the outgoing circuits simultaneously maintained by a source node its *circuit set*. The proposed techniques can also apply to other perspectives including destination nodes or the entire network. The reuse distance of circuit *C* can be also measured in time, i.e. the time elapsed between two consecutive uses of *C*.

C. Profiling Based on HPC Benchmarks

We search for circuit reuse opportunities in HPC applications by analyzing the distribution of reuse distance based on a group of representative benchmarks [16, 17]. As a node progresses its workload, it assigns an index number to each of its circuit requests. Upon a new request, the difference between the current index and the last index of the requested circuit is a *sample* of reuse distance for that circuit. To cover a wide distance range, we use power-of-two based bin divisions, i.e. [0], [2^0], [2^1, 2^2), [2^2, 2^3), etc.

The resulting distance histograms are shown in Fig. 7a and Fig. 8. Each application leads to a different reuse pattern. Applications such as miniMD show 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 its irregularity in the temporal pattern. Overall, the results show a high probability for the circuit to be reused within a small distance. For instance, reuse distances in miniMD with a value smaller than 8 comprise 90% of the samples. Applications such as miniMD, miniFE and GTC even show a high percentage for distances ranging from 0 to 2. Fig. 8 presents the time distance distribution and shows that many circuits are reused within tens of microseconds.

IV. PREDICTING REUSE DISTANCE

Upon a request miss, the circuit that is the least likely used in the near future should be replaced. One key step to optimize such replacement at *runtime* is to predict the reuse distance of the replacement candidates.

Traditional methods predict by looking at the currently-collected reuse distance distribution of *a circuit* and selects the bin with the highest frequency. However, such *maximum likelihood based predictor (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.

Here we utilize a *transition matrix based predictor (TMBP)* that avoids the above drawbacks (Fig. 7b). TMBP models the temporal aspect of the reuse distance sequence observed for a circuit using a Markov chain model and offers a prediction based on the sequence's transition pattern [17]. 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 value to the current one increments by 1. Upon predicting the next reuse distance, TMBP finds the bin to which the current bin has the greatest transition probability. Such a Markov chain is maintained per circuit.

Fig. 9 shows how our two prediction techniques lead to different prediction accuracies across the applications and problem sizes. Each time a circuit is used, we predict its distance until the next use. If this prediction has the same *log2* value as the next reuse distance observed, then the prediction is considered accurate; otherwise, it is not accurate. In the case of miniMD, Fig. 9 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. 7a, where the distribution of miniMD transforms from a single-tower shape into a multi-tower one (when observing 64 nodes versus 128 nodes). In comparison, the accuracy of TMBP remains high, with a gain of 40% and 36% over MLBP observed in the cases of miniMD and HPCCG, respectively.

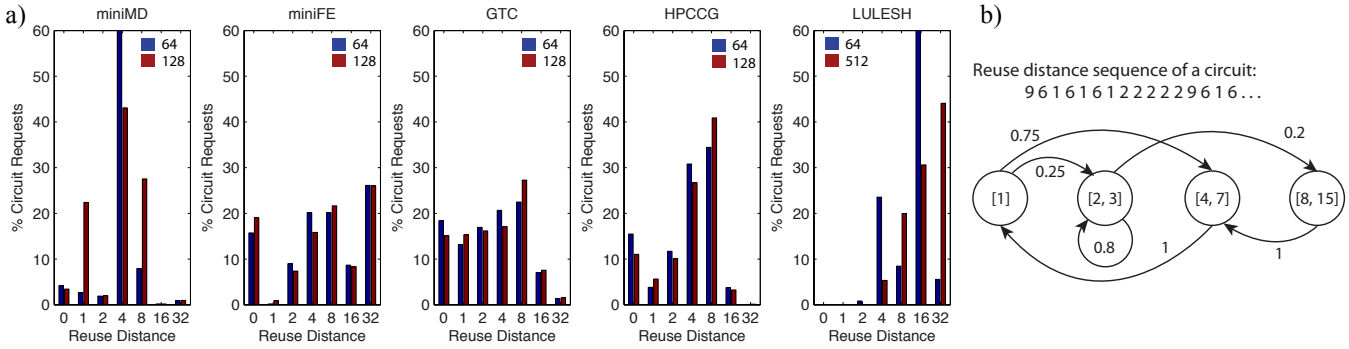


Fig. 7. (a) Distribution of reuse distances for HPC benchmarks (blue for 64 nodes and red for 128 nodes). Each bin corresponds to a range between the current label (included) and the next label (excluded); same for Fig. 8. The results show a majority of circuit reuses within a distance less than 8 (except LULESH). (b) Example for transition matrix Based preditor. (*TOP)* Reuse distance sequence of a circuit. (*BOTTOM*) Modeling of the sequence transition using a Markov chain. Each state of the Markov chain corresponds to a bin in the distribution histogram.

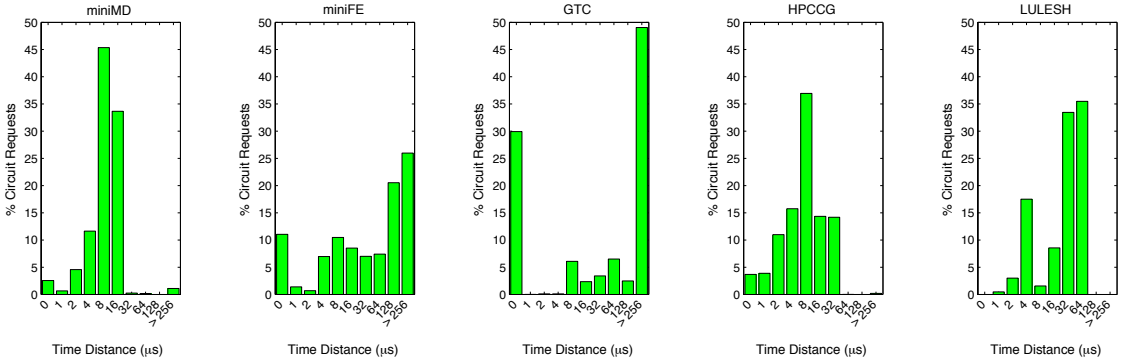


Fig. 8. Distribution of time-based reuse distances for HPC benchmarks (64 nodes). For miniMD, GTC and HPCCG, the majority of reuses are within 16 µs.

V. Managing Circuit Replacement

Predicting circuit reuse distances allows us to approximate an optimal replacement algorithm because we can attempt to preempt future communication patterns with appropriate circuit configurations. In this section, we describe two ways of using the reuse distance information for circuit replacement.

A. Circuit Replacement Policies

The farthest next use (FNU) policy selects the circuit that will be reused in the farthest future. Similar to [15], 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. Not every circuit has a positive ETA, and 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 [15], we also use the *decay time* if the credibility of the ETA prediction is not high enough. In FNU, the circuit with the largest value for ETA or decay time will be replaced.

In the minimum reuse score (MRS) policy, a score is given for each circuit's 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.

B. 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. 10 shows, 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 generally better than MRS, except for LULESH. From the distribution, we know that LULESH shows more likelihood towards long reuse distances. FNU replaces these long-distance circuits, subsequently overlooking most reuse opportunities.

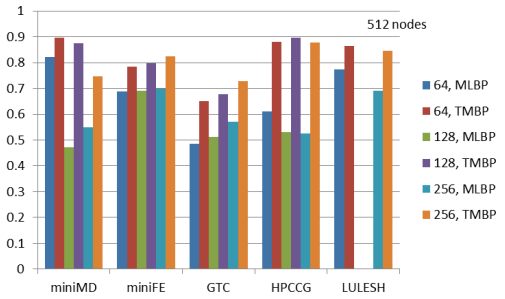


Fig. 9. Reuse distance prediction accuracy of transition matrix based predictor (TMBP) versus maximum likelihood based predictor (MLBP), across different benchmarks and numbers of nodes. TMBP shows as much as 40% and 36% higher accuracy than MLBP for miniMD and HPCCG, respectively.

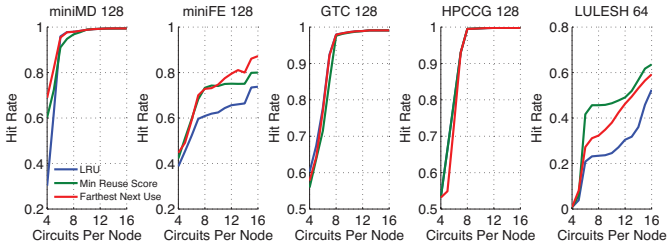


Fig. 10. Circuit hit rates for replacement policies of least recently used (LRU), farthest next use (FNU) and minimum reuse score (MRS), across different benchmarks (128 nodes) and when different numbers of circuits are provisioned per node. FNU and MRS show much better or comparable hit rate than LRU. Maximum hit rate increases of 40%, 16% and 22% are observed in cases of miniMD, miniFE and LULESH, respectively.

VI. Conclusion

In this work, we demonstrate dynamic reconfiguration of 20 Gbps WDM optical circuits using a SiP switch and a SiP demux interfaced to high-speed FPGAs. We measure a PHY synchronization time of 1.2 µs and an FPGA-controlled thermal tuning delay on the order of 200-500 µs for FPGA-enabled SiP link initialization, and subsequently explore architectural solutions for avoiding this setup penalty. The investigation of circuit reuse distances based on HPC benchmarks provides evidence for the temporal locality of circuit reuses and the opportunity to amortize setup overheads. Inspired by previous cache optimization techniques, we investigate the performance of reuse distance based circuit replacement techniques. Our TMBP is shown to provide much higher prediction accuracy than previous MLBP prediction for HPC communications. Based on the reuse distance prediction, the two replacement policies—FNU and MRS—also effectively increase the circuit hit rate compared to the LRU policy and therefore work to mitigate the setup penalty.

Acknowledgments

This work was 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 (AFOSR).

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] N. Ophir, K. Bergman, "Analysis of high-bandwidth low-power microring links for off-chip interconnects [invited talk]," *SPIE Photonics West*, Feb. 2013, pp. 8628-22.

[6] 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.

[7] K. Padmaraju and K. Bergman, "Resolving the thermal challenges for silicon microring resonator devices," Nanophotonics 2 (4), Sep 2013

[8] 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 Paralle Virtural Machine and Message Passing Interface. Springer, 2006, pp. 76-85

[9] 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.

[10] 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.

[11] X. Zhu, K. Padmaraju, L.W. Luo, M. Glick, R. Dutt, M. Lipson, and K. Bergman, "Fast Wavelength Locking of a Microring Resonator," *IEEE Optical Interconnects Conference* 2014 MB4 (May 2014).

[12] R. Gerstenberger, M. Besta, and T. Hoefler, "Enabling highly-scalable remote memory access programming with mpi-3 one sided," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, ser. SC '13. New York, NY, USA: ACM, 2013, pp. 53:1–53:12.

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

[14] T. Shiraishi, Q. Li, Y. Liu, X. Zhu, K.Padmaraju, R. Ding, M. Hochberg, K. Bergman, "A reconfigurable and redundant optically-connected memory system using a silicon photonic switch," in *Optical Fiber Communication Conference*. OSA, 2014, pp. Th2A–10.

[15] 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.

[16] M. A. Heroux, D. W. Doerfler, P. S. Crozier, J. M. Willenbring, H. C. Edwards, A. Williams, M. Rajan, E. R. Keiter, H. K. Thornquist, and R. W. Numrich, "Improving performance via mini-applications," *Sandia National Laboratories, Tech. Rep. SAND2009-5574*, 2009.

[17] 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.

[18] K. Wen, *et. al.*, "Reuse Distance Based Circuit Replacement in Silicon Photonic Interconnection Networks for HPC," *2014 IEEE 22nd Annual Symposium on High-Performance Interconnects*, 2014.