Modeling Performance and Energy Consumption of Silicon Photonic Interconnection Networks via Analytical Cache Models

Ke Wen[1], Jeremiah J. Wilke[2], Sébastien Rumley[1] and Keren Bergman[1]
[1] Department of Electrical Engineering, Columbia University, New York, NY
[2] Sandia National Laboratories, Livermore, CA

***Challenge Addressed*—Silicon photonics (SiP) occupies an entirely different portion of the interconnect design space from electrical switches. Huge bandwidth and energy benefits can be realized, but at extra latency costs associated with establishing point-to-point circuit links. Aiming at addressing this issue, the proposed technologies are sufficiently disruptive to have broad impacts across the entire software/hardware stack of HPC. Optimizing use of silicon photonics may demand new hardware support in the NIC, new runtime systems to manage connections, or even new programming model features to give application developers the ability to optimize circuit reuse. These design challenges must be first tackled with modeling and simulations. This work proposes use of analytical cache models to co-optimize performance and energy consumption of SiP interconnects.**

I. Introduction

Increased parallelism and data intensity have added to communication demands within high performance computing (HPC) systems. Silicon photonic (SiP) interconnects [1-2] have been shown capable of providing large bandwidth end-to-end connections over warehouse distances. These connections, eliminating the need for per-hop electronic switching, significantly reduce communication latencies and energy consumption. Despite these advantages, SiP interconnects also have a set of peculiarities and special requirements. For example, resonance based devices such as microring resonators require wavelength tuning [3] and thermal stabilization [4] to reach desired operating wavelengths. These requirements, inducing potentially long circuit initialization delays, can easily negate the aforementioned latency improvement.

Therefore, a SiP interconnection network—with a circuit-switched nature—must operate in such a way that a communication request can immediately find an established circuit (we call this a *circuit hit*); otherwise, the request sees a *circuit miss* and has to suffer from the setup penalty. In scenarios such as remote memory access (RMA) networks, the role of photonic circuits highly resembles that of cache lines in processor microarchitectures.

Enabled by the integration capability of SiP technologies, a disruptive design choice for maximizing circuit hit rates is to allow each node to maintain (and update) a set of circuits towards a subset of frequently accessed destinations. Thanks to the temporal locality present in the communication pattern of many applications, this design permits to *reuse* an established circuit for a number of temporally near requests, thus effectively amortizing the circuit initialization overheads. The proposed technologies may have broad impacts across the entire software/hardware stack. SiP may demand new runtime systems to manage connections, or even new programming models to give application developers directly the ability to optimize circuit reuse. These design challenges are also accompanied with several performance/cost trade-offs:

(1) How many circuits should be provisioned per node to achieve the best performance-energy balance?

(2) How to design the replacement policy for the case of circuit misses?

(3) Can on/off circuits save energy while retaining the same level of performance (e.g. hit rate)?

(4) Similar to cache prefetching, is it possible to replace or prefetch circuits in a predictive manner?

These questions must be tackled with modeling and simulations. The approach given here fundamentally recasts the problem in terms of cache reuse policies in a runtime system. The approach first models the temporal locality and circuit reuse behavior of an application, and reduces them to a simple model. Then, for efficient circuit replacement, we develop prediction approaches capable of predicting circuit reuses *at runtime*. For scalable evaluation of replacement policies, we propose coarse-grained discrete event simulations as well as *analytical cache modeling*, to which reuse profiles are input and circuit miss rates are predicted. The idea of using on/off links has been explored via the programming model [5]. Here we also show approaches for optimizing the performance-energy consumption tradeoff in the proposed technology.

II. Modeling Circuit Reuse Behaviors of Applications

A. Reuse Distance Profiling

This work captures temporal reuse behavior of an application by profiling its *circuit reuse distance (CRD)*. CRD describes how frequent a circuit is used and can be either count-based or time-based. While the count-based metric can provide guidance for replacement policy designs, the time-based metric indicates whether a circuit should be turned off to save energy or maintained to improve the hit rate. Profiling an application gives us a distribution of its CRDs and hence a better understanding of its temporal locality pattern.

Several reuse distance based models exist in cache performance modeling, which can estimate the missed rate given a cache size. Similarly, such models can be used to predict the circuit miss rate given the number of circuits provisioned per node and a specific circuit reuse profile. The estimated performance can be listed with the energy consumed by the provisioned circuits, and a desired balance can be found.

B. Prediction Model

Predicting reuse distance is also a key step for replacement policy design. A common practice is to select the CRD bucket with the highest frequency in the distribution. While such maximum likelihood based predictor (MLBP) works well for distributions that are concentrated, it performs poorly for those having two or more buckets with comparable frequencies. We propose a transition matrix based predictor (TMBP), which models the temporal aspect of a *CRD sequence* and predicts

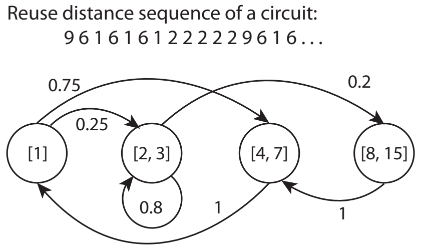

Fig. 1. Modeling transition behavior of a reuse distance sequence (top) using a Markov chain (bottom). Each state of the chain corresponds to a distance range in the distribution histogram.

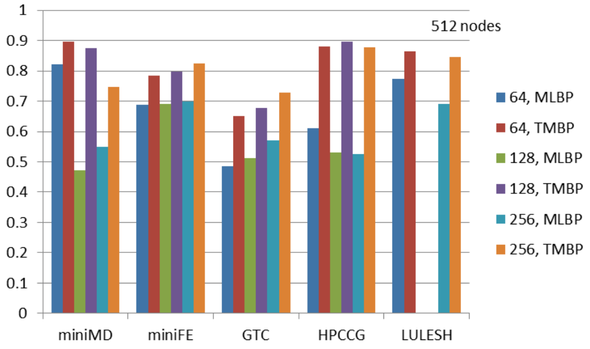

Fig. 2. Reuse distance prediction accuracy of Transition Matrix Based Preditor (TMBP) versus Maximum Likelihood Based Predictor (MLBP), across different benchmarks and numbers of nodes. TMBP (even bars) shows significantly higher prediction accuracy than MLBP (odd bars).

based on its temporal transition probabilities. A Markov chain is used to model the CRD sequence of a circuit (Fig. 1), where the states correspond to the CRD buckets and the transition matrix indicates the probability of the CRD transiting from one bucket to another. Each time a CRD sample is collected, the matrix element representing the transition from the last bucket to the current one increments by 1. Upon predicting the next CRD, TMBP finds the bucket to which the current CRD has the greatest transition probability. Our preliminary result shows that TMBP has much higher prediction accuracy than MLBP and better consistency as the system scales (Fig. 2) [6-7].

III. Scalable Evaluation of Replacement Policies

Course-grained simulations using tools such as the SST macroscale simulator [8] can evaluate performance of circuit replacement policies within real application context, either through on-line simulation of application skeletons or post-mortem replay of MPI traces collected with the DUMPI tool. For modular simulation toolkits like SST, only straightforward changes to existing models are required. It can also leverage existing mini-apps, particularly the Mantevo suite [10].

While such representative benchmark-based evaluations can work as early-stage design hints, simulations can be expensive and evaluation for a broader application spectrum therefore clearly benefits from synthesizable and reliable models. Authors in [9] proposed an analytical model for predicting performance of cache replacement policies (i.e. missed rate) by inputting a simple *circular sequence profile* of the application. The model is based on a deduction that for an *A*-way associative LRU cache, the last access of a circular sequence with *d* distinct accesses will result in a cache miss if $d > A$, or a cache hit otherwise. It also abstracts each replacement policy by a *replacement probability function*, which specifies the probability of replacing a line at different LRU stack positions. It will be interesting to collect similar circular sequence profiles from applications' circuit uses, input them to the above analytical model, and see if modeling results match simulation results. Furthermore, without the need for modifying benchmark codes, synthesizability of this approach also enables efficient investigation of behavior changes that can benefit the application more from circuit reuses.

IV. Energy Consumption Tradeoff

Although roles of circuits and caches share many similarities, notable differences persist. An obvious one is that maintaining a circuit explicitly costs time-proportional energy consumption (e.g. laser power). Though maintaining circuits as long as possible helps reduce miss rates, such reduction comes at a price of excessive energy consumption. Therefore, unused circuits must be proactively turned off (PTO). One method is to predict the time-based CRD of a circuit—if its energy cost for maintenance is greater than the potential benefit, the circuit will be turned off. However, 100% prediction accuracy is not possible, causing a performance-energy tradeoff that can be explored with modeling and simulation. Previous analytical cache models can be modified to include the PTO feature. A time-based CRD can be attached to each circular sequence as its time span. If the CRD exceeds a threshold, the circular sequence will directly contribute to a miss.

It should be also noted that proposed approaches offers vast opportunities for application-circuit runtime co-design. The presence of temporal locality in applications can be potentially reinforced with awareness of underlying circuit reuse opportunities. Also, privileges can be given to programmers to guide PTOs for energy-efficient computing. One can exploit the simulation and modeling capabilities presented above to quickly iterate the co-design process.

V. Assessments

Maturity: The approach builds on previous understanding of memory caching models. Initial works [6-7] have already demonstrated how simple circuit reuse policies can greatly improve interconnect performance.

Uniqueness and Novelty: The approach given here fundamentally and uniquely recasts the circuit management problem in terms of cache reuse policy designs. The prediction model and the idea of synthesizing circuit reuse behaviors for broad-spectrum evaluation are both novel. The idea of using on/off links has been previously explored, although most often in the context of electrical switches than optical interconnects.

Applicability: Although intended for understanding silicon photonics, the cache models here are generally applicable to latency-hiding problems. The notion of circuits, although specific to optical interconnects here, serves as a general abstraction for point-to-point communication.

Effort: The approach integrates with existing simulation capabilities. It requires new NIC models implementing cache reuse policies, but this is straightforward in modular toolkits like SST. The research program can also leverage existing mini-apps. More demanding will be achieving simulations at extreme-scales relevant to exascale research. Newly implemented parallel discrete event simulation (PDES) capabilities have already managed to simulate over 1 million threads on 200K nodes for coarse-grained models. The simulations, while demanding, are at least now feasible.

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] 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.

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

[5] 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.

[6] D. M. Calhoun, K. Wen, X. Lee, *et al.*, "Reuse Distance Based Optimization of Silicon Photonic Circuit Switched Interconnection Networks," *accpeted, 18th IEEE High Performance Extreme Computing Conference (HPEC), 2014*.

[7] K. Wen, D. M. Calhoun, S. Rumley, *et al.*, "Reuse Distance Based Circuit Replacement in Silicon Photonic Interconnection Networks for HPC," *submitted to 22nd Annual Symposium on High-Performance Interconnects (HOTI), 2014*.

[8] Janssen C L, Adalsteinsson H, Cranford S, *et al*. "A simulator for large-scale parallel computer architectures," *International Journal of Distributed Systems and Technologies (IJDST),* 2010, 1(2): 57-73.

[9] F. Guo, Y. Solihin, "An analytical model for cache replacement policy performance," in *SIGMetrics/Performance'06*, 2006.

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