Modeling and Simulation Environment for Photonic Interconnection Networks in High Performance Computing

Madeleine Glick[1], Sébastien Rumley[2], Robert Hendry[2], Keren Bergman[2] and Raj Dutt[1]
1: APIC Corporation, 5800 Uplander Way, Culver City, CA 90230, USA
2: Department of Electrical Engineering, Columbia University, 500 W. 120th Street, New York, NY 10027, USA
glick@apichip.com

ABSTRACT
Computer network architects are concerned with reducing the execution time of applications while maintaining or reducing power consumption of the system. In order to demonstrate the advantages of photonic interconnection networks at the scale of high performance computer clusters and warehouse scale data centers, system level simulations are required. To achieve an appropriate level of accuracy both for device detail and system scale, a simulation suite, rather than a single simulation tool will be most efficient. We present our work on this topic and results for rack scale photonic interconnection networks for high performance computing.
Keywords: optical networks, optical interconnects, simulation.

1. INTRODUCTION AND MOTIVATION

Solutions are required to the newsworthy power consumption [3], latency [4] and bandwidth challenges [5] of high performance computing (HPC). Low latency, previously primarily viewed as a requirement for supercomputers, is becoming a more significant metric as the data center size increases to warehouse scale and web search results gather information from more servers [6, 7]: beyond a given scale, the parallel efficiency (and thus the general efficiency) is directly related to the latency.

Silicon photonics (SiP) [1, 2] has the potential to address these challenges and offer improved network functionality. SiP based device capabilities and manufacturability are advancing rapidly in several key areas including: ring resonator temperature stabilization [8], integrated switch fabrics [9, 10] and hybrid silicon technology [11]. Integrated silicon (and InP based) photonics reduce power consumption and cost in addition to providing high speed switching capabilities. With off chip bandwidths reaching 100's GB/s [12] high bandwidth photonics is clearly an option to reduce I/O bandwidth bottlenecks. However, proof of concept demonstrations at the device or link level are not sufficient to convince computer architects to incorporate photonics into their next generation system designs. On the other hand, system experiments from computer architects tend to focus on commercially available products, whose price and specifications can easily be obtained [1, 2]. Enabling interaction between SiP device designers and computer architects is necessary for realizing novel systems with integrated silicon photonics. Consequently, we must take the extra steps to demonstrate, through modeling and simulations (and, where possible, test-bed validations), the effects of incorporating photonic technology in computer networks. This means, in addition to showing the impact on network metrics such as latency and bandwidth, developing tools that illustrate the execution time improvement of target applications.

The optimal way to incorporate SiP into a large scale system depends on the communication requirements of the executed applications. We therefore need modeling and simulation tools to understand the interactions between real time applications and the hardware [24]. More specifically, we need to compare the various possible associations of a data-plane (topology, mix of electrical or optical components) with one or more resource allocation schemes (centralized compared to distributed, synchronous compared to. asynchronous, time multiplexing compared to statistical multiplexing, etc.). But at the same time, optical network models must not be overly simplistic as there is a strong interdependence between network-level design decisions, such as choice of topology, and device-level design decisions, such as the coupling coefficients. The choice of the number of WDM wavelengths also has consequences on network design and device design. This presents us with the major challenge of developing modeling and simulation capabilities potentially spanning from the impact of physical designs on signal quality to the performance of given application traffic on a silicon photonics based interconnect.

The best solution to this challenge is not immediately obvious. Software development efforts in the community are varied, with different tools focusing on different aspects. In this article, we first review these efforts. We briefly present PhoenixSim 1.0, which was one of the first tools to deal with the challenge of integrating parameters inherent to the physical layer with system-level traffic simulation. Based on these elements, and our conclusion that it is difficult to efficiently integrate all details and metrics into the same tool, we present the first blocks of the evolution of PhoenixSim. We have found that a preferable approach is to build a suite of tools, each tool focusing on a given time scale, level of detail, and/or metric of interest [13].

2. SYSTEM SIMULATION AND MODELING TOOLS

2.1 Related work

Researchers in the field have studied the integration of silicon photonics into computing systems with various strategies. DSENT [20] takes a CMOS-integration approach and focuses on the circuitry required to drive the optical lanes. This provides accurate estimations of the system power and area consumption, however, only under fixed network load. Other groups focus on integrating optical network models with the generic hardware system simulation, using SystemC [21], which allows functional verification as well as extended timing measurements. Models of optical networks have also been incorporated in the gem5 [22] and SESC [23] cycle-accurate simulators, allowing for application performance measurements. Although these efforts result in major advances in optical network understanding, they have their own fields and metrics of interest, and neglect other effects. For example, DSENT doesn't consider traffic variations, while gem5 and SESC do not consider the optical signal quality.

2.2 The PhoenixSim 1.0 experience

The PhoenixSim 1.0 simulator was a first step toward increased integration between these different worlds. It is an event-driven simulator that captures the interaction of every message with every network component. These interactions can be *message-to-network*, as when a message is inserted in a buffer, or *network-to-message*, as when a device partially depletes the message energy. In this way, the impact of each device is accounted for in order to carefully reflect the optical signal quality. Resulting performance metrics of interconnection networks can be analyzed both at the physical level (physical impairments) and the system level (traffic dynamics). A change in the network configuration (typically the number of network clients) can be measured both in terms of bandwidth performance and optical signal quality. PhoenixSim 1.0 has been used for various studies. In particular, Hendry et al. [15] used it to compare electronic and photonic solutions for Fast Fourier Transform, matrix multiplication and projective transform applications for high performance embedded computing. These simulations produced metrics for power consumption, performance (operations per second), and efficiency.

The PhoenixSim 1.0 approach of jointly analyzing traffic performance and signal quality is similar to the one used by Y. H. Chen et al. [25], although their focus is on the interaction between the messages, the power consumption and the thermal dissipation. While providing very valuable information, the joint simulation approach is computation intensive. In [25], the authors report a simulation time (wall-clock time) of 600 s to simulate a 64 node architecture for 30 ms. This approach becomes unwieldy for simulating large data-center sized networks, preferably for a duration of at least 10 seconds.

2.3 Toward the PhoenixSim suite

With this in mind, we decoupled the physical layer study from the statistical response study. We use a physical layer model (PhotonIc Link Optimisation Tool - PILOT) that maximizes the number of wavelengths used on a link while ensuring transmission quality. Insertion loss, crosstalk, and power penalty are calculated for each device along the photonic link. In addition to including detailed device parameters, such as coupling coefficients, PILOT calculates each type of loss as a function of the wavelength channel spacing, yielding an accurate prediction of the maximum feasible throughput of a link. PILOT integrates device models reported in the literature and laboratory measurements, as well as projections based on these models. It can also be configured to follow conservative or optimistic trends. In the future, we will work toward giving PILOT the capability to support several symbol rate and/or modulation formats.

In parallel, we focus on the traffic dynamics of the network ranging from board to rack scale systems [17, 18] with LWSim (Lightweight Sim), a discrete event simulator which deals with traffic dynamics only (queues, drops, protocols). LWSim helps us to develop and understand how the network topology and the resource reservation/allocation mechanisms impair the baseline bandwidth and latency. LWSim integrates timing measurements obtained in the laboratory or reported in the literature. For instance it includes the time required to switch the state of a comb switch ring. We plan to obtain finer timing measurements of buffers and serializers/deserializers through ad-hoc experiments (e.g. VHDL or SystemC simulations) that can then be introduced as parameters. LWSim can be driven with random traffic and can also run application skeletons, which inject traffic depending on the speed at which they receive previous messages.

We are working toward integrating LWSim and PILOT in an automated design process. For this, the Javanco framework [19] serves as a common substrate. A baseline network topology is first created within Javanco. PILOT tells us the highest bandwidth supported by the topology without compromising signal quality, LWSim then tells us how the topology deals with random or sporadic traffic. Based on these results, the baseline topology can be modified toward achieving better application performance. Javanco utilizes graph theory functionalities to reason about the topologies, without involving discrete-event simulation. With these three elements (PILOT, LWSim, Javanco), we have a coherent set of software tools that we call the PhoenixSim suite (Figure 1). With it, starting from an initial abstract design, we can first describe the architecture using the Javanco data structure, then analyze this architecture from the signal and traffic point of view.

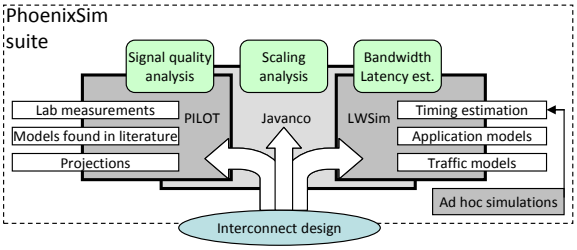

Figure 1. Schematic of the PhoenixSim suite.

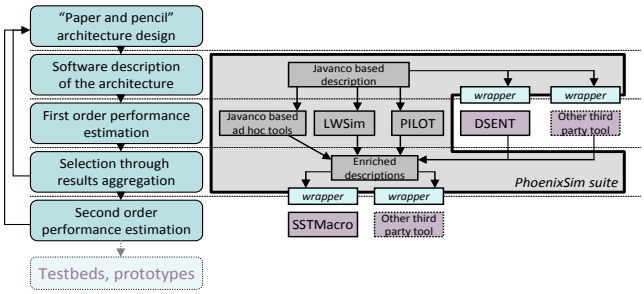

Figure 2. Schematic of the position of the PhoenixSim suite in the research and development process

Our current suite is designed to allow us to quickly design a network architecture incorporating novel photonic devices and test the performance of this architecture for executing an application. It also allows us to analyse potential scalability of this architecture (Figure 1). It thus provides "first order" performance estimations. For the sake of flexibility, LWSim is not a parallel simulator, and therefore has scaling limits. For full scaling, we are working towards integrating our efforts with those achieved by the Sandia National Laboratories with SSTMacro [24]. We are also working towards closer integration with third party tools. In particular, we aim to use DSENT [20] to obtain power estimations. Figure 2 summarizes our target research process.

3. APPLICATION OF THE PhoenixSim SUITE TO INTER-BOARD INTERCONNECTS

To demonstrate the value of including photonic interconnection networks in HPC clusters we are extending our studies to the board, rack and cluster level. In [18] we simulated a rack scale interconnection network with a silicon ring resonator based SPINet (Scalable Photonic Interconnection Network) switch fabric and explored various protocols. We have also used PhoenixSim 1.0 and LWSim to investigate rack to rack networks based on a star coupler topology using tunable transceivers [17].

Based on the results obtained in [18], we are currently investigating architectures with multiple interconnected SPINets to obtain a HPC-scale sized interconnect. PILOT is used to evaluate the maximum scalability of the SPINet interconnect for a given line rate and number of wavelengths. We also use it to analyze the number of switches that can be cascaded considering noise and amplification. Starting from the same architecture, generated with Javanco, we analyze the impact of design decisions including interconnection topology, and the oversubscription factor. In Figure 3, an aggregation of four 32x32 switches is simulated utilizing either 10 or 12 ports devoted to inter-switch traffic, and leaving 22 or 20 ports per switch available for connecting clients (88 and 80 in total). This corresponds to oversubscription factors of 180% and 240% respectively. Each client sends Poisson traffic at 10% of the link rate. We see that for a slight latency penalty, the scalability of the architecture can be increased. Figure 4 depicts the simulated architecture (16x16 SPINet switches are displayed for the sake of readability) and highlights how the distinct elements of our PhoenixSim suite are exploited.

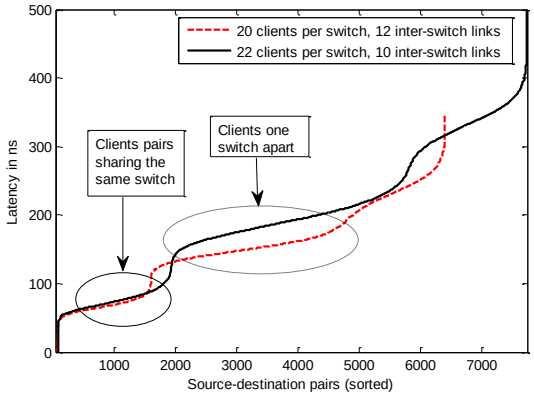

Figure 3. Average latency observed between each source-destination pair in a 4-SPINet architecture, for different number of inter-switch links (180% or 240% oversubscription). Source-destination pairs are sorted according to their corresponding latencies.

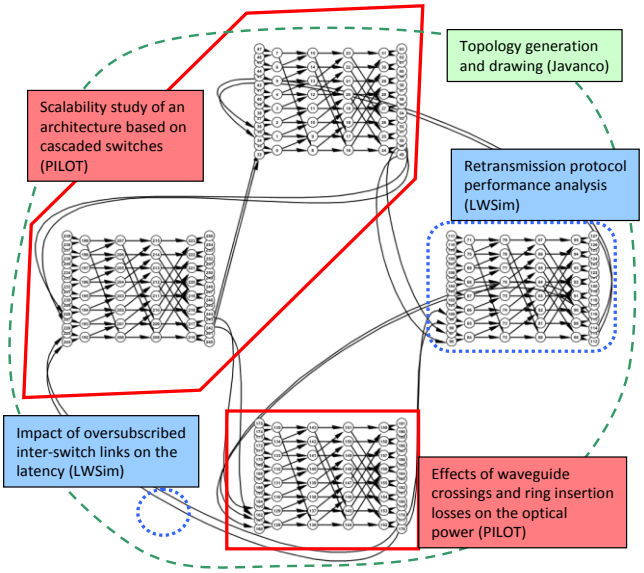

Figure 4. Schematic showing the use of the software tools in the study of a 4-SPINet architecture

4. CONCLUSIONS

In order to determine the system and network level effects of incorporating photonics into large scale computer clusters system level simulations are required. Rather than integrating all the possible models into one simulator, we find that a preferable method consists of using a suite of simulators, each one focusing on a given time scale, level of detail, or metric of interest. With Javanco, LWSim and PILOT, we have the first elements of such a suite. We present these tools and show how they contribute to simulate and optimize network performance of photonic interconnection networks for high performance computing.

REFERENCES

[1] N. Farrington, *et al.*, "Helios: a hybrid electrical/optical switch architecture for modular data centers," *ACM SIGCOMM Computer Communication Review*, pp. 339-350, 2010.

[2] G. Wang, *et al.*, "Your Data Center Is a Router: The Case for Reconfigurable Optical Circuit Switched Paths," *Hot Topics in Networks* (HotNets-VIII) October 22-23, 2009.

[3] J. G. Koomey, "Estimating total power consumption by servers in the US and the world," 2007.

[4] R. Kohavi and R. Longbotham, "Online Experiments: Lessons Learned," *Computer*, pp. 85-87, September 2007.

[5] S. Kandula, *et al.*, "The nature of data center traffic: measurements & analysis, " *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. ACM, 2009.

[6] D Zats, *et al.*, "DeTail: Reducing the Flow Completion Time Tail in Data Center Networks," *Sigcomm'12*, 2012.

[7] M Alizadeh, *et al.*, "Data Center TCP (DCTCP)," *Sigcomm'10*, 2010.

[8] K. Padmaraju, *et al.*, "Integrated Thermal Stabilization of a Microring Modulator," *Optical Fiber Conference* (OFC) 2013.

[9] B. Lee, *et al.*, "Four and Eight-Port Photonic Switches Monolithically Integrated with Digital CMOS Logic Drivers and Circuits," *Optical Fiber Conference* (OFC) 2013.

[10] P. DasMahapatra, A. Rohit, R. Stabile, K. A. Williams, "Broadband 4x4 Switch Matrix using Fifth-order Resonators," Optical Fiber Conference (OFC) 2013

[11] M. J. R. Heck, *et al.*, "Hybrid Silicon Photonic Integrated Circuit Technology," to appear in *IEEE Jour. Sel Topics Quantum Electronics*.

[12] J. R. Bautista, "Interconnect Challenges in a Many Core Compute Environment," *Proceedings of the 17th IEEE Symposium on High Performance Interconnects*-Volume 00, 2009.

[13] J. Shalf, D. Quinlan, and C. Janssen, "Rethinking hardware-software codesign for exascale system," *Computer*, pp, 22-30, 2011.

[14] J.Chan, G. Hendry, K. Bergman, L. P. Carloni, "Physical-Layer Modeling and System-Level Design of Chip-Scale Photonic Interconnection Networks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 30, pp.1507-1520, 2011.

[15] G. Hendry, *et al.*, "Circuit-Switched Memory Access in Photonic Interconnection Networks for High-Performance Embedded Computing," *Supercomputing* (SC), 2010.

[16] S. Kandula, S. Sungupta, A. Greenberg, P. Patel, R. Chaiken, "The Nature of Datacenter Traffic: Measurements & Analysis," *ICM*, 2009.

[17] Q. Li, R. Hendry, J. Chan, K. Bergman, M. Glick, R. Dutt, "Network Simulation of Passive Optical Broadcast-and-Select Network for Avionics Applications," *Government Microcircuit Applications and Critical Technology Conference* (38th Annual GOMACTech Conference), 2013.

[18] G. Dongaonkar, S. Rumley, Q. Li, K. Bergman, M. Glick, "Ultra-low Latency Optical Switching for Short Message Sizes in Cluster Scale Systems," *IEEE Optical Interconnects*, May, 2013.

[19] S. Rumley, *et al.* "Javanco, a software framework for optical network modelling and optimization," accepted for publication, IEEE ICTON 2013.

[20] C. Sun, *et al.*, "DSENT -- A Tool Connecting Emerging Photonics with Electronics for Opto-Electronic Networks-on-Chip Modeling," *6th Symposium on Networks-on-Chip* (NOCS), 2012.

[21] M. Briere, *et al.*, "System level assessment of an optical NoC in an MPSoC platform," *Design, Automation & Test in Europe Conference & Exhibition*, 2007.

[22] A. Van Laer, *et al.*, "Full System Simulation of Optically Interconnected Chip Multiprocessors Using gem5," *Optical Fiber Conference* (OFC) 2013.

[23] N. Kirman, *et al.*, "Leveraging optical technology in future bus-based chip multiprocessors," *39th Annual IEEE/ACM International Symposium on Microarchitecture*, IEEE Computer Society, 2006.

[24] C. L. Janssen, *et al.*, "A simulator for large-scale parallel computer architectures." *International Journal of Distributed Systems and Technologies* (IJDST) 1.2, pp.57-73, 2010.

[25] Y.H. Chen, C. Sun, V. Stojanovic, "Scalable Electrical-Optical Thermal Simulator for Multicores with Optical Interconnects," *IEEE Optical Interconnects*, May, 2013.