Modeling Silicon Photonics in Distributed Computing Systems: from the Device to the Rack

(Invited Paper)

Sébastien Rumley, Dessislava Nikolova, Robert Hendry, Ke Wen and Keren Bergman
Lightwave Research Laboratory, Columbia University, 530 West 120th Street, New York, NY 10027
sr3061@columbia.edu

Abstract: SiP interconnects are envisioned for large scale distributed computing platforms. In such applications, optical systems involving millions of devices have to be modeled. We review key model transformation methods that allow scalable modeling.

OCIS codes: (200.4650) Optical interconnects; (130.6750) Systems

1. Introduction

Optical systems have dominated the long haul telecom market for a few decades now. Medium to long distance optical transmission systems are complex, and can be expensive to produce. Nevertheless, their costs are offset by the incapacity of non-optical systems to provide similar bandwidth-distance products, or to the least to provide them at a low operational cost. Indeed, optical systems show incomparable energy dissipation over long distances. However, as photonics expands to the datacom market and tries to get into the inter-chip or even on-chip communication segments, competition against electrical system becomes fierce. The distances at stake are much smaller (so non-optical systems are less impeded), and in these markets, times for return on investment are much shorter, so procurement costs (i.e. fabrication costs) do matter. Fortunately, with the advent of optical systems integration platforms, one can envision scaled productions of integrated optical systems. Furthermore, the Silicon Photonics (SiP) platform opens the door to mass production using CMOS foundries and know-how [1]. This leads to a drastic decrease of the transmission system's cost, and opens a new horizon for photonics technologies. Indeed, data exchange cost is becoming a dominant one in large computer systems, and scaling in computing power (mainly through increased parallelism) might be ineffective if efficiency in data-movement is not sufficiently improved. Using photonics in short distance communications could offer the necessary improvement in communication speed, energy consumption and procurement costs. Photonics might well be the only option to interconnect 100,000 compute nodes with Terabit/s links, and reach the Exaflop/s goal [2].

There has been significant effort in the last decade from research and industry to develop SiP devices. Generation after generation, devices become more agile as their behavior is better understood and the fabrication process better mastered, and this will likely continue in the next few years. Devices such as microrings, Mach-Zehnder interferometers, or germanium photodetectors can hence still acquire maturity. One of the biggest open challenges is improving the wall plug power efficiency of the external laser sources and amplifiers, possibly through hybrid integration of SiP and gain-capable III-V components.

Developing these nanoscale devices presents highly challenging engineering and technological problems *per se*. In the context of large scale distributed systems, however, one also has to deal with optical systems whose complexity is orders of magnitude higher. A Dense Wavelength Division Multiplexing (DWDM) link is already a complex system, which is subject to interaction between wavelengths (crosstalk, self-phase modulation, etc.) and typically involves up to a thousand optical devices. If optical switching is additionally introduced to limit the number of costly Electro-Optic or Opto-Electric conversions, the optical system is turned into an interconnection system whose device count can be millions or more. In this context, it is crucial to know which device parameters should be emphasized at the interconnect design phase. Adding to the challenge, there are strong interrelations between the interconnection system, the computing elements and the executed application. It is therefore desirable to co-design these three elements, and, consequently, to know which silicon photonics parameters should be brought in the co-design exploration space.

In this paper, we relate our experience in developing an integrated and scalable modeling approach.

2. Modeling approaches

2.1 From the geometry/materials to the impact on light propagation

Due to fabrication delays and cost, it is not realistic to analyze the impact of design parameters on SiP devices behavior by successive fabrication and characterization of prototypes only. Furthermore, characterization itself takes time and does not unveil all of the intrinsic behaviors of the devices under test, since all properties cannot be directly measured. Therefore, to develop understanding of the device "mechanisms" and the sensitivity of these mechanisms to parameters, detailed electromagnetic field or multiphysics simulations are needed. *Numerical techniques* for this include the Finite Element Method (FEM) and the Finite-difference time domain (FDTD) method. These methods divide the space into small elements and solve the field equations for each of them. They are capable of capturing the fine details of, for instance, the electrodynamics, but do not scale well for big devices with many small elements.

Fortunately, we do not always need all the physical quantities at all time. By using some approximations and parameterization of the more complex dependences, one can derive a simpler set of equations which describe well the device behavior in a signal transmission context. The ring coupling model proposed by Yariv [4] is a typical example. All geometrical and material characteristics of a ring and its coupling with one or more waveguides are described with only a few parameters (coupling coefficients κ and t, signal attenuation α, effective refraction index n_{eff} and ring length L). This simplifies the analysis and drastically reduces the design space. As a result, the exploration of this space becomes scalable. *Numerical techniques* are subsequently used to find device designs matching the desired parameters (typically by changing the materials and/or the device geometry).

Other higher abstraction models can be exploited to analyze how the parameters (as κ, t, α, n_{eff}) can be tuned, to obtain active devices such as configurable filters, modulators, switches. FEM or FDTD methods can similarly be leveraged to analyze which physical designs best accomplish this tuning, and how fast it can occur (transition times).

2.2 From light propagation to signal transmission and link bandwidths

Once the impact on light propagation (mainly on the electric field intensity) is known, we can lean toward modeling devices as part of a communication system. In this context, one becomes interested by the way devices affect the carried information symbols. A signal processing approach can be adopted to transform optically passive devices in linear time invariant filter equivalents [5]. These filters, parameterized by the devices parameters (as, again, κ, t, α, n_{eff} in the ring resonator case), can then be sequentially applied to a frequency representation of the source signal. One then obtains the signal attenuations, distortions, and crosstalk values along the system. Actively controlled devices can be represented by different filters, each representing a possible device stable state. Transitions between states are not considered: one always assumes that devices are in the state represented by the filter at the characteristic instant. Transition times, however, can be estimated through FEM; for modulators, they bound the maximum symbol transmission rate (in Baud).

Thereafter, optical power budgets can be established from which the minimal optical power per wavelength required to ensure transmission can be deduced [6]. Ultimately, this allows fixing a bound on the number of wavelengths available for transmission, as the total power propagating along a silicon waveguide will typically be bounded to prevent non-linear effects [6][7]. Assuming a simple on-off keying modulation scheme with 1bit/symbol, these modeling steps lead us to an estimation of the bandwidth supported by a communication system composed of SiP devices represented as filters. Moreover, this modeling approach lets us optimize SiP device parameters for maximizing bandwidth (and/or energy efficiency). With these optimal parameters *for a given architecture* at hand, we can go back to finite element methods and investigate how components offering these values can be designed.

2.3 From the link bandwidths to message latencies

Knowing the bandwidth of a link (maximal one, or the one maximizing another parameter) is far from being sufficient to characterize the system performance in terms of “data-movement”. Many other aspects affect the overall performance. However, taking a different perspective allows us, again, to drastically reduce the design space. Indeed, seen from the CPU (or programmer) point-of-view, an interconnect is ultimately nothing else than a component introducing latency in data-displacement operations. We are therefore mostly interested by the way interconnect designs affect these latencies, in order to optimize for shorter ones. This can be obtained with wider bandwidths, which translate in shorter serialization times. The other unavoidable component of the latency, the propagation time, is determined by the physical distance, something hard to optimize. However, in many case, additional contributions add to this base latency (i.e. serialization + propagation). First, at the link level, transmission can only begin once a clock has been recovered and once the receiver has locked on the framing. If thermally sensitive optical devices are used, thermal stabilization must additionally precede clock and frame locking. With packet based transmission schemes, this initialization happens prior to any transmission, and therefore an initialization time systematically adds to latency. Second, at the network level, two or more transmissions might contend for network resources, which must be arbitrated. As the network is distributed over small but non negligible distances, an arbitration protocol is required. Various such protocols have been proposed, leading to distinct contributions to latency [8].

Link initialization times are not easy to predict. They depend both on the behavior of the protocol state machine (implemented with registers and discrete logic components), and on the evolution of the (analog) optical received signal – thus on the transition times of optical components. Bounds can be derived by summing up worse case transition times on one hand, and worse case protocol state advancements, on the other hand. Physical implementations (test-bed) can be used to validate these bounds. With respect to network congestion related latencies, those depend not only on the optical devices characteristics and on the resource arbitration protocol. They additionally depend on the network state, which is itself dependent on the traffic requests issued in the (more or less) recent past, and on the network structure – different network designs lead different sets of communicating pairs to compete for shared resources. To estimate these congestion latencies, event-driven simulation models must be leveraged. Assuming different types of network solicitations (low vs.

high load, uniform across the communicating pairs vs. a few pairs at a time, etc. – generated by traffic generator components in the simulator), the joint impact of the network architecture (and of the bandwidth of its links) and of the protocols on the performance (i.e. latency) can be quantified; and so can be the costs (power consumption, area footprint, etc.) associated with a one architecture.

Once the first cost/performance metrics obtained, one can start an optimization process, changing the protocol and/or the network architecture. As the network architecture changes, so does the sequence of filters applied to the optical signal. Filter parameters must therefore be recomputed, leading to new bandwidths. New devices also have to be designed to obtain the desired high level-parameters (as, one more time, κ, t, α, n_{eff} in the ring resonator case).

2.3 From the message latencies to parallel application performance

We may be able to optimize an interconnect for a type of traffic solicitation, but we still miss the final element of our scalable approach for optics based data-movement modeling. Hence, in a distributed computing context, each application may present its own type of network solicitation, which will challenge the protocol/architecture in a unique way. Additionally, latencies induced by the interconnect affect the proceedings of the application, and therefore its upcoming solicitations.

There no other options than to include at least partly the application in the model to reach this ultimate modeling goal. This can be achieved by replacing the traffic generator components by the application itself. But since this basically means simulating all aspect of application execution (CPU, OS, memory hierarchies), this approach lacks the desired scalability. Simplifications must therefore be operated at the application level, to capture only the data-movement aspect of the program. We reviewed such possible simplifications in and proposed a novel one in [9].

3. Software implementation

The integrated modeling approach, presented in abstract terms in the previous sections, must also be translated in a software implementation. This step alone is particularly challenging. Functionalities from very diverse fields of modeling (i.e. from multiphysics modeling to signal processing, to discrete-event simulation) must be provisioned. If signal processing or discrete-event simulation operations can be achieved in home-brewed software, numerical solvers on one hand, and CPU/application simulators, on the other hand, are extremely complex software. They cannot be developed on purpose, and existing packages (COMSOL; SST-Macro or SST-Micro) must therefore be procured.

Once all the critical software functions are at hand, generally in several software blocks, these blocks must additionally be *interfaced* among themselves, forming a tool suite. Otherwise, results from a block (e.g. the device parameters) must be manually introduced or adapted for the next tool (e.g. signal processing part). Our software efforts toward developing such a suite are detailed in References 10 and 11. In the future, one may see all the components described here integrated in a single software, able to operate an automated cross-layer optimization of an optical interconnect for an application. Internally, however, this software will iterate from one model to another, attacking the problem from successive sides, at least initially.

4. Conclusions

The integrated modeling approach presented in the previous section is still in a prototype state. Additional aspects need to be explored: on the signal side, for instance, the model should support gain capable devices, while on the “application side”, the discrete-event simulation approach, even with simplified assumptions, might not be scalable enough, calling for fluid (e.g. leaky-bucket) or packet-train approaches instead. Nevertheless, it covers the most significant aspects of the inclusion of photonics in large scale distributed computing, it enables a fast exploration of the design space, and therefore should ultimately support our initial co-design objective.

5. References

[1] I O'Connor, G. Nicolescu, “Integrated Optical Interconnect Architectures for Embedded Systems,” Springer, 2013.

[2] S. Rumley, et al, “Impact of Photonic Switch Radix on Realizing Optical Interconnection Networks for Exascale Systems” Optical Interconnects, 2014.

[3] A. Beling, et al., "InP-based waveguide photodiodes heterogeneously integrated on silicon-on-insulator for photonic microware generation", Optics Express, Vol. 21, Issue 22, pp. 25901-25906 (2013)

[4] A. Yariv, "Critical Coupling and Its Control in Optical Waveguide-Ring Resonator Systems", IEEE Phot. Tech. Letters, 14(4), 2002.

[5] C. Kaalund, G.-D. Peng, "Pole-Zero Diagram Approach to the Design of Ring Resonator-Based Filters", IEEE/OSA JLT 22(6), 2004.

[6] R. Hendry, et al. "Physical layer analysis and modeling of silicon photonic WDM bus architectures", HiPEAC workshop, 2014.

[7] N. Ophir, et al. "Silicon Photonic Microring Links for High-Bandwidth-Density, Low-Power, Chip I/O", IEEE Micro, 2013.

[8] K. Wen, et al. "Reducing Energy per Delivered Bit in Silicon Photonic Interconnection Networks", Optical Interconnects, 2014.

[9] S. Rumley, et al. "A Synthetic Task Model for HPC-Grade Optical Network Performance Evaluation", IA^3 workshop, 2013.

[10] M. Glick, et al. "Modeling and Simulation Environment for Photonic Interconnection Networks in High Performance Computing", IEEE ICTON, 2013.

[11] S. Rumley, et al. "Fast Exploration of Silicon Photonic Network Designs for Exascale Systems", ASCR/DOE workshop on Modeling & Simulation of Exascale Systems & Application, 2013.