DimLightSim: Optical/Electrical Network Simulator for HPC Applications

Hugo Meyer, Jose Carlos Sancho
Barcelona Supercomputing Center
{hugo.meyer; jose.sancho}@bsc.es

Abstract *– Optical Packet Switches (OPS) and Optical Circuit Switches (OCS) provide the needed low latency transmissions in today large data centers and HPC systems. These switches can deliver lower latency and higher bandwidth than traditional electrical-based switches. Although light-based transmission has its advantages over electrical-based transmissions, in optical devices packet collisions are possible and this can generate retransmissions. In this work we present an optical network simulator called DimLightSim. DimLightSim models communication events in optical devices at packet level by replaying real application traces. Different experimental evaluations have been made using DimLightSim in order to compare current datacenter networks with the fully optical Architecture-on-Demand (AoD) proposed in the Lightness project. Initial results helped to foresee the impact in HPC applications execution time. In terms of performance improvement, the AoD architecture can outperform Infiniband-based network up to 19%.*

I. IINTRODUCTION

Data centers are growing in size and complexity to accommodate the ever-increasing demand of High Performance Computing (HPC) applications. One of the most challenging issues when scaling out a data center is the network infrastructure. As the size of data centers increases, higher volumes of data have to be transported among thousands of servers very fast. It is predicted that applications will need in the order of several Terabit/s of bandwidth in the near future. In addition, to provide enough network bandwidth there is also need to provide fast access to data, specially to HPC applications, where, for many applications, low latency network is critical to achieve high scalability.

Optical-based network has currently been explored to overcome this bandwidth and latency bottleneck in data centers. The deployment of optical devices leverage on Dense Wavelength Division Multiplexing (DWDM) allows the transmission of more than a hundred of wavelength channels operating at 10, 40, 100 Gb/s and beyond.

Basically, current optical switching architectures are based on Optical Circuit Switching (OCS) and Optical Packet Switching (OPS). OCS is capable to accommodate long-lived high-capacity smooth flows with little latency whereas OPS is ideal for dynamic traffic such as HPC. In order to take advantage of the benefits from both OCS and OPS, a novel architecture-on-demand (AoD) function programmable data center network architecture with the integration of OCS and OPS was recently proposed [1]. AoD is allocating applications to different optical switches depending on the communication characteristic of applications. HPC applications are desirable to be allocated to OPS as their traffic is dynamic. The OCS may not be suitable for HPC traffic in some scenarios, since the mirror reconfiguration time of OCSs is around 25 ms. On the other hand, packet collisions may occur when using OPSs since packets cannot be stored in these switches.

In this work we describe an optical packet-level network simulator named *DimLightSim* that has been designed to model the behavior of fully-optical Datacenter Networks. *DimLightSim* have been designed in order to evaluate the impact of optical network components in HPC applications. *DimLightSim* allows foreseeing how variations in latencies, packet retransmissions, bandwidth, among others, affect execution time of applications.

II. DIMLIGHTSIM: PACKET LEVEL OPTICAL SIMULATION

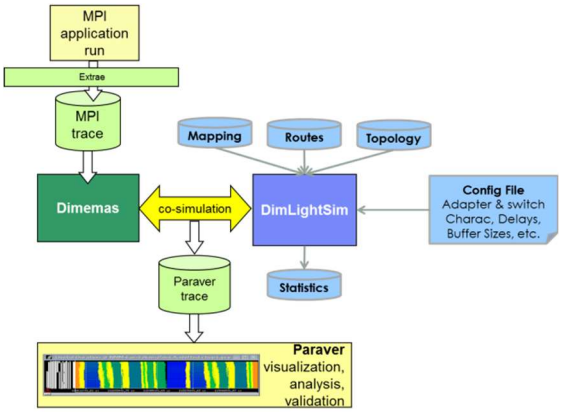

Fig. 1. Optical Network Simulator Framework.

The simulation framework of *DimLightSim* is composed of four open-source tools: a) *Extrae*: extracts information that includes timestamps of events such as message transmissions and other runtime calls; b) *Dimemas*: it reproduces the events from the trace. Communication events are forwarded to *DimLightSim*. c) *DimLightSim*: is the packet level network simulator that models optical devices. It has been developed using the Omnet ++ framework. d) *Paraver*: is a visualization and analysis tool of the computation and communication events.

Fig. 1 depicts the simulation framework and how the different elements interact. *DimLightSim* does the MPI process mapping, setup the preferred network topology, and the routing information as well in order to forward packets from source servers to destination servers. *DimLightSim* allows users to configure the network topology and the routing information in switches.

DimLightSim and *Dimemas* participate in a co-simulation where they take turns during the simulation execution. *Dimemas* starts the simulation execution processing the communication and computation events in the application

trace. *DimLightSim* is waiting for messages from *Dimemas*. When the next event in the application trace is a message transmission then *Dimemas* creates a message and forwards it to *DimlightSim*. *DimLightSim* proceeds with the simulation execution till the transmission is finished or it is interrupted by *Dimemas* because other communication event needs to be scheduled.

III. Experimental Evaluation

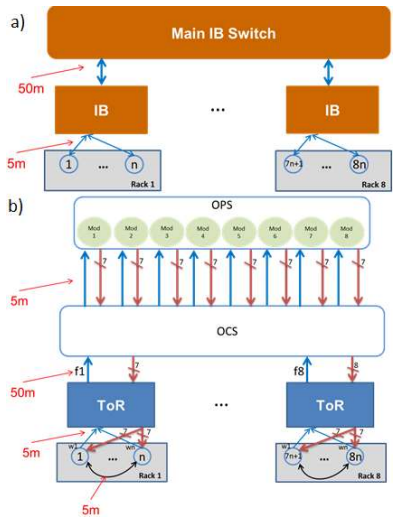

Fig. 2. Experimental Networks. a) IB network with 8 racks, 8 Top-of-the-Rack (ToR) switches and a main IB Switch. b) Optical AoD network with 8 racks, 8 ToRs connected to an OCS and one OPS.

DimLightSim has been validated in [2] by comparing real transmission times with simulated times. It is also very important to highlight that all the parameters of the devices used in *DimLightSim* correspond to real measures taken from the optical devices [3].

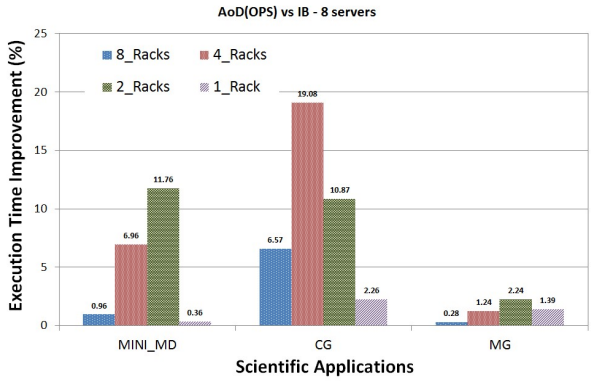

Fig. 3. Performance Comparison of OPS and Infiniband Switching using DimLightSim.

Fig. 2 shows the experimental configurations used to compare an IB-based network with an optical AoD network. Considering the experiments made using AoD infrastructure, NICs are able to communicate directly between each other when residing in the same rack and for communication between racks, all traffic goes through the OPS switch. Optical Top-of-the-Rack (ToR) switches are in charge of multiplexing/demultiplexing optical wavelengths to fibers in a negligible time. The next configuration parameters were set: NIC Delay=300ns; OPS Delay=25ns; IB switch Delay=200ns; Bandwidth = 8 Gbps. The cables lengths are depicted in Fig. 2.

Fig. 3 depicts results obtained using *DimLightSim* when comparing Infiniband (IB) switching with OPS assuming the usage of the AoD network infrastructure. According to the obtained results, OPS-based networks can outperform IB-based networks in up to 19% in terms of execution time. The showed results consider also the penalty of packet retransmission in the execution time when using OPS.

Further results and proposals that base their research in *DimLightSim* can be found in [3, 4].

IV. Conclusions

Foreseeing the impact of new network technologies in HPC applications is highly important in order adapt or configure properly these new improvements. In this work we presented *DimLightSim,* which is able to mimic the behavior of optical networks at a packet level and enable us to analyze how the technology improvements and its limitations affect execution of applications and resource usage. *DimLightSim* has been used in the Lightness project [5] and other works to drive design decisions and foresee performance impact in applications. Obtained results helped to determine the benefits of optical networks when comparing to electrical-based networks. In particular, OPS can outperform in up to 19% Infiniband-based networks.

Acknowledgment

This work has been supported by the FP7 European Project LIGHTNESS (FP7-318606).

References

[1] S. Peng, D. Simeonidou, G. Zervas, R. Nejabati, Y. Yan, Y. Shu, S. Spadaro, J. Perello, F. Agraz, D. Careglio, H. Dorren, W. Miao, N. Calabretta, G. Bernini, N. Ciulli, J. C. Sancho, S. Iordache, Y. Becerra, M. Farreras, M. Biancani, A. Predieri, R. Proietti, Z. Cao, L. Liu, S. J. B. Yoo, *A novel sdn enabled hybrid optical packet/circuit switched data centre network: The lightness approach*, in*: Networks and Communications* (EuCNC), 2014 *European Conference* on, 2014, pp. 1{5. doi:10.1109/EuCNC.2014.6882622.

[2] Hugo Meyer, Jose Carlos Sancho, Wang Miao, Harm Dorren, Nicola Calabretta, and Montse Farreras. *Performance Evaluation of Optical Packet Switches on High Performance Applications*. In Waleed W. Smari, editor, *Proceedings of the 2015 International Conference on High Performance Computing & Simulation* (HPCS 2015) Amsterdam, the Netherlands, pages 356–363. IEEE Computer Society, 2015.

[3] Wang Miao, Jun Luo, Stefano Di Lucente, Harm Dorren, and Nicola Calabretta, "Novel flat datacenter network architecture based on scalable and flow-controlled optical switch system," *Opt. Express* 22, 2465-2472 (2014).

[4] Hugo Meyer, Jose Carlos Sancho, Milica Mrdakovic, Shuping Peng, Dimitra Simeonidou, Wang Miao, and Nicola Calabretta. *Scaling architecture-on-demand based optical networks.* In *Proceedings of the 17th International Conference on Distributed Computing and Networking*, ICDCN '16, pages 10:1–10:10, New York, NY, USA, 2016. ACM.

[5] *Low Latency and High Throughput Dynamic Network Infrastructure for High Performance Datecentre Interconnects* (Lightness) European Project, 2012 [online]. Available: www.ict-lightness.eu.