Combining Hardware and Simulation for Datacenter Scaling Studies

Sarah Ruepp*, Artur Pilimon*, Jakob Thrane, Michael Galili, Michael Berger, and Lars Dittmann

DTU Fotonik, Dept. of Photonics Engineering
Technical University of Denmark
{srru,artpil,jathr,mgal,msbe,ladit}@fotonik.dtu.dk

Abstract— **Datacenter networks are becoming crucial foundations for our information technology based society. However, commercial datacenter infrastructure is often unavailable to researchers for conducting experiments. In this work, we therefore elaborate on the possibility of combining commercial hardware and simulation to illustrate the scalability and performance of datacenter networks. We simulate a Datacenter network and interconnect it with real world traffic generation hardware. Analysis of the introduced packet conversion and virtual queueing delays shows that the conversion efficiency is at the order of a few microseconds, but the virtual queuing may have significant implication on the performance analysis results.**

***Keywords*— Datacenter, simulation, system-in-the-loop**

I. Introduction

Datacenters and datacenter networks (DCNs) are becoming increasingly important to our current and future communication infrastructure. Network connectivity is omnipresent from a multitude of devices, and they all generate data that must be stored and processed. This puts an enormous strain on datacenters and requires novel approaches to ensure scalability of the underlying DCN infrastructure.

Unfortunately, current DCN architectures are not easily scalable, and current solutions impose unsustainable overheads in terms of capacity, connectivity and energy consumption requirements [1]. It is thus essential to develop fundamentally new hardware technologies, internal datacenter network connection infrastructures joined with advanced mechanisms for control and service orchestration.

A challenge faced by researchers is that real world datacenters often are closed systems, and building up commercially-sized datacenters simply for research purposes is generally highly unfeasible from both an economic and a footprint wise perspective. On the other hand, simply conducting datacenter research by means of mathematical analysis or simulation does not visualise the effect of using real components (e.g. processing delays, setup constraints, timings, etc.), and may thus provide unrealistic results when concepts are to be ported to real networks. In other words, what may work on a single computer for research purposes may not be feasible in a commercial datacenter infrastructure.

In this work, we thus showcase the approach of combining hardware and simulation for datacenter performance scaling studies using Riverbed's System-in-the-Loop (SITL) tool [2]. This gives the possibility to use commercial components and their respective properties whenever available, and to evaluate different scalability and performance aspects without having to acquire enormous amounts of expensive components.

The remainder of this paper is organized as follows: section II deals with optical datacenter networks. In section III, the combined hardware and simulation setup is explained. Section IV presents the results and the paper is concluded in section V.

II. Optical Datacenter Networks

One of the challenges the datacenter industry is facing is to move away from vendor-specific, hierarchical, statically controlled and poorly scalable infrastructures. In the context of the EC FP7 project COSIGN [1], the partners pursue the development of scalable, low-latency, cost-effective, versatile multi-technology datacenter networks, that can combine the benefits of introducing optics and SDN (Software Defined Networking) into the datacenter ecosystem (shown in Figure 1).

Datacenter network architectures have traditionally been built over variations of the Fat-Tree, BCube, D-Cell, flattened butterfly, etc. [3]. Newer proposals contain a Ring-of-Rings (RoR) architecture [4], which benefits from a high internal robustness due to the inherent protective ring architecture.

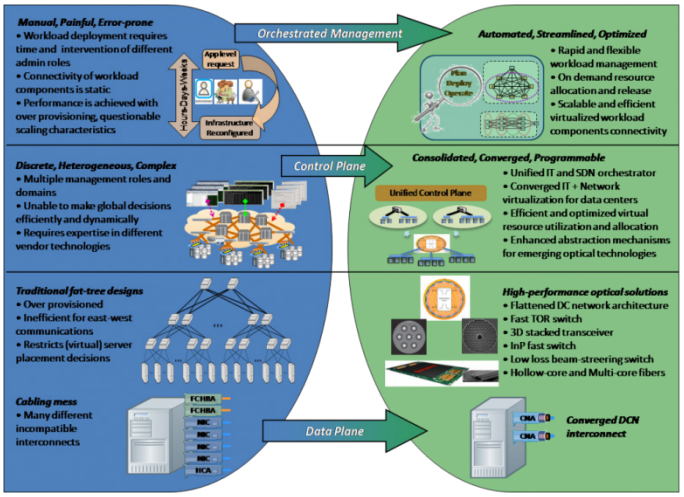

Figure 1: COSIGN datacenter evolution [1]

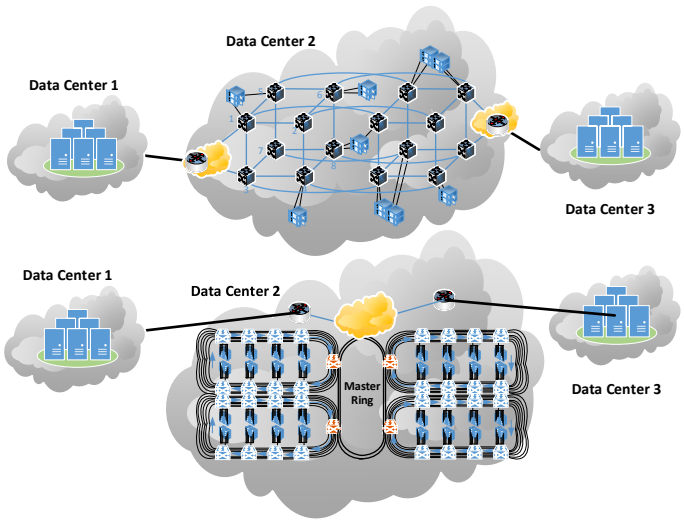

Figure 2: HyperCube and Ring-of-Rings architectures

Additionally, HyperCube and HyperX structures [5] have been widely used in the high performance computing (HPC) industry, and due to their benefits in terms of low latency and high scalability propose themselves as promising candidates for future datacenter architectures.

Furthermore, the introduction of optical components and optical switches can significantly increase the amount of data that can be transmitted within a datacenter. Figure 2 illustrates how Top of Rack (TOR) switches in datacenters are interconnected using either the Ring-of-Rings or the HyperCube topology in an inter-DC communication scenario.

III. System In The Loop Simulations

In order to facilitate the performance analysis and scaling studies of datacenters, simulation and especially integration of real hardware into the simulations, Riverbed's System-in-the-Loop (SITL) tool [2] provides some very powerful features.

The tool provides the linkage between the "real world" and the simulation that is running inside a computer. Any type of device, e.g., a router or a switch, can be linked to the simulation via the workstation's Ethernet port. Let's assume a packet that is generated by a real server, processed in a simulated datacenter network, and terminated by another real server. Once the packet is generated on the real server, it will be transmitted via Ethernet to the workstation running the simulation. Most importantly, the real time and the simulation time are running continuously, ensuring that timing is kept consistently throughout the entire system. The real packet is then converted into a simulation packet that will then be processed throughout the simulation. When the packet has passed through the desired simulation path, it is mapped at the external interface and converted back to a real packet. The real packet is then sent from the workstation through the Ethernet interface towards the real equipment. An abstracted packet translation procedure is illustrated in Figure 3.

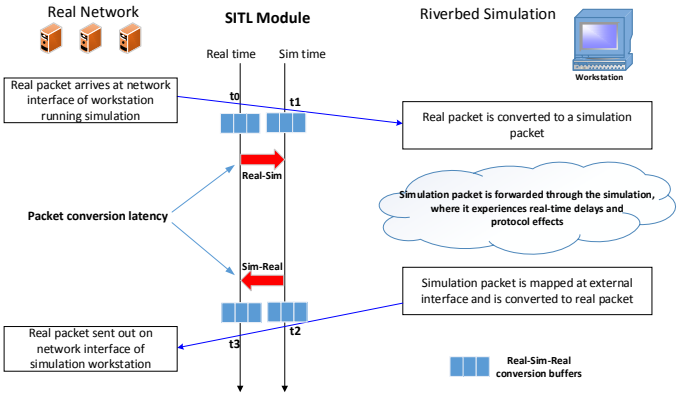

Figure 3: Packet flow between real and simulation equipment

IV. Simulation Scenario

In this work, the Ring-of-Rings and the HyperCube architectures are analysed. Unfortunately, having such a setup containing real datacenter hardware available for experimental and research purposes is often unrealistic due to high cost and space (footprint) requirements.

We are using two state of the art traffic generators, namely Xena Testers [7], to generate traffic on different layers between Datacenter 1 and Datacenter 3. The traffic is passed through Datacenter 2, which is modelled using an internal architecture of either HyperCube or Ring-of-Rings.

Every topology is composed of a multitude of TOR switches, each of them having multiple servers attached. Thus, the architectures are built as simulation models to illustrate scalability without having to acquire multiple TOR switches.

The packet translation efficiency is highly dependent on the translation level needed (see Figure 4). Some traffic may not be terminated within the DCN, hence certain payload types will be just copied as a block of bits, not touching the corresponding headers, resulting in faster translation. The simulation configuration (setup) and real hardware is shown in Figure 5, where the simulation model and real equipment are linked via the aforementioned virtual SITL gateways.

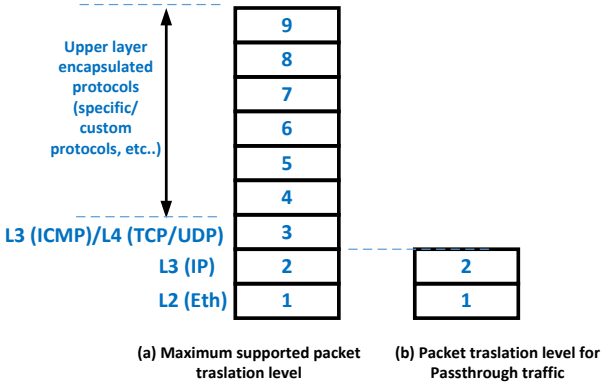

Figure 4: Packet translation depth (level) at the real/simulated interface

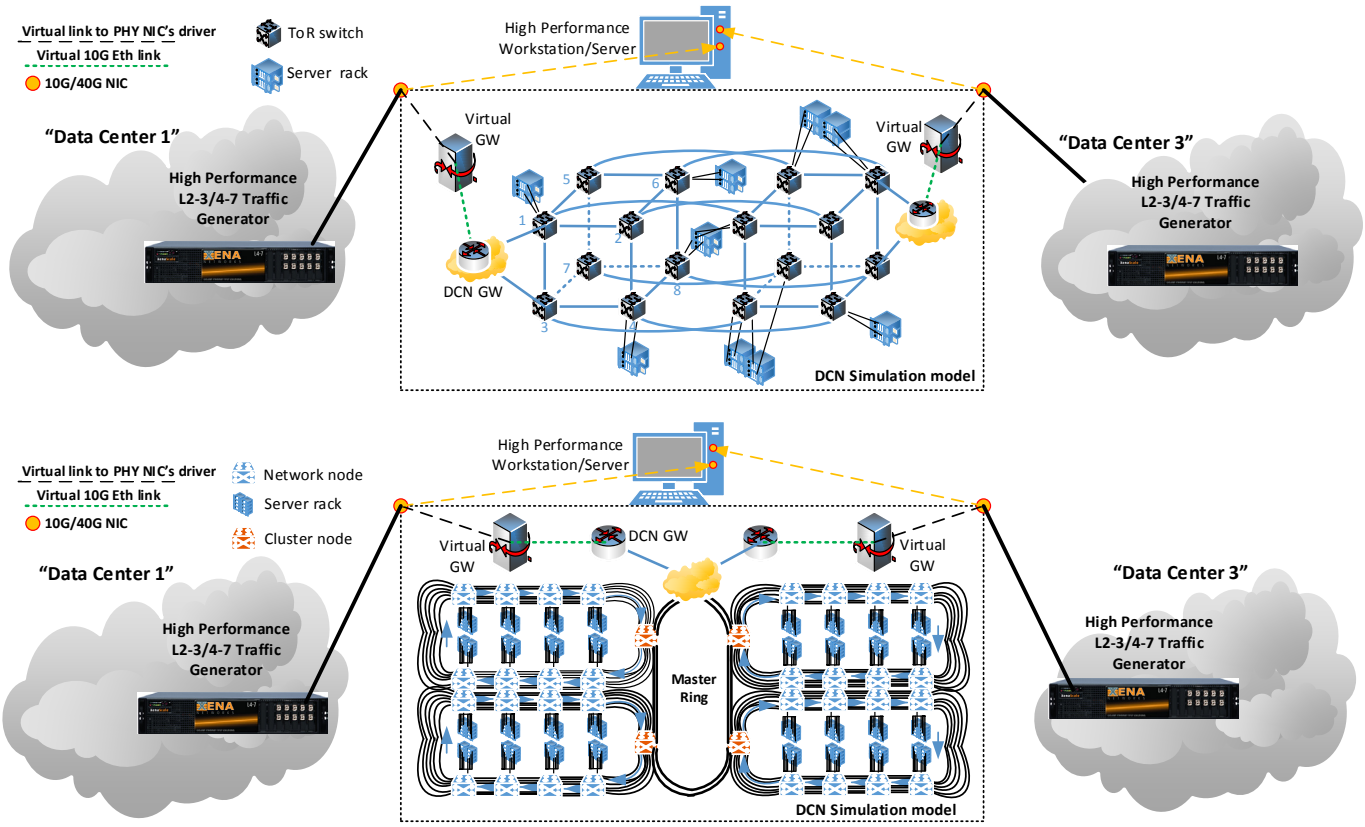

Figure 5: Experimental setup combining real equipment and simulation

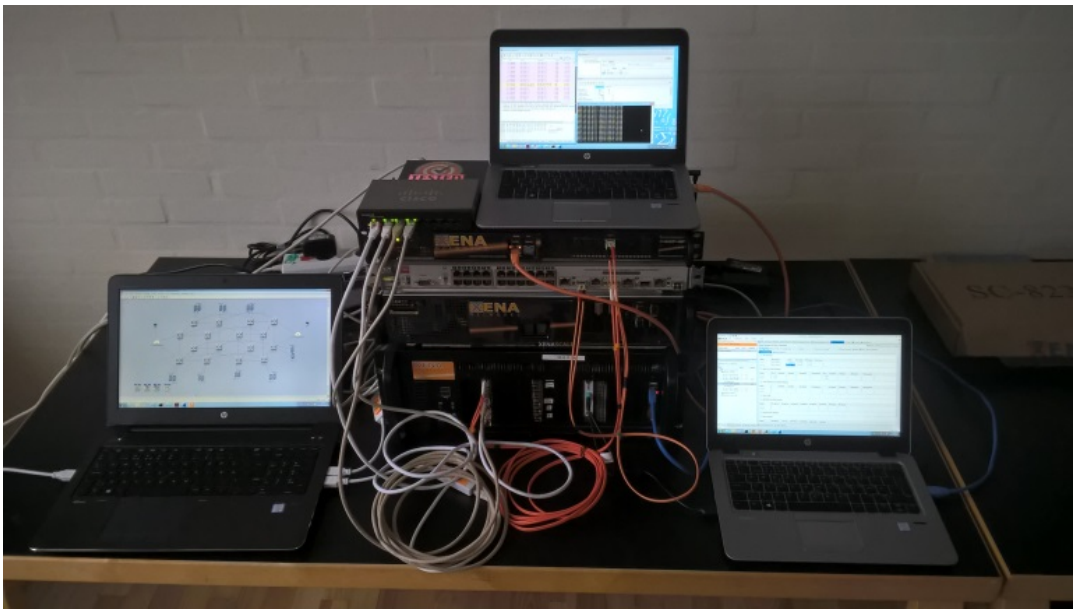

Figure 6: Picture of the experimental setup

The entire system setup is shown in Figure 6. Note that in the depicted experimental setup we used portable computers for the initial demonstration tests. In this case a maximum rate of 80 and 120 Mbit/s was achieved in our tests for the ICMP and TCP traffic, respectively, on a 1GE network interface due to encountered Windows socket buffer overflow (operations on non-blocking sockets that cannot be completed) as described in [8]. Large scale experiments will run on a dedicated set of more powerful servers. The "workstation" on the left (black) is running the simulation model of datacenter 2. The Xena testers in the middle are emulating datacenter 1 and datacenter 3 by generating different predefined traffic patterns.

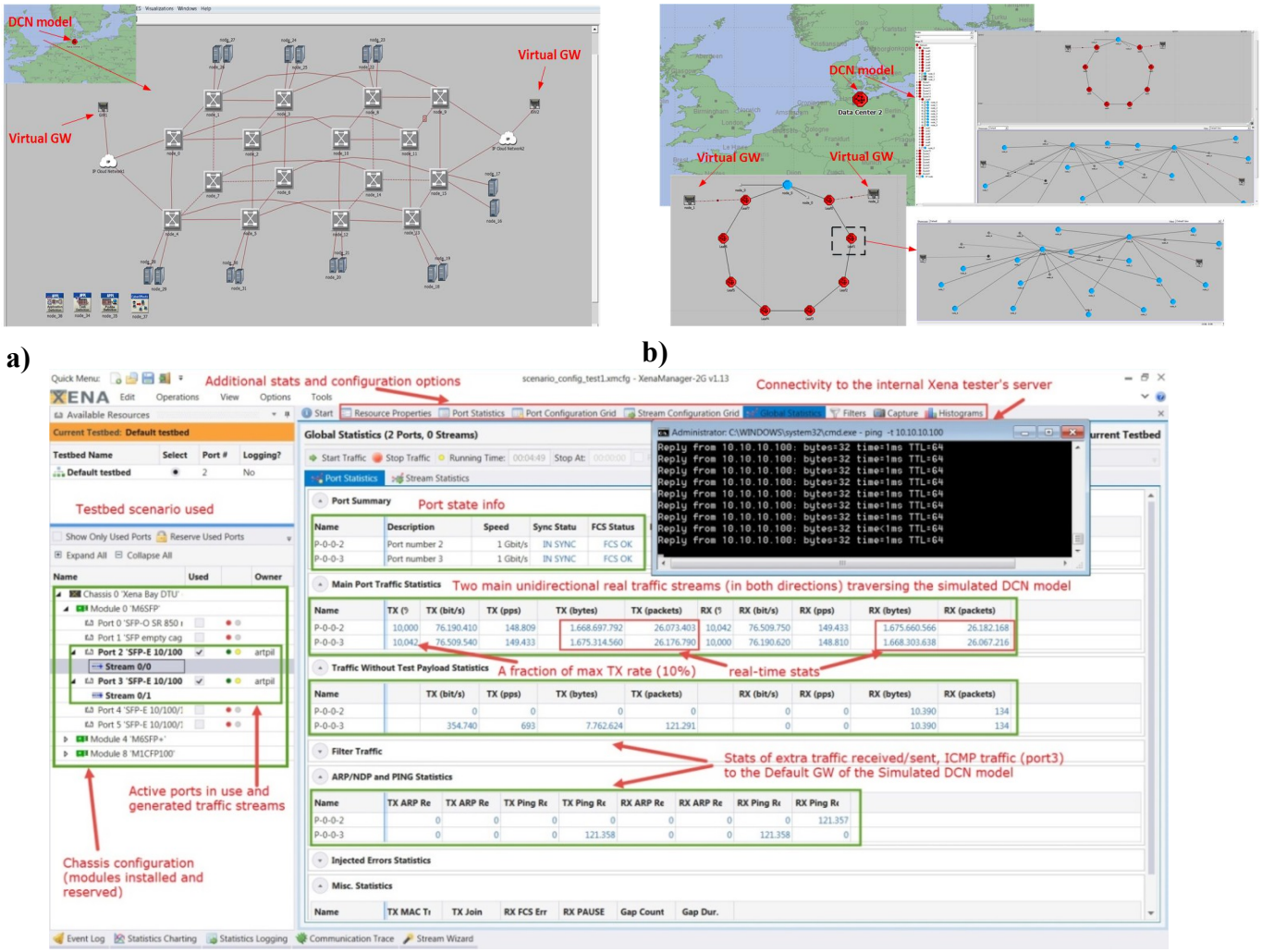

Figure 7: Software components of the testbed setup: a) Simulation model of HyperCube-16 architecture; b) Simulation model of Ring-of-Rings architecture; c) Picture of Xena L2-3 tester

The main benefits of using these high performance testers are: a) possibility to generate realistic application layer communication patterns (*pcap*-based replay); b) possibility to emulate millions (when we need scale) of concurrent traffic flows by utilizing multiple available transceiver modules (1GbE, 10GbE, 40GbE and 100GbE rates, depending on the modules used). For a realistic experimental case study it is sufficient enough to have a few 10G or 40G modules, since present day workstations or servers can be equipped with 10G or even 40G NICs (Network Interface Card) to provide a reasonable uplink/downlink for our modelled DCN.

The laptop on top of the system is analysing packets running a Wireshark tool. The software for the Xena testers is executed on the laptop on the right side of the picture.

Detailed screenshots from the individual computers and models can be seen in Figure 7, showing the simulation models and Xena tester software interface, respectively.

It is important to point out that evaluation of the packet translation latency parameter at the real-simulated network interface is crucial in the context of DCN performance analysis, because of much more stringent timing requirements, compared to conventional networks. In modern (and future) DCN environments there are several critical factors, which set DCNs aside into a different "networking" category, namely ultra-low latency and high throughput requirements, ultra-short duration of vast majority of the intra-DC traffic flows (in the order of a few 10s or 100s of *ms*), significantly larger east-west (internal, remaining within a DC) traffic volumes as compared to south-north (external) traffic. That is the reason for assessing the feasibility of using such a hybrid system first.

V. Results

One of the most important measures in datacenter networks is latency. We thus measure the time that it takes to traverse the SITL gateway nodes in both directions, namely the conversion delay on this virtual interface. This parameter is of paramount importance when it comes to the further performance evaluation of the simulated DCN topology in terms of delay, because it directly affects the accuracy of the obtained measurement results.

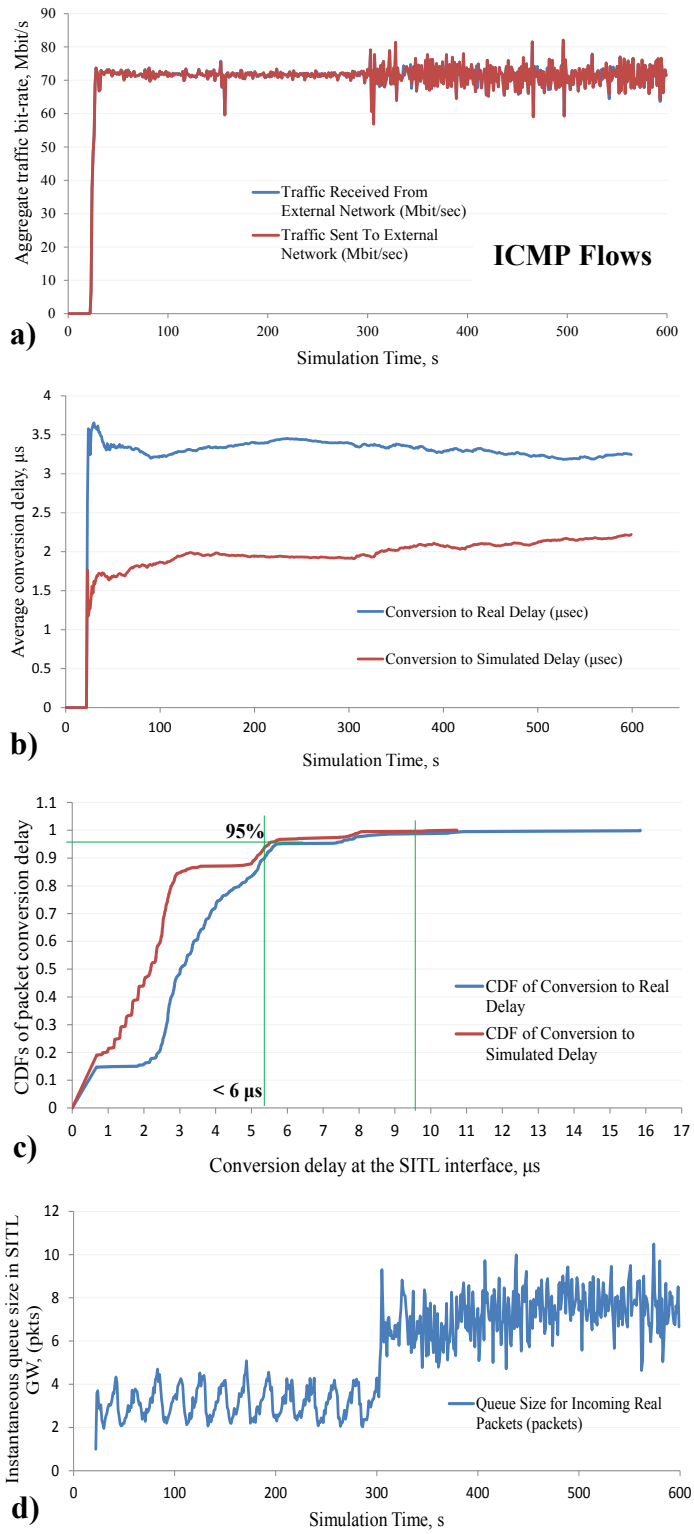

Figure 8: Results of Simulation model's stress-testing. Packet conversion efficiency at the SITL GW interface. Test1

The stress-testing was performed by loading (symmetric bidirectional traffic) the transit (simulated) DCN, datacenter 2, with a large number of high bit-rate ping flows, launched sequentially using a script. The results are illustrated in Figure 8. We observe that under the load of ~ 75-80 *Mbit/s* (see Figure 8*a*), the time-average conversion delay for the incoming traffic (real-to-simulated) is fluctuating around 2.0 *µs* per packet, while in the opposite direction (simulated-to-real) it is almost 60% higher. However, this is a relative difference, since this result shows a time-average value. It's more interesting to look at the Cumulative Distribution Function (CDF) of this parameter, being a better indicator. As it can be seen in Figure 8 (*c*), for around 95% of collected samples, the per packet conversion delay is less than 6 *µs* on average for both directions, with the worst case scenario being below 10 µs. As a result this penalty must be taken into consideration while evaluating the performance metrics of a DCN under consideration. The evolution of the queue size in virtual SITL gateway is shown (Figure 8*d*) to be relatively low (3-8 packets on average), but this parameter is very important, since it will affect the queueing delays under much higher traffic loads.

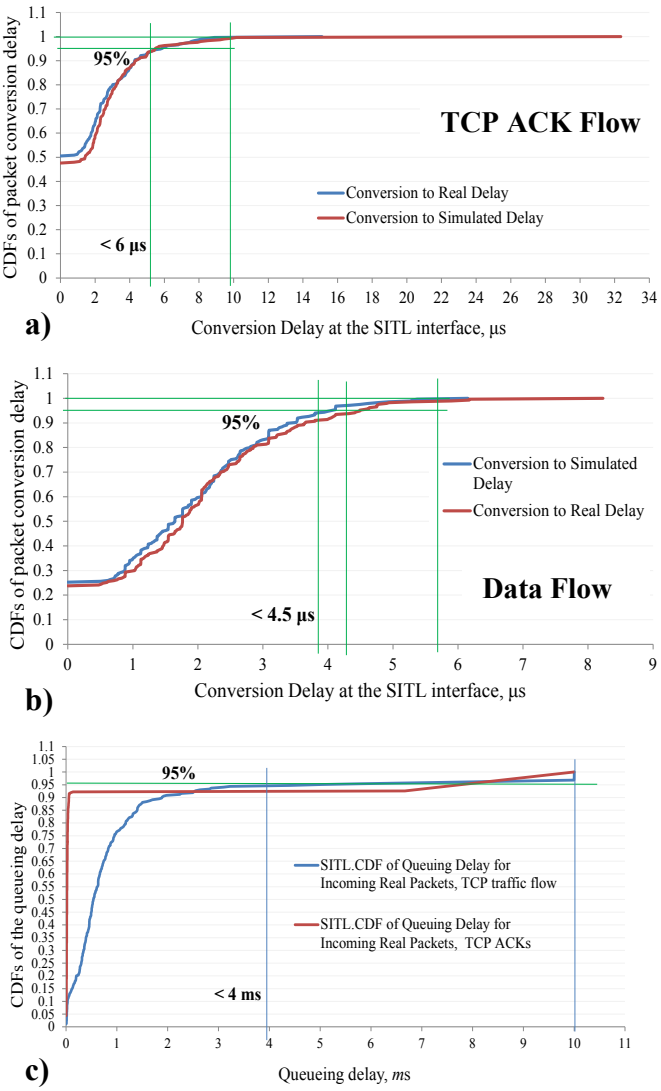

Figure 9: Results of Simulation model's stress-testing. Packet conversion efficiency at the SITL GW. Test2

In another test the SITL gateways were loaded with asymmetric traffic flow, namely by emulating a transmission of a large data file (3.41 *GB*) via TCP connections (data in the forward direction, streams of ACKs in the reverse) with the average data rate of 120 *Mbit/s*. Transmission profile followed a bursty traffic profile configured.

We analysed the CDFs of bidirectional packet conversion and queueing delays, presented in **Figure 9** (*a* – CDF for the TCP ACK flow, *b* – Data traffic flow, *c* – queueing delays for the incoming real packets). As can be seen, statistically, the conversion delay is at the order of a few *μs* in both directions, and 95-percentile latency is in the same range for both flows. However, considering the proportion of packets corresponding to each flow (Data and TCP ACKs), the average number of Data packets per second (pps) was around 14000, whereas for the TCP ACK stream it was ~ 50% of that, namely around 7000 pps. This proportionality is expected, since by default TCP connections were using a *delayed ACK* mechanism. Thus, this dependency is reflected in the conversion delay results in **Figure 9** (*a*) and (*b*), where the conversion delay of the 50% of the TCP ACK packets is less than 1 μs, while 50-percentile of the data packets experience delays twice as large (~ 2 μs).

When it comes to the packet queueing in SITL, **Figure 9** (*c*) shows the delay experienced by more than 92% of ACK stream packets is under 1 *ms* (at the order of a few *μs*), while data packets are queued up to 4 *ms* (95-percentile), with the worst case scenario of ~ 10 *ms*. The latter values are relatively high and will have a serious impact on the performance results.

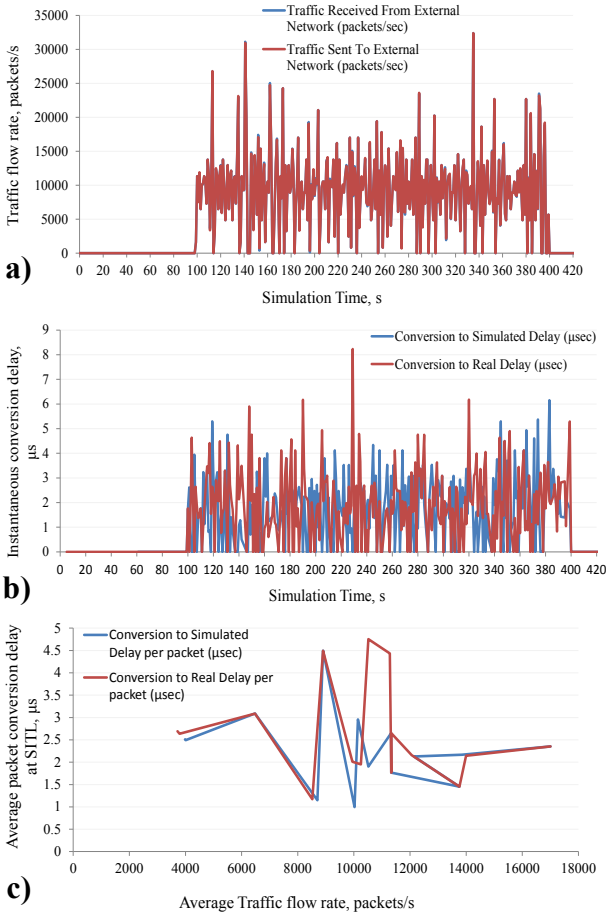

Figure 10: Dependency of the packet conversion latency on the traffic flow rate in virtual SITL gateway node

We evaluated the dependency of the packet flow rate on the average conversion delay by statistically sampling the packet rates and corresponding (by simulation timestamp) conversion latency using the obtained distributions (Figure 10 *a, b*) and the preliminary results (see Figure 10 *c*) show that there is no clear link between the packet rate and conversion delay incurred, and the stochastic nature may be a result of several other factors, such as specifics of packet capture by the WinPcap [9] (libPcap for Linux) module, implementation of the conversion functions (code) and characteristics of the NIC installed (buffering, protocol checksum offload, etc.).

VI. Conclusion

In this paper, we have detailed an approach for combining real hardware and simulation for the purpose of evaluating the performance and scalability of datacenter networks. We describe how the Riverbed System-in-the-Loop (SITL) tool can be used to interconnect real world and simulation under continuous timing constraints, without having to invest in vast amounts of expensive hardware. Our results show that the SITL gateway adds a conversion delay in the order of microseconds as well as load-dependent buffering delays that must be taken into consideration for any latency measurements.

The approach presented here, showcasing the combination of real hardware and simulation, has an enormous potential in the emerging integration of optics in the datacenter world, where scaling and latency effects must be studied without necessarily having access to real datacenter infrastructures.

Acknowledgment

This work has been partially supported by the EC FP7 project “COSIGN, grant no. 619572”, and the Innovation Fund Denmark project “Layer 4-7 Testing at 100 Gbps”.

References

[1] EU FP7 Project Cosign, Combining Optics and SDN In next Generation data centre Networks, http://www.fp7-cosign.eu/

[2] Riverbed System in the Loop Tool (SITL). www.riverbed.com (formerly known as OPNET Inc.)

[3] D. Abts, J. Kim. “High Performance Datacenter Networks: Architectures, Algorithms, and Opportunities”, in Synthesis Lectures on Computer Architecture (2011)

[4] A. M. Fagertun, M. Berger, S. Ruepp, V. Kamchevska, M. Galili, L. K. Oxenløwe and L. Dittmann (2015). “Ring-based All-Optical Datacenter Networks”, in Proc. of European Conference on Optical Communications (ECOC) , Valencia, Spain

[5] J. Ho Ahn, N. Binkert, A. Davis, M. McLaren, R. S. Schreiber. “HyperX: Topology, Routing, and Packaging of Efficient Large-Scale Networks”, in Proc. of the Conference on High Performance Computing Networking, Storage and Analysis - SC '09 (2009)

[6] SITL session, Opnetwork conference 2010, www.opnetwork.com

[7] Xena Networks, www.xenanetworks.com, Xena Bay L2-3 and Xena Scale L4-7 Testers

[8] Microsoft, “Windows Sockets Error Codes,” 2017. [Online]. Available: https://msdn.microsoft.com/en us/library/windows/desktop/ms740668

[9] The Windows packet capture library, https://www.winpcap.org/.