



**Low latency and high throughput dynamic network infrastructures
for high performance datacentre interconnects**

Small or medium-scale focused research project (STREP)

Co-funded by the European Commission within the Seventh Framework Programme

Project no. 318606

Strategic objective: Future Networks (ICT-2011.1.1)

Start date of project: November 1st, 2012 (36 months duration)



Deliverable D5.3

Demonstration of DC applications over LIGHTNESS system and trials

Due date: 31/10/2015

Submission date: 10/11/2015

Deliverable leader: IRT

Author list: Alessandro Predieri (IRT), Matteo Biancani (IRT), Giacomo Bernini (NXW), Roberto Monno (NXW), Nicola Ciulli (NXW), Wang Miao (TUE), Nicola Calabretta (TUE), Salvatore Spadaro (UPC), Fernando Agraz (UPC), Albert Pagès (UPC), Jose Carlos Sancho (BSC), Yi Shu (UNIVBRIS), Yan Yan (UNIVBRIS), George Saridis (UNIVBRIS), George Zervas (UNIVBRIS), Reza Nejabati (UNIVBRIS), Dimtra Simeonidou (UNIVBRIS)

Dissemination Level

<input checked="" type="checkbox"/>	PU: Public
<input type="checkbox"/>	PP: Restricted to other programme participants (including the Commission Services)
<input type="checkbox"/>	RE: Restricted to a group specified by the consortium (including the Commission Services)
<input type="checkbox"/>	CO: Confidential, only for members of the consortium (including the Commission Services)

Abstract

The aim of this document is to report the experimental assessment of the LIGHTNESS system, including performance evaluation of data centre services running over the SDN control plane and the optical data plane prototypes. Moreover, this deliverable presents the final LIGHTNESS demonstration carried out at the ECOC 2015 conference.

The experimental assessment of the whole LIGHTNESS system has been successfully carried out. All components developed in WP3 and WP4, including two applications running on top of the SDN control plane, have been integrated and evaluated in the testbed, thus enabling the successful demonstration performed at ECOC 2015. In particular, multicast connectivity services have been showcased as crucial services in current data centre environments.

Table of Contents

Table of Contents	3
0. Executive Summary	5
1. Data centre services and applications	6
1.1. Multicast services in data centres: rationale	6
1.2. Sensitivity to multicast in HPC applications	7
2. Experimental assessment of the overall LIGHTNESS architecture	10
2.1. Overall inter-cluster architecture	10
2.2. All-optical reconfigurable data plane testbed setup	11
2.3. All-optical reconfigurable control plane testbed setup	14
2.4. Multicast Virtual Data Centre composition and monitoring	15
2.5. Experimental demonstration and results	17
3. LIGHTNESS system demonstration	22
4. Conclusions	29
5. References	30
6. Acronyms	32

Figure Summary

Figure 1.1: MPI_AllGather collective operation	8
Figure 1.2: MPI_AlltoAll collective operation.....	8
Figure 2.1: The overall architecture of LIGHTNESS DCN solution	11
Figure 2.2: Experimental setup for intra-/inter-cluster communication.....	12
Figure 2.3 Multicasting with OCS scheme	13
Figure 2.4 Multicasting with OPS scheme	13
Figure 2.5 Time traces for OPS label generation, where “M” represents multicast communication is enabled.....	14
Figure 2.6: OpenFlow extensions.....	14
Figure 2.7: VDC composition GUI (top), Power Monitor and OPS Rate Monitor (bottom).....	17
Figure 2.8: BER measurement for intra-rack, intra/inter-cluster scenarios with OCS/OPS schemes	19
Figure 2.9: DMA-to-DMA latency for (a) OCS scheme; (b) OPS scheme	20
Figure 2.10: DMA write and ready latency VS DMA length measured with 256bytes Ethernet packet frame length	20
Figure 2.11: Throughput plot during OCS/OPS switchover	21
Figure 2.12: Message exchange captures from SDN-enabled control plane	21
Figure 3.1 LIGHTNESS demonstration testbed at ECOC2015	23
Figure 3.2 Hardware installed at ECOC2015 LIGHTNESS booth	24
Figure 3.3 LIGHTNESS VDC composition application GUI.....	25
Figure 3.4 LIGHTNESS OpenDaylight GUI	26
Figure 3.5 LIGHTNESS monitoring VNF application GUI	27
Figure 3.6 LIGHTNESS demonstration is performed to Thibaut Kleiner (EC)	28

0.Executive Summary

The continuous growth in size and deployment of cloud and data centre applications is posing more and more challenges in order to support unprecedented amounts of network traffic at data centres on the one hand, and new communication patterns for highly distributed applications on the other. In particular, multicast services are even more becoming crucial within data centres to efficiently cope with highly intensive east-west network traffic in support of either high performance computing jobs or distributed data centre services.

Following this trend, LIGHTNESS has concentrated its last experimentation and validation efforts to assess its full optical SDN enabled data centre network architecture against multicast connectivity services to be provisioned across multiple optical technologies. In particular, all the components developed in LIGHTNESS have been fully integrated and experimentally assessed in an end-to-end data centre system, encompassing physical servers equipped with the novel programmable optical Network Interface Cards (NICs) and grouped in clusters, interconnected following the LIGHTNESS architecture approach by means of optical Top of the Rack (ToR) switches, Optical Packet Switching (OPS) prototype switches and Optical Circuit Switches (OCS). The LIGHTNESS SDN control plane software has been also deployed and installed in this end-to-end testbed to enable the provisioning of data centre network connectivity by means of the extended OpenFlow protocol. Two applications have been used to validate the LIGHTNESS concepts. A Virtual Data Centre (VDC) composition application to provide multicast virtual network slices in the LIGHTNESS full optical data centre, and a monitoring virtual network function (VNF) to collect performance and statistics of installed VDCs and dynamically trigger either technology (i.e. OPS/OCS) or multicast/unicast switch-over. These final experimental activities have also included the collection of results and performances of the whole LIGHTNESS system for BER, latency and throughput measurements, among others.

As a final assessment and validation step for the work carried out in these three years, LIGHTNESS prepared a demonstration event at the ECOC 2015 conference, in Valencia, Spain. Here, in a dedicated project booth, the full LIGHTNESS system was integrated on site and successful demonstrations were performed during the three days event at ECOC. Many attendees experienced the LIGHTNESS demonstration, and lot of positive feedbacks from operators, vendors and researchers were collected as a further prove that the work carried out in the project had a substantial impact in the community and that it is aligned with the trends of future data centres.

1.Data centre services and applications

Multicast services are becoming more and more predominant within data centres due to the continuous growth of emerging distributed applications that need intensive computation and communication resources. Following this trend, LIGHTNESS experimentally assessed the proposed optical flat data centre network architecture against multicast connectivity services, as a further means of validation of its outcomes and benefits with respect to state-of-the-art monolithic approaches.

This section provides a rationale behind this choice, introducing the importance of multicast services and network connectivity in data centre and HPC applications.

1.1. Multicast services in data centres: rationale

The rapid growth of network traffic in data centres experienced in the recent years shifted communication patterns from being predominantly north-south to mostly east-west. Network traffic that was mostly entering and exiting data centres, now is mostly serving emerging applications that need rack-to-rack communications. This increase of east-west traffic has introduced the need of complex and distributed communication patterns that can be supported with multicast services.

Many data centre applications that use distributed file systems for data storage and MapReduce type of processing algorithms rely on multicast communication patterns, such as publish-subscribe services for data dissemination, web cache updates, system monitoring. Multicast traffic is also frequent in other data centre applications such as Virtual Machine (VM) provisioning and software upgrading, where amount of data are transmitted among hundreds of servers to update software packages and installations. Moreover, multicast traffic facilitates and enables one-to-many VM migrations, as a key function for backup and disaster recovery purposes. The usage of multicast communication for these types of applications provides also benefits in terms of efficiency and by increasing the capacity of data centre networks while lowering down operation costs (e.g. at the level of cost per VM per hour).

Today data centre networks, irrespectively of architectures and deployments, do not natively support multicast traffic for east-west communications patterns, and IP multicast is often used in all-electronic (packet) switched data centre networks. However, IP multicast normally requires complex configurations on all the switches and routers within a data centre, and in addition, it faces severe scalability challenges, largely in terms of the number of supported multicast groups and their

robustness against network failures. These challenges are present at both control and data plane level. Moreover, existing multicast protocols lack of recovery strategies in case of network failure conditions: a single point of failure may affect many multicast trees, and reverting a multicast tree often needs lot of network communication and redundant states in switches. In this context, as way to improve scalability, emerging data centre overlay virtualization strategies, such as VXLAN [1] and NVGRE [2], are able to bridge multiple subnets into large layer 2 VLANs while translating broadcast in the virtualized subnet into multicast in the physical network. Despite these mechanisms, IP multicast is not supported in the majority of current data centre networks and application layer solutions are often used to implement multicast services. These methods are inherently inefficient since they introduce large connection overheads, increasing latency while sending multiple copies of the same data. This is more than true in all-electrical packet switched data centre networks, where, due to the switching cost and cabling complexity, providing non-blocking multicast services is a further challenge and operators are often forced to rely on over-subscription.

In this complex scenario, the LIGHTNESS approach of a full optical flat data centre network able to support scalable multicast services at the optical level by leveraging on a flexible SDN approach that allows to avoid a full implementation of IP multicasting functions (mostly at the control plane level) provides a solution to address the emerging data centre multicast services requirements.

1.2. Sensitivity to multicast in HPC applications

Multicast is a collective communication operation where one source sends data to more than one destination. This is one of the most frequently used collective operations on High Performance Computing (HPC) applications. Typically, multicast is the basis to implement high-level collective communication operations in modern parallel programming models like Message Passing Interface (MPI). MPI offers `MPI_AllGather` and `MPI_AlltoAll` collective communication primitives that are implemented using multicast. These operations differ in the way data is distributed among the involved processes.

In particular, `MPI_AllGather` is an operation where every process is sending the same data to every other process, and receiving processes place data in some corresponding places in the receiving buffer depending on the sender process. Figure 1.1 illustrates the data transfer of three processes for the `MPI_AllGather` using three data elements per process. Each process performs a multicast operation to distribute data to the other processes. This operation ends with each process having the exact same data in its receive buffer, and each process contributes a single value to the overall array.

This operation is extensively used in HPC workloads like Matrix multiplication, LU factorization, and Linear algebra operations [3]. In addition, `MPI_AllGather` collective operations are also found key to achieve higher performance on Big Data applications [6]. A 33% performance improvement can be achieved using this collective operation for the K-means clustering application.

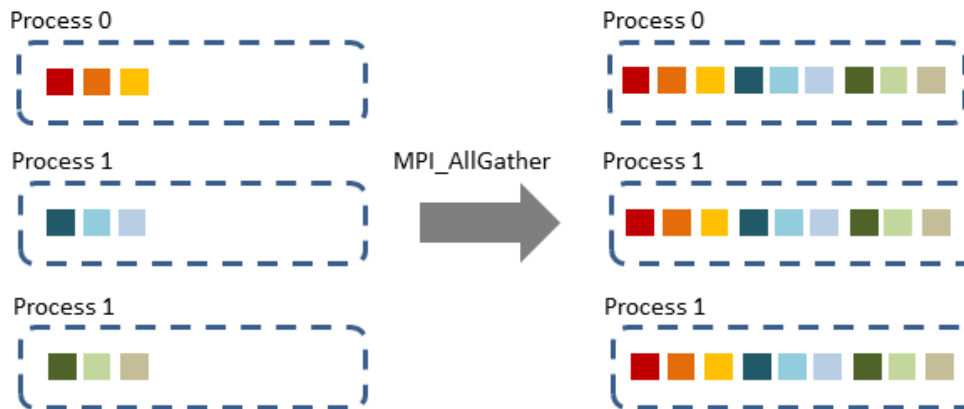


Figure 1.1: MPI_AllGather collective operation

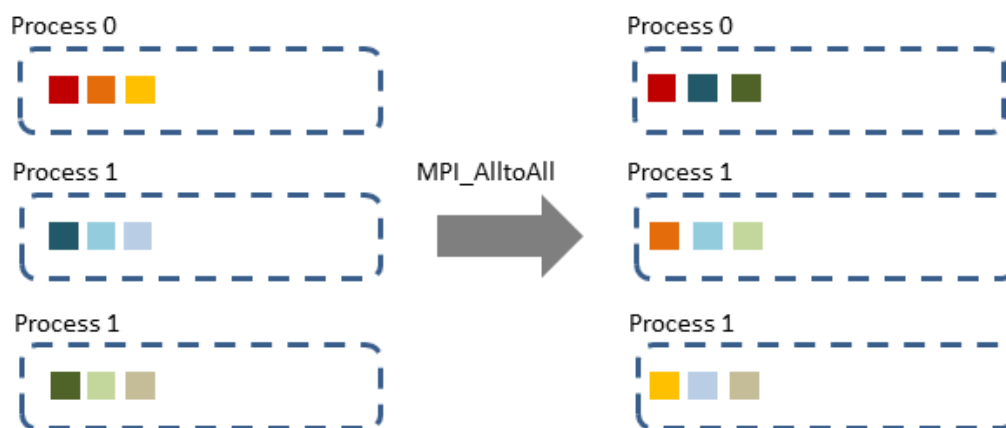


Figure 1.2: MPI_AlltoAll collective operation

MPI_AlltoAll is a different collective operation from the previous one. This operation does not send the same values to each other process as in the previous one. Instead of providing a single value that should be shared with each other process, each process specifies one value to give to each other process. In other words, with n processes, each must specify n values to share. Then, for each processor j , its k 'th value will be sent to process k 's j 'th index in the receive buffer. This is useful if each process has a single, unique message for each other process.

MPI_AlltoAll is extensively used to perform Fast Fourier Transforms (FFT). HPC workloads usually require this operation for signal processing such as digital filtering. For example, a good use case of this operation is weather HPC applications where data often contain two distinct cycles, diurnal (daily) and annual (yearly), and one might want to remove one to study the other in isolation. Also, economical applications are using FFT to remove unwanted periodicities and reveal secular trends.

In these applications MPI_AlltoAll is the most dominant operation in order to achieve higher performance. In particular, for P3DFFT application [5] the MPI_AlltoAll accounts for two thirds of the execution time on InfiniBand networks [4].

These collective operations are dominated by different factors depending on the size of the data transferred. When the data transferred is small the cost of these operations is dominated by the software overhead of sending the messages. This could be optimized by combining multiple small

packets into a single big one. On the other hand, for large data transfers the cost is dominated by network contention. And network contention can be minimized by using a flattened network topology like the one proposed in LIGHTNESS.

2. Experimental assessment of the overall LIGHTNESS architecture

In these last WP5 validation activities, LIGHTNESS experimentally assessed a fully SDN-programmable intra data centre network (DCN) architecture including network function virtualization (NFV) capabilities for effective network control. It is the first demonstration of NFV and SDN functionalities, such as monitoring and database migration, entirely integrated with an advanced all-optical physical layer. Our data plane is capable of performing OCS/-to-OPS switch-over and vice versa on-demand, providing intra-DCN connectivity with low deterministic latency and variable bandwidth granularity. Multicast enabled and hybrid optical VDCs are provisioned in the LIGHTNESS system leveraging on control functions exposed by the SDN controller, while the OCS/OPS and multicast/unicast switch-over is performed on, according to performance metrics collected by a monitoring NFV application.

2.1. Overall inter-cluster architecture

The overall DCN architecture is presented in Figure 2.1. Looking the DCN design with a bottom-up approach, servers within the same racks are interconnected via a novel optical Network Interface Card (NIC) and an all-optical top-of-the-rack (ToR) switch to the programmable DCN [7].

The FPGA-based hybrid NICs [8] employ SDN-enabled hybrid OCS/OPS interfaces that support programmable OPS/OCS composition and transmission of optical packets with associated labels or Ethernet frames. It eliminates the electronic ToR switch and interconnects the intra-rack servers with ultra-low latency. The functionality of such novel NIC includes network interface functions, programmable aggregation and segregation functions, OPS/OCS switch and layer 2 switch functions. Thus, according to the specific service requirements, the FPGA-based OPS/OCS NIC can be provisioned by an SDN controller on demand.

A large port-count space switch is used as the all-optical ToR switch, which connects all the NICs in the same rack and provides connectivity between them and the access outside of the rack. This scheme offers lower interconnection latency and potential reduction in power consumption of the overall network, due to absence of O/E/O conversions. It also supports full bandwidth transparency, since optical switches are totally agnostic of the link bitrate, the network protocol or the modulation format that is being used.

The programmable DCN is configured by an optical programmable system deploying Architecture-on-Demand (AoD) concept [9]. The AoD configuration utilizes Polatis beam-steering fibre switches [9] as the optical backplane, which connects various optical sub-systems and all the ToRs.

Optical Packet Switch (OPS) nodes, with multicasting and optical packet reception/contention monitoring capabilities, and all the ToR switches of a single cluster are connected to the high-radix optical backplane (implemented by a Polatis switch) with OCS/OPS multicasting capability. The back plane hosts optical function block plug-ins (i.e. wavelength selective switches (WSS), EDFAs, OPS, splitters for OCS multiplexing) and offers ToR-to-ToR connectivity.

With all clusters utilizing the same intra-cluster infrastructure, another Polatis switch is used as the interconnection between clusters. Each cluster is connected to the inter-cluster fibre switch with multiple fibres or a SDM fibre [10].

With such architecture, different network configurations can be obtained by setting appropriate cross-connections between inputs/outputs and modules in the optical backplane. Thus, synthetic node architectures can be dynamically provisioned involving only the required transmission and functionality.

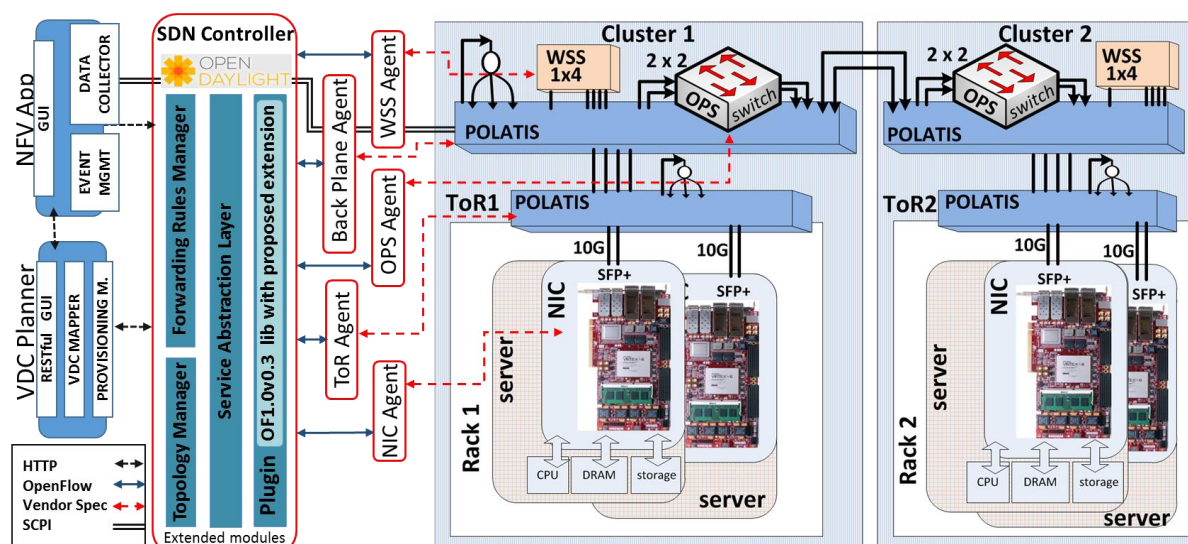


Figure 2.1: The overall architecture of LIGHTNESS DCN solution

On top of this inter-cluster full optical DCN, an SDN controller is deployed to configure and reconfigure the physical layer topology, by dynamically provisioning appropriate cross-connections in the optical backplane according to different applications' requirements. In addition, according to the specific service requirements, the SDN controller is in charge of dynamically provisioning all the optical devices deployed in the DCN, the FPGA-based OCS/OPS, the optical ToRs, inter-cluster switch and OPS switch, in support of unicast and/or multicast communication patterns between servers. The interactions between the SDN controller and the optical devices are mediated by dedicated control agent entities (one per each device) that allow to have an OpenFlow enabled DCN while keeping the vendor specific control interface exposed by each device. Two applications have been also developed for the purpose of these final experimental validation activities. First, a VDC composition application running on top of the SDN controller, able to provision multicast and unicast virtual network slices in the LIGHTNESS DCN leveraging on control primitives offered by the SDN controller. Second, a monitoring NFV application, that is a virtual network function (VNF) able to collect OPS statistics and Polatis port power status and trigger on-demand the switch-over of provisioned services.

2.2. All-optical reconfigurable data plane testbed setup

The all-optical testbed setup for these experimental activities is illustrated in Figure 2.2. Four rack-mounted PowerEdge T630 servers are equipped with an FPGA-based NIC board utilizing 10G SFP+ transceivers, providing the interface to the optical network. The FPGA-based OPS/OCS hybrid NIC

using NETFPGA SUME development board, has been designed to plug directly into a server, and replace the traditional NIC. In the prototype design, it has an 8-lane Gen3 PCIe interface for DRAM communication, one 10Gbps interface for getting commands from SDN control OpenFlow agent and sending feedback, two OPS/OCs hybrid 10Gbps ports for inter-rack/cluster communication and an OPS label pin interface connecting to the OPS label generator. The SFP+ transceivers' channels are in the 1550nm region and ITU grid-spaced, in order to be compatible with the LCoS-based WSS and SOA-based OPS switches, which both normally operate in that frequency band.

All servers are connected to a 192×192 port Polatis circuit switch, which acts as a ToR switch and as an optical backplane on top of each cluster. A 1×4 optical power splitter, two 1×4 Wavelength Selective Switches (WSS) and one SOA-based 4×4 OPS [11], which can logically perform as two 2×2 switches, are attached to that optical backplane.

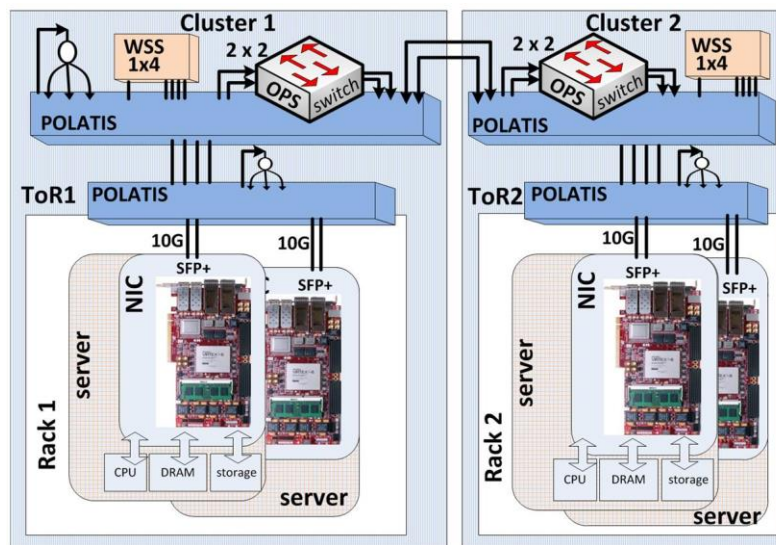


Figure 2.2: Experimental setup for intra-/inter-cluster communication

The OPS is based on modular WDM architecture that allows scalability of the number of ports beyond 128×128 while the highly distributed control and parallel packet processing allows port count independent 20ns reconfiguration time. A fully equipped 4×4 prototype including optical label processing, optical switching fabric and controller has been realized in the LIGHTNESS framework [14]. The modular architecture allows the 4×4 OPS prototype to logically perform as two 2×2 OPS. The electrical label bits generated by each NIC, are encoded in an in-band optical RF tone label by a prototyped label generator. The in-band optical labels are then coupled to each of the optical packets. At the OPS node, the optical label of each packet is filtered out, processed and matched with the look-up-table by the switch controller in order to determine the packets destination.

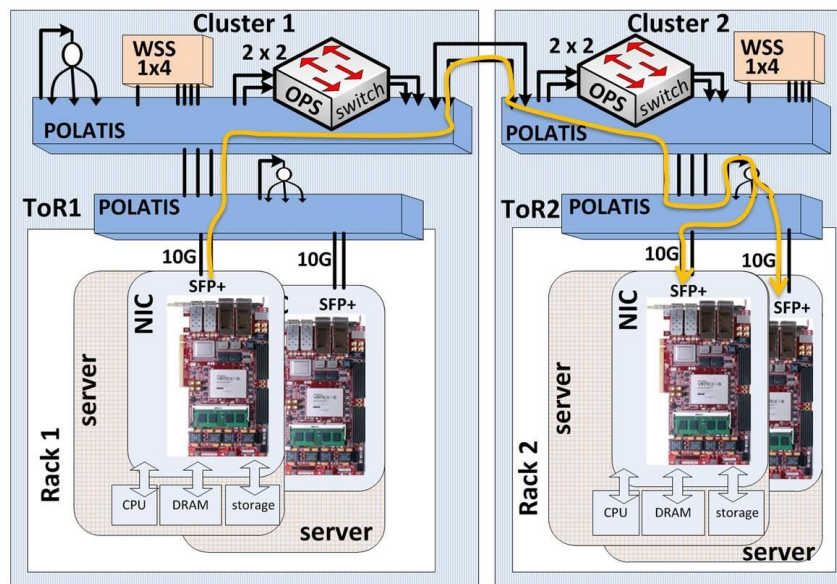


Figure 2.3 Multicasting with OCS scheme

The 1×4 optical power splitter is used to accomplish OCS one-to-four multicasting scenarios. The WSSs are used for grooming inter-cluster traffic carried by channels from different servers or racks into an inter-cluster WDM super-channel. In the destination cluster, the local WSS de-multiplexes the super-channel and switches the channels to the receiving servers. As illustrated in Figure 2.3, traffic from Rack 1 in Cluster 1 can also be sent to Rack 2 in Cluster 2 through a splitter via OCS links.

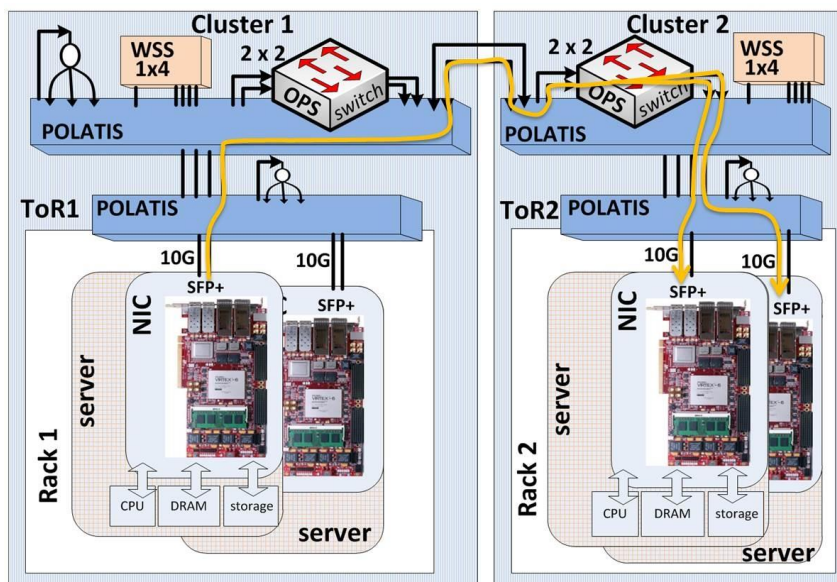


Figure 2.4 Multicasting with OPS scheme

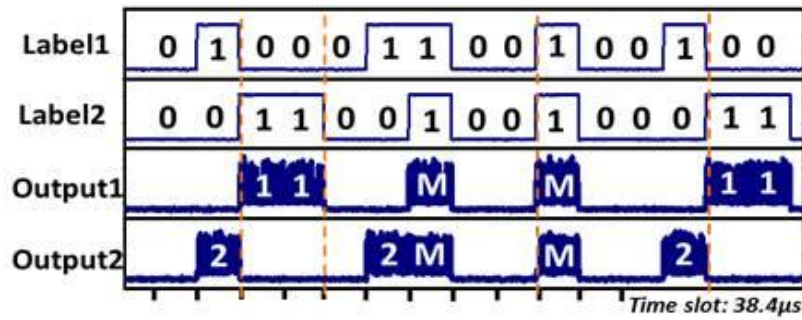


Figure 2.5 Time traces for OPS label generation, where “M” represents multicast communication is enabled.

For OPS multicasting scenario, as shown in Figure 2.4, traffic from Rack 1 in Cluster 1 is sent to Rack 2 in Cluster 2 through the WSS (if WDM channels are enabled) and the OPS module via multi-stage Polatis links. The WSS is adopted for grooming inter-cluster channels from different servers/rack to the destination cluster while band switching is enabled for OPS. The OPS switch is able to rapidly switch optical packets with a reconfiguration time of 20ns. The label bits are generated by each NIC, then an optical RF tone label is created by the label generator and attached to each optical packet. At the OPS, the label is extracted processed and matched with the look-up table by the switch controller to determine the packets destination. Multicasting is enabled when two label bits have been set as “11” as shown in the time traces of Figure 2.5.

2.3. All-optical reconfigurable control plane testbed setup

The LIGHTNESS enhanced OpenDaylight (ODL) is used as the SDN controller, and OpenFlow (OF) agents for Polatis, WSS, OPS switch and FPGA-based hybrid OCS/OPS NIC were also developed in WP4 to enable further SDN-based programmability [12], as shown in Figure 2.1 left. The agents interact with the underlying devices through different interfaces (i.e., TL1, Raw socket and Ethernet frame) to perform a set of actions such as capabilities/attributes collection, configurations, and monitoring. The OF protocol is significantly extended to enable the communications between the ODL controller and the optical data plane via the OF agents, as specified in deliverable D4.2 [13]. Figure 2.6 shows the implemented OF extensions in the fields of ofp_capability, ofp_match and ofp_action. The ofp_match is extended for OPS to support the use of labels, while the ofp_action is extended for FPGA-based optical NIC to enable the optical packet label configurations.

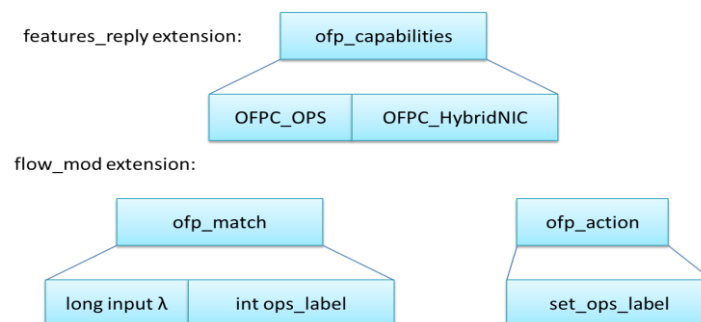


Figure 2.6: OpenFlow extensions

Also, some of the ODL internal software modules (i.e. the ones included in the OpenDaylight box in Figure 2.1) are extended to support the LIGHTNESS optical network device specific features. For example, regarding WSS ports, the Switch Manager and Service Abstraction Layer (SAL) were extended to enable OpenDaylight record the supported wavelength and supported spectrum range respectively, both of which are used to validate its provisioning. Furthermore, the OPS optical packet statistics can be collected and maintained by the Statistics Manager. Last but not least, in order to properly configure the above optical devices, the Forwarding Rules Manager has been extended, among the others, to construct the required set of configuration information e.g. label & output for the OPS switch; central frequency, bandwidth & output for the WSS and match, label and output for the NIC (which is optional for OPS).

The FPGA-based optical NIC communicates with the OF agent through a bidirectional 10Gbps SFP+ Ethernet interface. The commands and information are encapsulated in a 1504 Byte Ethernet Frame (VLAN).

Furthermore, through the LIGHTNESS enhanced ODL, various applications can be deployed on top of the controller to provide enhanced functionalities and services leveraging on the RESTful interfaces exposed at the northbound side by ODL.

2.4. Multicast Virtual Data Centre composition and monitoring

For the purposes of the experimental test, two applications have been deployed on top of ODL: a Virtual Data Centre (VDC) composition, that is a further enhanced version of the one released in WP4 with [12], and a virtual network monitoring function (monitoring VNF), that is a new service developed within WP5.

The VDC composition aims to create and provision virtual network slices within the DCN as a multi-tenancy application and consists of a graphical user interface (GUI) developed in HTML/JavaScript that interfaces with a backend application developed in Python 2.7 able to interact with the ODL controller. The user can access to the GUI with any existing browser, and create dynamically a VDC request, which is shown in a graph and a table, as depicted in Figure 2.7. The parameters that the user can specify are: virtual machines (VMs) to be used in the virtual slice, virtual links to be created to connect those VMs, technology for each link (OPS vs. OCS) and multicast properties for VMs and virtual links. In particular, these multicast/unicast functionalities have been added in this new version of the VDC composition application and algorithms, that is now able to compute virtual slices taking into account these further constraints in support of optical multicast connectivity. Other parameters that are available include: required bandwidth in Mbps and bi-directionality of a given link. The application receives a set of requirements for the VDC and generates a bunch of flows to be pushed in the DCN by ODL, distributed among the different technologies (NIC cards, OPS Switches and OCS backplane). This set of flows are generated in JavaScript object notation (JSON) format and sent to the ODL controller through RESTful interface. The application shows a popup confirming the status of the request.

When a VDC has been deployed (in this experimental work, a multicast VDC using OPS resources), the user can request for a dynamic VNF to monitor various parameters on his virtual data centre. The

request is a basic HTTP GET request, made through a standard web browser, which is handled by one of the servers creating the monitoring virtual function (that is basically launch and properly configure a VM running the monitoring VNF) and redirecting the users web browser using a standard HTTP redirection response (301 Moved Permanently message, with 'Location' header parameter). The user's monitoring VNF starts retrieving network information by means of two different interfaces. RESTful northbound of ODL controller to gather information of the OPS packet counting and generate a graph of the OPS packet rate on the one end. Standard commands for programmable instruments (SCPI) to instruct Polatis switch to retrieve information of the multicast ports used in the test on the other end (output from optical splitter and the OPS switch). This information is plotted in a web interface (Figure 2.7), showing two graphs illustrating the optical power received in two ports (in dBm), and the packet rate in the OPS switch, with a theoretical maximum and a configurable threshold.

Two buttons in the bottom part of the monitor allow the user to dynamically trigger two different actions and switch-over functions:

- i. switch the multicast traffic to a second OPS switch by making a backup copy of the content between two servers inside the same rack when the optical power detected drops below an expected value
- ii. switch the multicast traffic using the optical splitter (OCS multicasting) whenever the OPS packet rate exceeds the threshold, meaning not sufficient OPS resources to cope with the desired VDC service.

Up to this point, the user has deployed his own VDC with multicasting capabilities through OPS switch. At the same time, the user has requested his own virtual monitor within his VDC, getting information with different polling timers (one second for power monitoring, twenty for OPS received packets to avoid time mismatches among OPS agent/controller/monitoring function). The monitoring VNF allows the VDC user to take two different recovery choices based on the monitored data:

- i. The user, whose service is experiencing high data loss caused by unexpected low power problems, decides that the packet rate within the OPS switch is sufficient to cope with his service. The OPS switch in the VDC should be avoided to reduce or remove packet losses, so the user start a recovery workflow to switch the service to other intra-rack server switching the previous VDC, to a new virtual network using a second OPS switch.
- ii. Based on a high packet per second rate shown by the monitoring VNF, the user decides that, before experience any loss within his VDC, a reconfiguration of the VDC to use a pure OCS network is required. The user can use both the VDC planner and the monitor function (which are connected), to reconfigure the network by moving the content of the first server to a backup one in the same rack and establishing the new set of connections through the optical splitter to maintain the multicast capabilities.

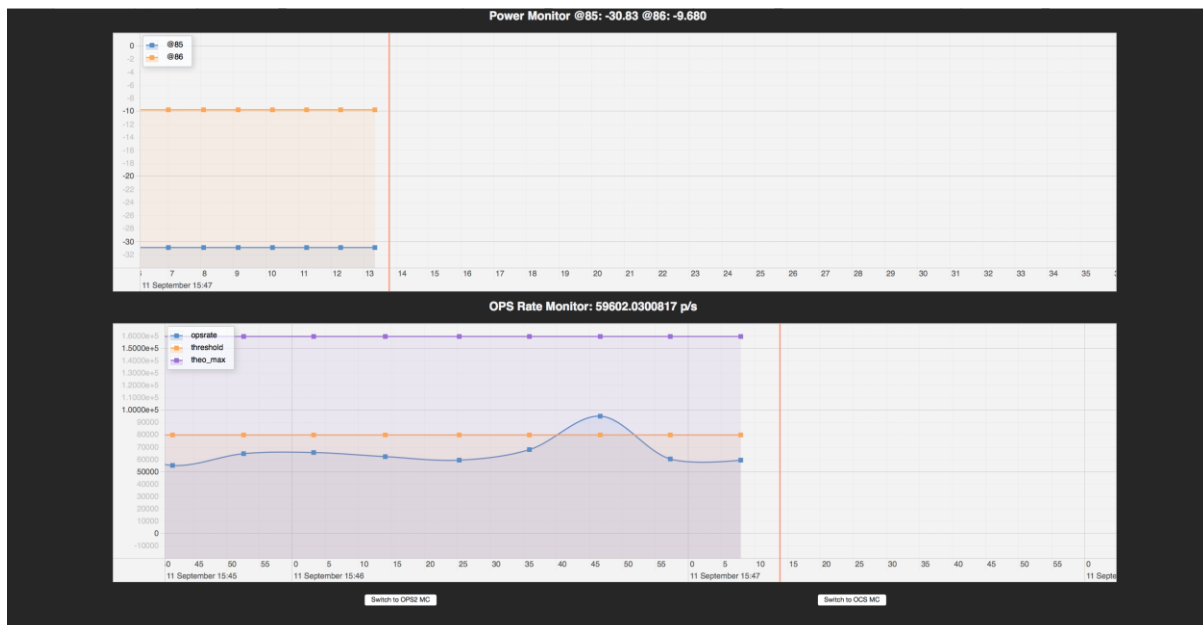
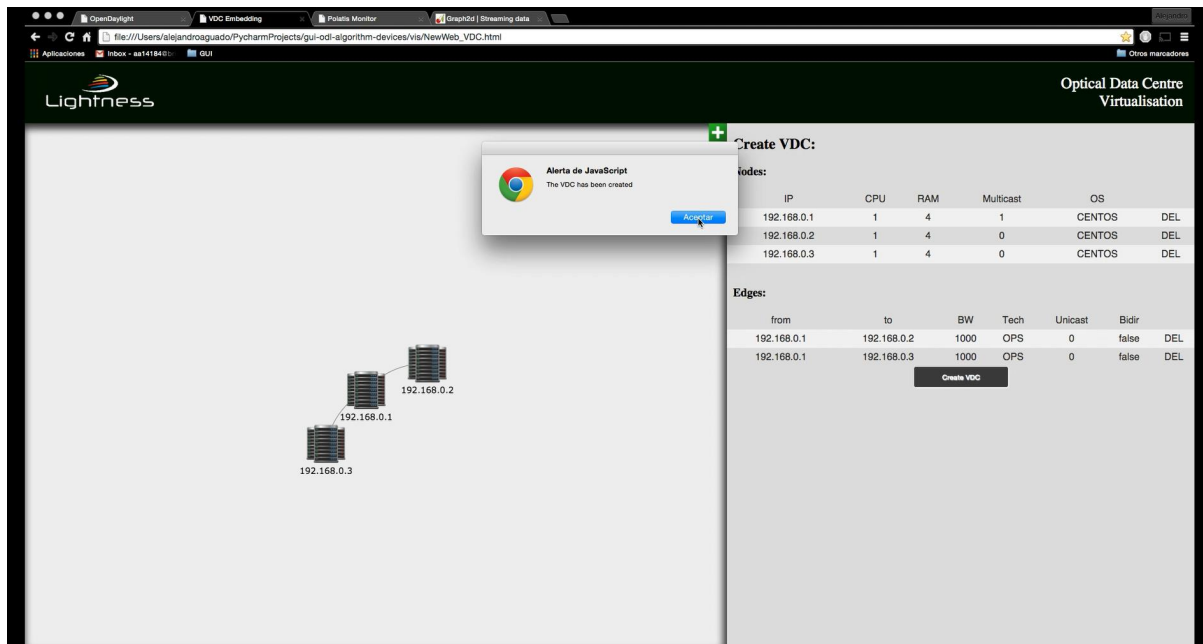


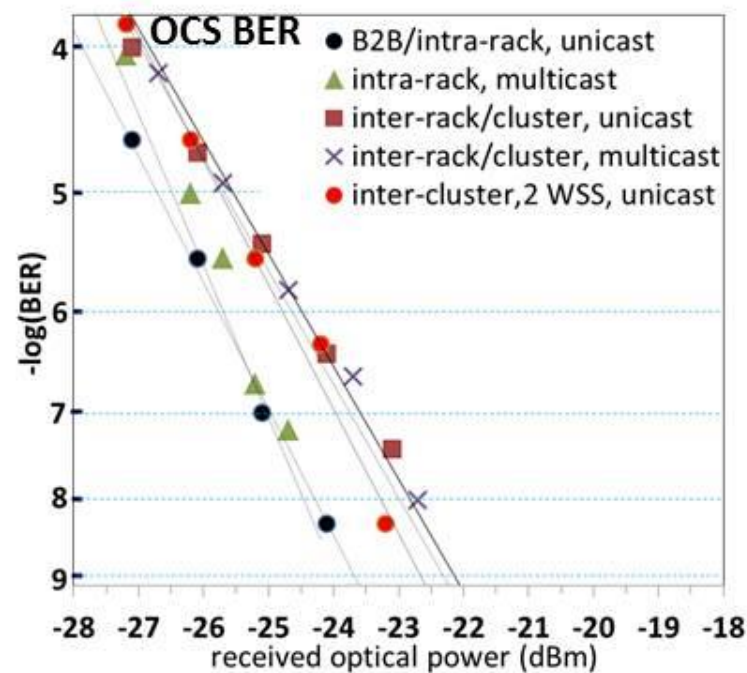
Figure 2.7: VDC composition GUI (top), Power Monitor and OPS Rate Monitor (bottom)

2.5. Experimental demonstration and results

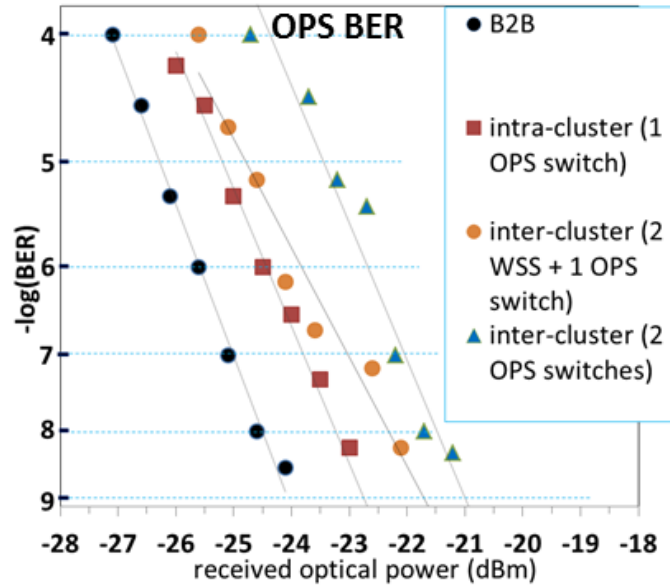
For this final experimental assessment, we combined all our available data plane and control plane resources, as presented in sections 2.2 and 2.3, and validated several intra-DCN interconnection scenarios based on VDC applications' and NFV functions' requests.

First of all, we evaluated the physical layer of the DCN for intra-rack, inter-rack and inter-cluster unicast and multicast communication by measuring BER for both OCS and OPS switching technologies using real traffic with scrambled PRBS payload generated by a traffic analyzer, as shown in Figure 2.8. The traffic analyzer feeds the FPGA-based NIC with the Ethernet traffic, and then NIC pushes the data

to one of its hybrid OCS/OPS ports. When OPS mode is chosen, the NIC, depending on the configuration it receives from the SDN controller, sets up the optical packet duration, encapsulates certain number of Ethernet frames and releases the optical packet while the label is generated and combined in parallel. Intra-rack communication is realized going from transmitting to receiving server through the optical ToR for unicast, and through an optical splitter in multicast operation. Inter-rack and inter-cluster are similarly realized by going through multiple Polatis OXC and/or optical power splitters and OPS switches, while for inter-cluster multiplexed interconnection signals propagate additionally through two WSS for WDM mux/demux and switching purposes. Minor penalties of <2 dB are observed for all OCS interconnection scenarios, as seen in BER curves of Figure 2.8(a). OPS BER plots in Figure 2.8(b) show <1 dB and <3dB penalties when passing through one (for intra-cluster) and two (for inter-cluster) switches, respectively.



(a)



(b)

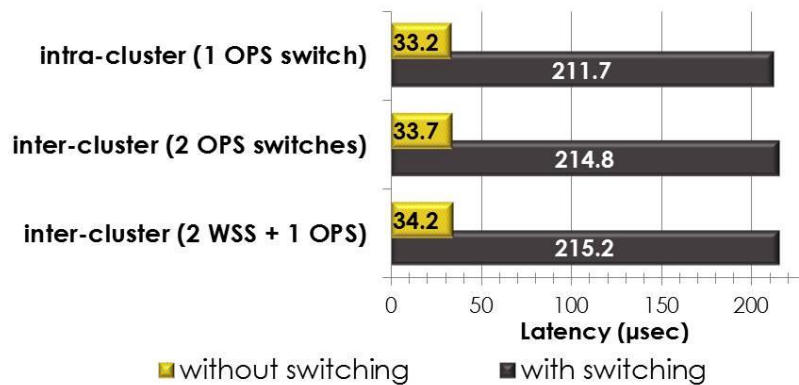
Figure 2.8: BER measurement for intra-rack, intra/inter-cluster scenarios with OCS/OPS schemes

In addition to BER testing of the physical links, we collected network Layer 2 results regarding the interconnection latency from one NIC's DMA to the destination NIC's DMA, excluding the DMA driver's actual delays, which is separately measured for different DMA lengths. Moreover, interconnection throughput is monitored and plotted, exhibiting OCS-to-OPS switch-over and vice versa.

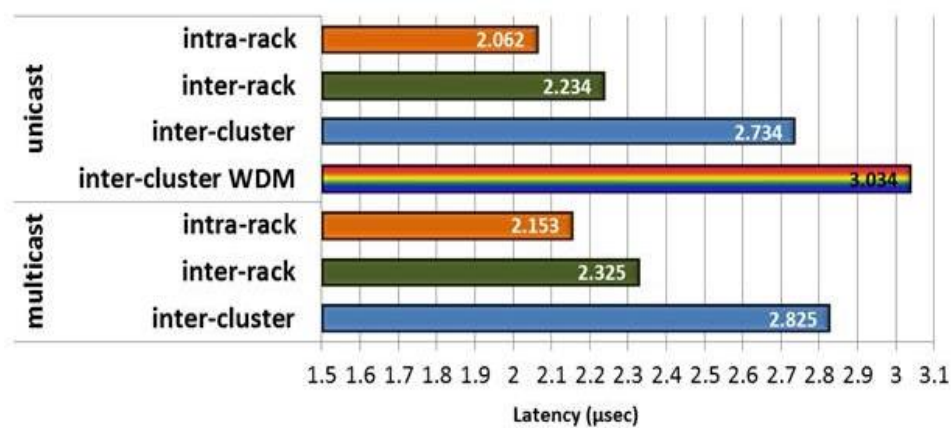
We measured DMA-to-DMA access latency again using the traffic analyzer and Ethernet traffic with PRBS payload. The test and measurements are based on the best possible latency with maximum bitrate, which are around 3 Gb/s for OPS switching, and around 8 Gb/s for OCS switching, respectively. All measured latency values include FPGA physical and logic delays, which can vary depending on the frame length, chosen transmission/switching scheme (OCS or OPS) and FPGA design. Figure 2.9 (a) shows unicast and multicast OCS access latencies for all the studied interconnection scenarios; whereas Figure 2.9 (b) shows intra/inter-cluster OPS access latencies with and without switching.

Results indicate different values of latency for OPS with and without switching. When no switching is performed, data does not need fragmentation and it is transmitted as it is. For OPS with switching though, optical packets are formed, a procedure which inserts significant delays due to segregation, aggregation, buffering and clock recovery. For instance, when transmitting/receiving 3 packets, the latency is tripled, $33.2 \mu\text{sec} \times 3$ equal to $99.6 \mu\text{sec}$. In the FPGA, the first buffer of segregation-aggregation part uses numerous FIFOs and store-and-forward techniques, which means it takes $25.6 \mu\text{sec}$ more time when OPS with switching mode is chosen. The same happens at the last buffer of aggregation inside the FPGA design, so another $25.6 \mu\text{sec}$ are added. Since we use two FPGAs (one Tx plus one Rx), the whole segregation-aggregation procedure takes place twice. Thus, in total, the delay is doubled, adding further $102.4 \mu\text{sec}$. Finally, roughly $99.6 + 102.4 \mu\text{sec}$ equal to $202 \mu\text{sec}$, as shown in Fig. 6 with switching. The receiver of the FPGA-based NIC needs to recover the clock from the receiving traffic if this was lost during the operation. The time of this recovering depends mostly on the network conditions and the signal performance. In the case of OPS switching, it needs $25.6 \mu\text{sec}$

to recover the clock before each packet. This time can be drastically reduced to <150 ns by using dedicated ASIC clock and data recovery [15].



(a)



(b)

Figure 2.9: DMA-to-DMA latency for (a) OCS scheme; (b) OPS scheme

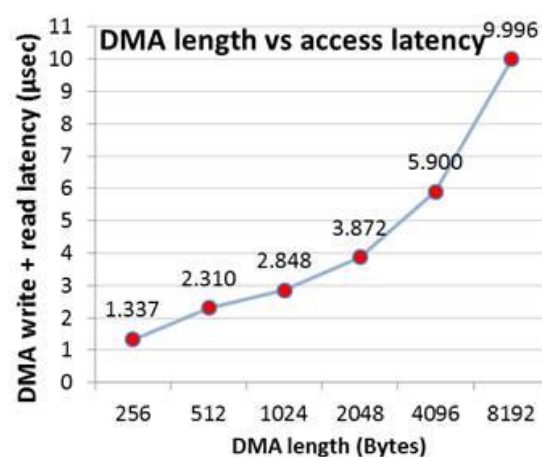


Figure 2.10: DMA write and ready latency VS DMA length measured with 256bytes Ethernet packet frame length

Finally, DMA driver's actual access latency was also measured, as shown in Figure 2.10, for different DMA lengths ranging from 256 to 8192 Bytes. It is the time interval we measured when we were

3. LIGHTNESS system demonstration

The LIGHTNESS team participated to the ECOC2015 exhibition, in Valencia (Spain), and successfully demonstrated the full LIGHTNESS data centre system by presenting on site all the outcomes of the project, from the optical data plane devices and prototypes (NIC and OPS) developed in WP3 and the SDN control plane with its applications on top implemented in WP4. ECOC is the largest conference on optical communication in Europe and one of the most prestigious and long-standing events in this field worldwide. ECOC stands for presentation of current scientific work as well as for major innovation and latest developments in optical devices in present and future networks. Every year, the conference is co-located with an exhibition event dedicated to the optic communication industry where hundreds of exhibitors presents new products, equipment and services. ECOC2015, the 41st edition of the conference, has been held in Valencia, Spain, from 28th of September to 1st of October 2015.

The LIGHTNESS demonstration presented the “SDN-enabled and Programmable Optical Data Centre with OCS/OPS Multicast/Unicast Switch-over” and was performed within a dedicated LIGHTNESS booth at ECOC2015 for the whole three days of the exhibition. A wide audience attended the LIGHTNESS booth and experienced the innovative full optical SDN enabled data centre network, where the optical testbed developed in the project was fully integrated on site with the SDN control plane enhanced with Virtual Data Centre composition and a monitoring virtual network function (VNF) applications. The demonstration was successfully performed to many people from telco industry (operators and vendors), academia and European Commission; all of them demonstrated high interest on the technologies and the approaches proposed and implemented in the project providing very good feedback on the overall achievements.

The scope of the demonstration was twofold: first, show the programmable transport, switching and OpenFlow configuration of data flows over the LIGHTNESS hybrid optical flat data centre network, with a full on-site integration of the LIGHTNESS extended OpenDaylight SDN controller and optical data plane, employing an OPS switch, an Architecture on Demand (AoD) OCS switch, programmable optical NICs and optical ToRs. Second, on top of the OpenFlow based provisioning features, the demonstration showed applications for: i) on-demand VDC provisioning and reconfiguration, with creation of multicast virtual slices using OPS resources, ii) monitoring Virtual Network Function (as an NFV application) to retrieve OPS statistics and OCS port power status, and automated OCS/OPS and multicast/unicast switch-over.

The testbed deployed for the demonstration at ECOC2015 is shown in Figure 3.1. It basically models the proposed LIGHTNESS full optical data centre network controlled and operated by the extended OpenDaylight controller, reflecting the configuration for the experimental assessments described in section 3. The testbed was composed by a set of hardware and software components.

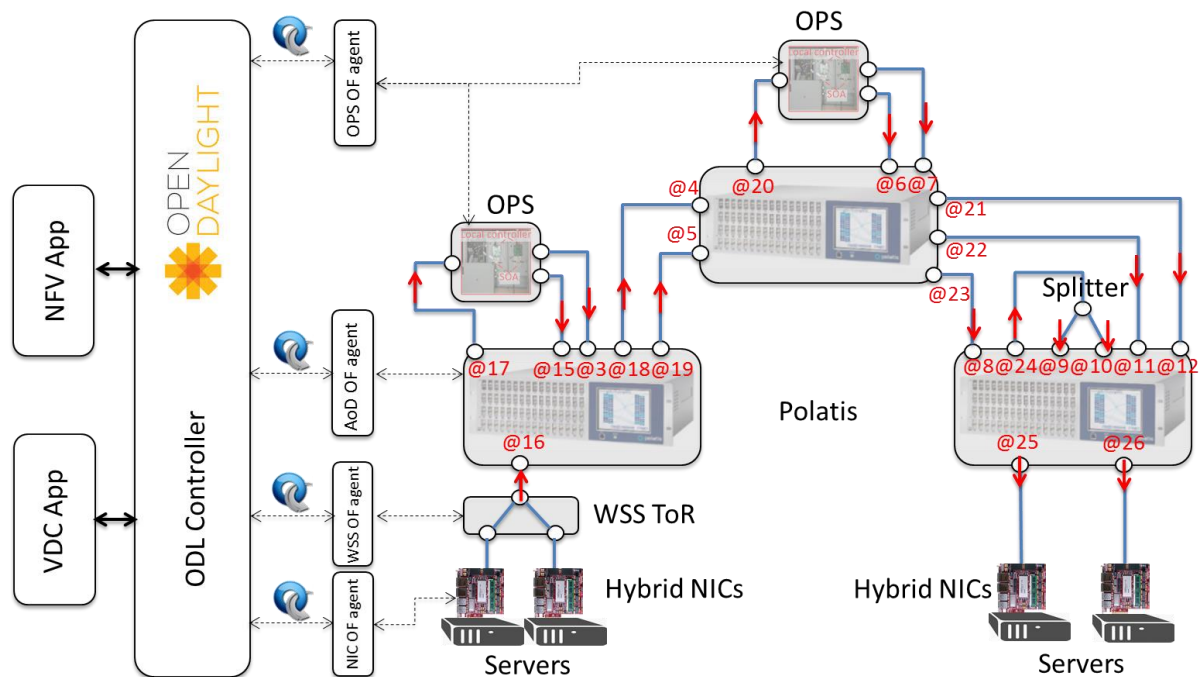


Figure 3.1 LIGHTNESS demonstration testbed at ECOC2015

At the hardware level, the testbed for this demonstration was built by the following components (also depicted in Figure 3.2):

- 4 Dell PowerEdge T630 servers
- 4 FPGA-based NIC boards utilizing 10G SFP+ transceivers, one for each server
- 1 4x4 OPS switch, logically split into 2 2x2 switches
- 1 Finisar 4000 1x4 ToR WSS switch
- 1 Polatis switch, logically split into 3 OCS/ToR optical switches
- 1 1x4 optical splitter

At the software side, the demonstration deployment included:

- 1 OpenDaylight SDN controller, running on site within a Virtual Machine
- 1 OPS OpenFlow agent, running on site in a dedicated laptop
- 4 NIC OpenFlow agents, running on site within the local servers
- 1 Polatis OpenFlow agent, running on site with a Virtual Machine
- 1 VDC composition application, running on site a Virtual Machine
- 1 monitoring VNF, running on site as a Virtual Machine to monitor the OPS switch counters for packet losses and the Polatis ports power

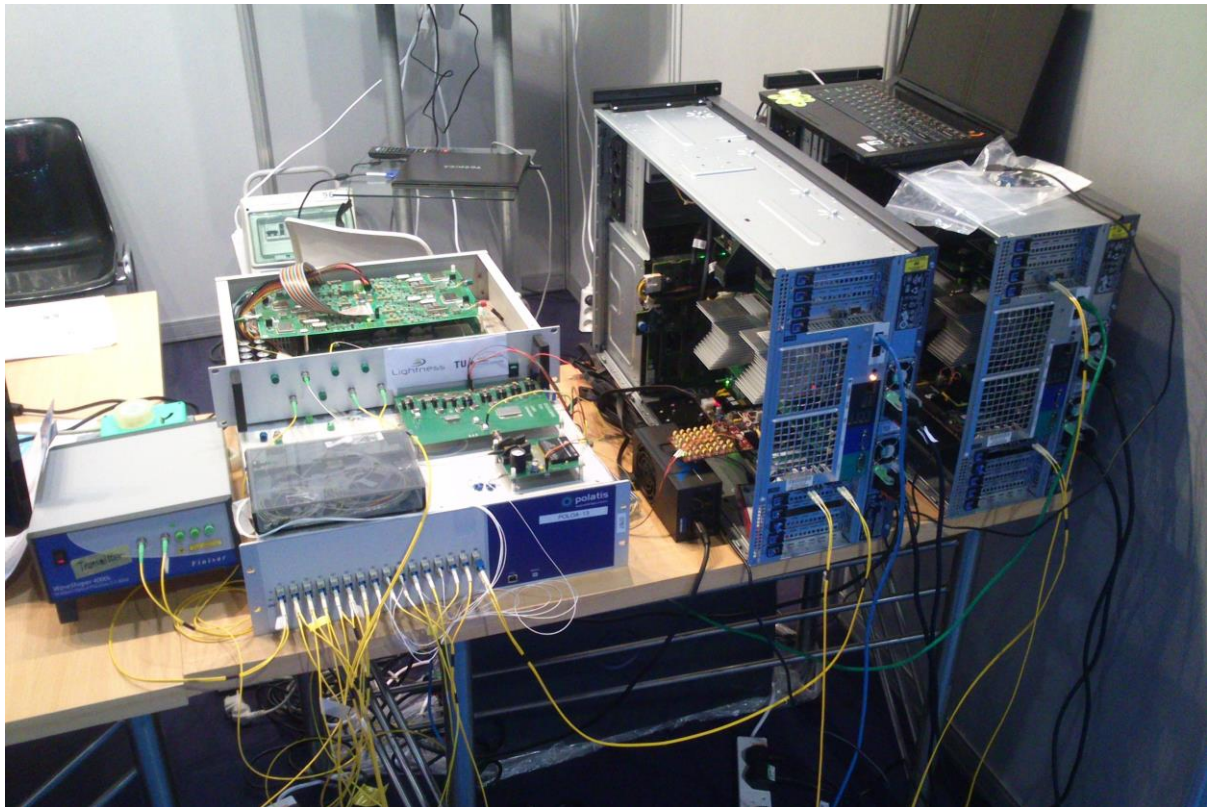
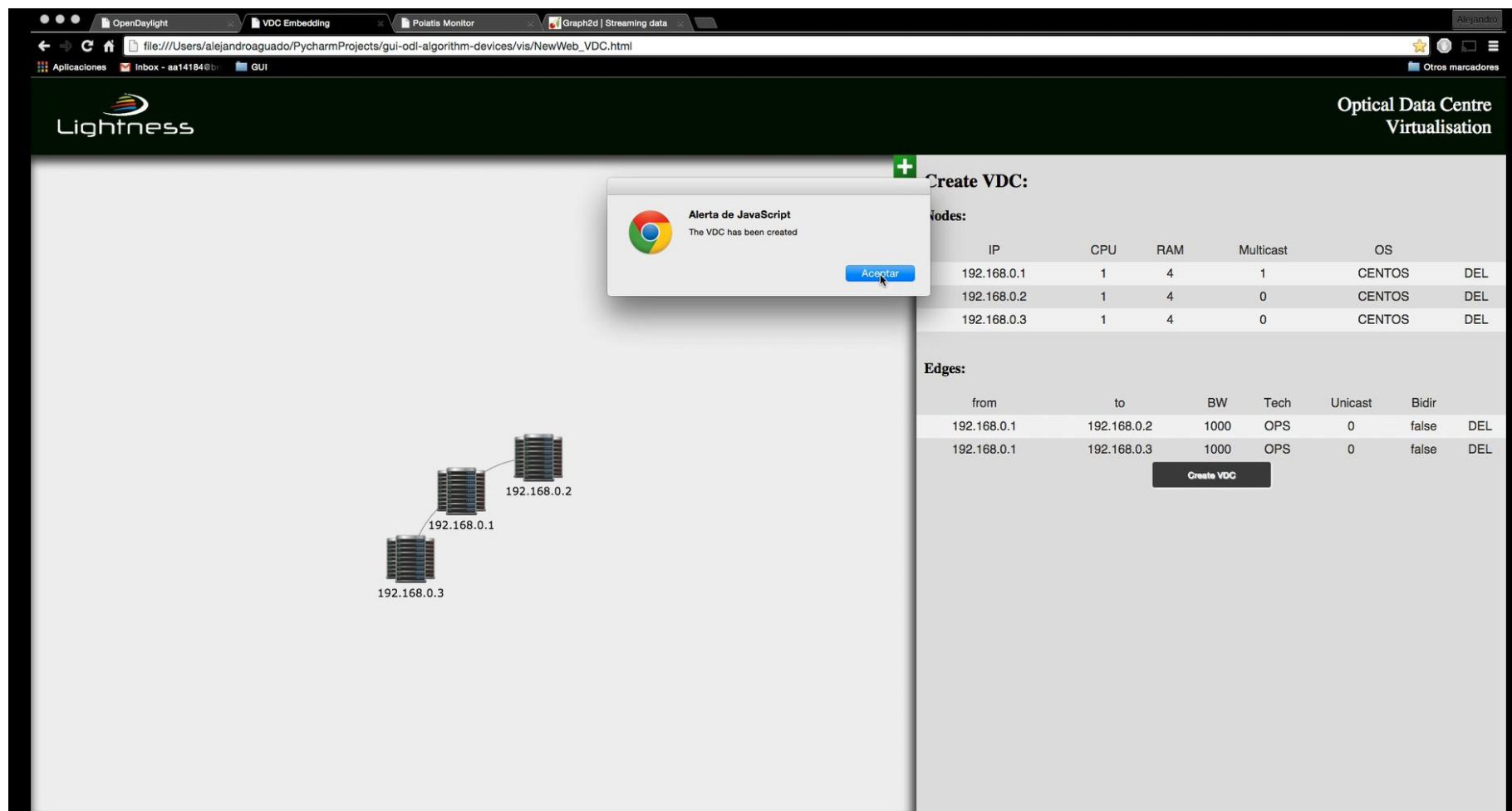


Figure 3.2 Hardware installed at ECOC2015 LIGHTNESS booth

The LIGHTNESS demonstration has been performed through the Graphical User Interface (GUI) exposed by the VDC composition application, that allows to trigger the creation of multicast OPS or OCS virtual slices leveraging on the control functions provided by the OpenDaylight controller. As shown in Figure 3.3, to create a new VDC, a set of parameters for virtual nodes (i.e. Virtual Machines belonging to the virtual slice) and virtual links (i.e. edges in the picture) has to be provided through the GUI, including: IP address of virtual node, source and destination virtual nodes for each virtual link, bandwidth for each virtual link, preferred technology for each virtual link (i.e. OPS vs. OCS), multicast capability of the virtual link. By clicking the “create VDC” button shown in Figure 3.3, the VDC composition application running on top of the controller computes the mapping between the virtual slice specified through the GUI and the optical physical resources available in the data centre network (as exposed by the OpenDaylight SDN controller and shown in Figure 3.4), and triggers the VDC provisioning. This is performed by requesting the OpenDaylight controller, through its northbound APIs, to configure and provision the computed optical resources for the given VDC. Once the VDC is provisioned, the monitoring VNF is automatically deployed to start collecting a set of performance and status metrics of the given VDC. For this demo, the monitoring VNF was able to monitor port power at the Polatis switch (through direct interaction with the switch), and OPS statistics for packet losses (collected through the OpenDaylight northbound APIs). The monitoring VNF GUI, shown in Figure 3.5, was equipped with a couple of buttons to trigger an automated switch-over of the VDC from OCS to OPS or from multicast to unicast.



Nodes:

IP	CPU	RAM	Multicast	OS	
192.168.0.1	1	4	1	CENTOS	DEL
192.168.0.2	1	4	0	CENTOS	DEL
192.168.0.3	1	4	0	CENTOS	DEL

Edges:

from	to	BW	Tech	Unicast	Bidir	
192.168.0.1	192.168.0.2	1000	OPS	0	false	DEL
192.168.0.1	192.168.0.3	1000	OPS	0	false	DEL

Create VDC

Figure 3.3 LIGHTNESS VDC composition application GUI

Demo OpenDaylight VDC Embedding Polaris Monitor Graphviz Streaming data

137.222.204.208:8080/#

Aplicaciones Inbox - aa14184@br GUI Otros marcadores

OPENDaylight admin

Devices Flows Troubleshoot

Nodes Learned

Nodes Learned

Search

Node Name	Node ID	Ports
OPS	OF 00:00:00:00:00:00:02	6
Polatis	OF 00:00:00:00:00:00:02:ff	32
NIC	OF 00:00:00:00:00:00:36:88	3

1-3 of 3 items Page 1 of 1

Static Route Configuration Connection Manager

Static Route Configuration

Add Static Route Remove Static Route

Search

Name	Static Route	Next Hop Address
0 items		

Subnet Gateway Configuration SPAN Port Configuration

Subnet Gateway Configuration

Add Gateway IP Address Remove Gateway IP Address Add Ports

Search

Name	Gateway IP Address/Mask	Ports
default (cannot be modified)	0.0.0.0/0	

1-1 of 1 item Page 1 of 1

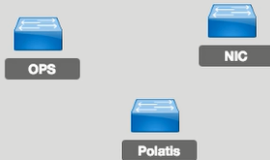


Figure 3.4 LIGHTNESS OpenDaylight GUI



Figure 3.5 LIGHTNESS monitoring VNF application GUI

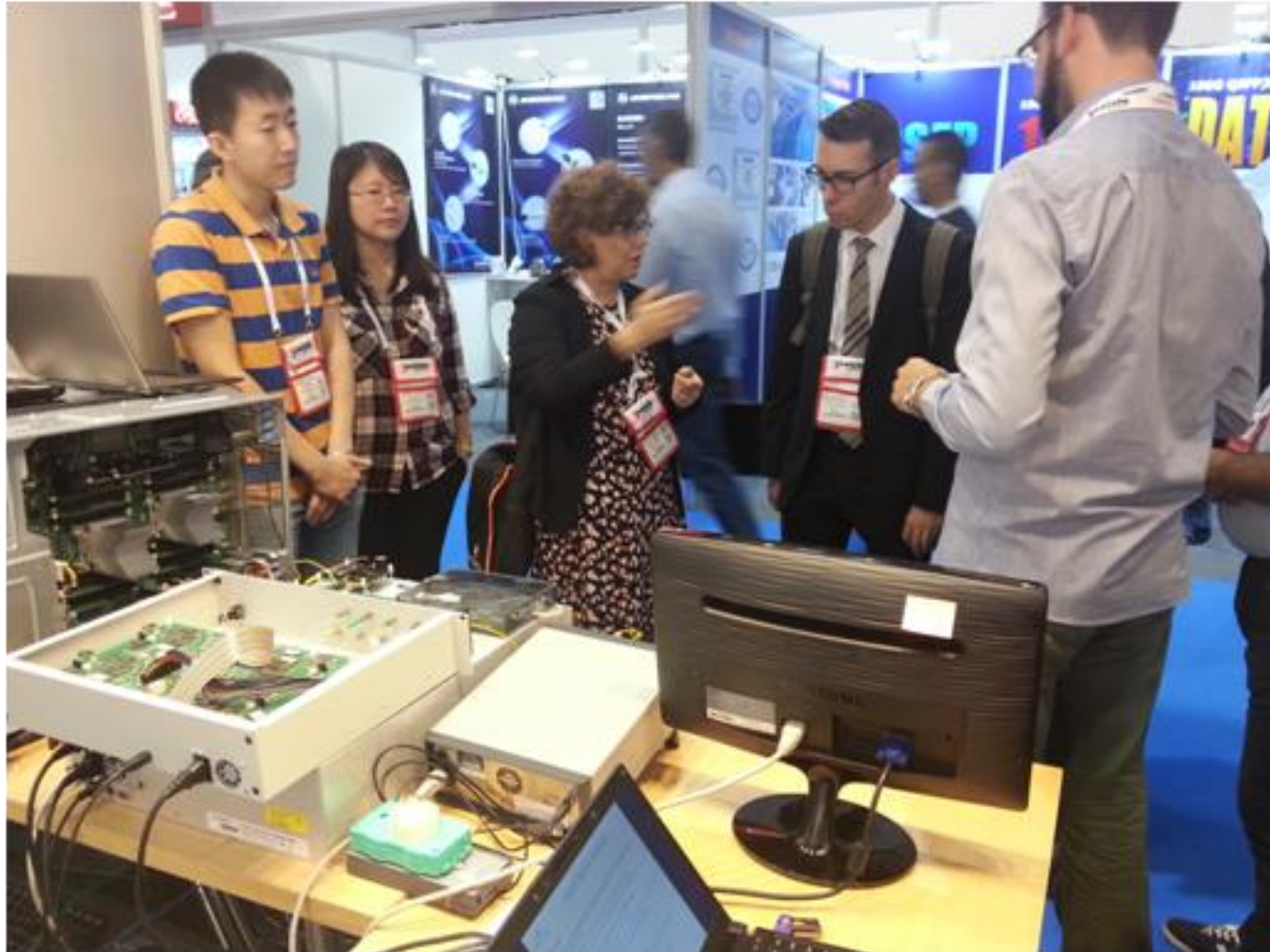


Figure 3.6 LIGHTNESS demonstration is performed to Thibaut Kleiner (EC)

4. Conclusions

This deliverable has reported on the experimental assessment of the LIGHTNESS system in the end-to-end data centre network testbed, where all data plane and control plane components have been integrated and evaluated. After the individual optical devices validation and performance evaluation carried out in WP3 and their individual integration with the SDN control plane carried out in previous WP5 activities, this document has reported the results of the final integration of all hardware and software pieces composing the proposed LIGHTNESS full optical flat data centre architecture. Thus, physical servers equipped with the LIGHTNESS programmable optical NIC have been interconnected following the LIGHTNESS inter-cluster architecture principles by a combination of optical ToR switches, OPS switches and AoD optical backplanes. The control and operation of this end-to-end optical testbed has been delegated to the SDN control plane developed in the project, whose software enhancing the OpenDaylight controller and providing OpenFlow capabilities to the optical devices by means of dedicated agents, has been installed in the testbed. A multicast VDC composition application and a monitoring VNF have been also deployed to validate the on-demand, dynamic provisioning of multicast virtual network slices with automated technology and multicast/unicast switch-over. An extensive set of performance results have been also collected on top of these dynamically provisioned multicast virtual network slices, including measurements of BER, end-to-end latency for layer 2 traffic and services, and throughput during technology switch-over.

Moreover, the demonstration event organized by LIGHTNESS at the ECOC2015 conference has been described in this deliverable, summarizing the technical scope, the on-site testbed and the demonstration workflow. In conclusion, this document as the final outcome of WP5 validation and demonstration activities, has presented the successful assessment and performance evaluation of the whole LIGHTNESS system (as it was developed in these three years project lifetime) against multicast services, that are considered in data centre and HPC environments as crucial services for emerging highly distributed applications.

5. References

- [1] M. Mahalingam, et al, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", IETF RFC7348.
- [2] P. Garg, et al, "NVGRE: Network Virtualization Using Generic Routing Encapsulation", IETF RFC7637.
- [3] Y. Yang and J.Wang. Near-optimal all-to-all broadcast in multidimensional all-port meshes and tori. IEEE Transactions on Parallel and Distributed Systems, 13(2), 2002.
- [4] Akshay Venkatesh, Sreeram Potluri, Raghunath Rajachandrasekar, Miao Luo, Khaled Hamidouche, and Dhabaleswar K. Panda. 2014. High Performance Alltoall and Allgather Designs for InfiniBand MIC Clusters. In Proceedings of the 2014 IEEE 28th International Parallel and Distributed Processing Symposium (IPDPS '14). IEEE Computer Society, Washington, DC, USA
- [5] P3DFFT User <https://code.google.com/p/p3dfft/>
- [6] Gunarathne, T.; Qiu, J.; Gannon, D., "Towards a Collective Layer in the Big Data Stack," in Cluster, Cloud and Grid Computing (CCGrid), 2014 14th IEEE/ACM International Symposium on , vol., no., pp.236-245, 26-29 May 2014
- [7] Shu Y, Zervas G, Yan Y, et al. Programmable optical packet/circuit switched data centre interconnects: traffic modeling and evaluation[C]//Optical Communication (ECOC), 2014 European Conference on. IEEE, 2014: 1-3.
- [8] Y. Yan, Y. Shu, G. M. Saridis, B. R. Rofoee, G. Zervas, D. Simeonidou, "FPGA-based Optical Programmable Switch and Interface Card for Disaggregated OPS/OCS Data Center Networks," in Proc. ECOC2015, Valencia, 2015.
- [9] Amaya N, Zervas G, Simeonidou D. Introducing node architecture flexibility for elastic optical networks, Journal of Optical Communications and Networking, 2013, 5(6): 593-608.
- [10] S. Jain, V. J. F. Rancaño, T. C. May-Smith, P. Petropoulos, J. K. Sahu, and D. J. Richardson, "Multi-Element Fiber Technology for Space-Division Multiplexing Applications," Opt. Express, vol. 22, no. 4, pp. 3787–3796, Feb. 2014

- [11] W. Miao et al., "SDN-enabled OPS with QoS guarantee for reconfigurable virtual data center networks," in *IEEE/OSA JOCN*, v.7, n.7, pp.634-643, (2015).
- [12] "Final LIGHTNESS network control plane prototype", deliverable D4.5, May 2015.
- [13] "The LIGHTNESS network control plane protocol extensions", deliverable D4.2, June 2014.
- [14] W. Miao, J. Luo, S. D. Lucente, H. Dorren, and N. Calabretta, "Novel flat datacenter network architecture based on scalable and flow-controlled optical switch system," *Opt. Express*, vol. 22, no. 3, pp. 2465–2472, Feb. 2014.
- [15] W. Miao, X. Yin, J. Bauwelinck, H. Dorren, and N. Calabretta, "Performance assessment of optical packet switching system with burst-mode receivers for intra-data center networks," in *2014 European Conference on Optical Communication (ECOC)*, 2014, pp. 1–3.
- [16] G. M. Saridis, S. Peng, Y. Yan, A. Aguado, B. Guo, M. Arslan, C. Jackson, W. Miao, N. Calabretta, F. Agraz, S. Spadaro, G. Bernini, N. Ciulli, G. Zervas, R. Nejabati, D. Simeonidou, "LIGHTNESS: A Deeply programmable SDN-enabled Data Centre Network with OCS/OPS Multicast/Unicast Switch-over", *ECOC 2015, PDP.4.2*, Valencia, Spain, September 27 – October 1, 2015.

6.Acronyms

AoD	Architecture on Demand
API	Application Program Interface
DC	Data Centre
DCN	Data Centre Network
FPGA	Field-Programmable Gate Array
GUI	Graphical User Interface
HPC	High Performance Computing
LUT	Look-Up Table
MPI	Message Passing Interface
NIC	Network Interface Card
OCS	Optical Circuit Switching
ODL	OpenDaylight
OF	Open Flow
OPS	Optical Packet Switching
OS	Operating System
QoS	Quality of Service
REST	Representational State Transfer
SB	Southbound interface
SDN	Software Defined Networking
ToR	Top of the Rack
VDC	Virtual Data Centre
VM	Virtual Machine
WSS	Wavelength Selective Switch