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 D3.2

Implementation results of the OPS switch, the OCS switch, and the TOR switch

Due date:	30/04/2014
Submission date:	10/06/2014
Deliverable leader:	TUE
Author list:	Wang Miao (TUE), Nicola Calabretta (TUE), Harm Dorren (TUE), Yi Shu (UNIVBRIS), Georgios Zervas (UNIVBRIS), Yan Yan (UNIVBRIS), Shuping Peng (UNIVBRIS), Reza Nejabati (UNIVBRIS), Dimitra Simeonidou (UNIVBRIS), Salvatore Spadaro (UPC), Fernando Agraz (UPC)

Dissemination Level

☒	**PU:**	Public
□	**PP:**	Restricted to other programme participants (including the Commission Services)
□	**RE:**	Restricted to a group specified by the consortium (including the Commission Services)
□	**CO:**	Confidential, only for members of the consortium (including the Commission Services)

Abstract

This document presents the results of the implementation and experimental evaluation of the hybrid data plane deployed in LIGHTNESS.

LIGHTNESS data plane integrates innovative optical switching technologies to provide scalable, flexible, high bandwidth and low latency optical interconnectivity within the data centre network (DCN) infrastructure. To aggregate and intelligently classify the traffic generated by the servers, a Network Interface Card (NIC) has been designed and implemented. A pure optical top of rack (ToR) switch approach has been considered to deploy high-bandwidth and low-latency DCN. The ToRs are interconnected with each other through the architecture-on-demand (AoD) switch, building a flat DCN. The AoD enables a flexible interconnection of optical circuit switching (OCS) and optical packet switching (OPS) technologies for handling long-lived/short-lived data flows respectively. This flat DCN allows for direct ToR-to-ToR and even server-to-server communication.

This document provides the implementation scheme of the LIGHTNESS data plane based on the design proposed in deliverable D3.1. System performance has been investigated including the experimental evaluation of each sub-system. Experimental assessments on the required hardware for controlling the optical switches as well as interfacing the switches with the control plane are reported in detail. The results will be utilized as a valuable feedback for the architecture optimization in WP2 and for the developing of the prototypes in WP3.

Table of Contents

Figure Summary

Table Summary

0. Executive Summary

To accommodate the exponential increase of data traffic driven by emerging services and applications, data centres (DCs) are required to provide more powerful capabilities. Current multi-level DC network infrastructure has the limitations of bandwidth bottleneck and latency that investigations on novel technologies for DCN have been widely carried out. The LIGHTNESS project proposes a flat and scalable data plane architecture which supports high-bandwidth, dynamic and on-demand network connectivity in combination with a unified network control plane.

This deliverable mainly focuses on the implementation and performance evaluation of the data plane in LIGHTNESS. Based on the design scheme reported in D3.1, the switching nodes including ToR, OCS and OPS are numerically and experimentally investigated. By utilizing high-ended hybrid NIC plugged into each server, the traffic is aggregated and diverged in either long-lived data flows or short-lived packet flows. With the integration of advanced transponders, the inter-rack traffic could be directed to OCS/OPS switching node through the optical ToR switch. Moreover, the intra-rack server-to-server communication is supported with the mesh connectivity between NICs. The AoD-based based on OCS backplane and OPS switches interconnect all the ToR switches. This AoD is able to handle the long-lived traffic and the short-lived traffic. The control plane can dynamically configure the NIC, AoD based on OCS and OPS through the southbound interface, which enables highly efficient intra-DC network.

The hybrid NIC is implemented with a Field-Programmable Gate Array (FPGA)-based platform which interfaces with servers for traffic aggregation/segregation and optical transceivers for intra-/inter-rack communications. Benefiting from this, the ToR switch only contains optical components avoiding extra processing efforts. Based on different technologies, four scenarios have been proposed and analysed for the implementation of optical ToR switches interconnection based on AoD. The filtering effects, the crosstalk between adjacent channels as well as sensitivity have been characterized for the AoD to accommodate various optical components. The modular distributed controlled OPS node is also investigated in terms of packet loss and latency. As the key elements of the prototype, the label processor and label generator are implemented with compact PCB (Printed Circuit Board) design and the performance has been evaluated.

To fulfil the controlling mechanism enabled by software defined network (SDN) framework, each switching node will be accessed by the control plane through the dedicated control plane southbound interface. OpenFlow (OF) has been chosen as the protocol for the communication and an OpenFlow agent will bridge the control plane with each LIGHTNESS data plane optical device. The implementations of the OpenFlow agents are presented in this deliverable.

1. Introduction

1.1. Motivation and scope

Data centre network is continuously growing in size and complexity to accommodate the increasing demand of high performance computers (HPCs) and data-intensive applications [1,2]. Most current DCNs are configured in a hierarchical structure due to limited port-count and speed of electrical switches. However, within this architecture, reaching higher interconnectivity level aiming at future DCN will greatly deteriorate the performance of the bandwidth and the latency [3-5]. In addition, power consumption, heat dissipation, as well as space occupancy will also become critically limiting figures of merit in the design of future DCNs.

The LIGHTNESS project, aiming at solving above issues, has proposed an advanced flat data plane architecture for intra-DCN which integrates both OCS and the OPS optical switching technologies. The improvements of capacity, power consumption and latency performance are benefited from the photonic-based solutions. Hybrid NIC located in each server supports the switching over of the traffic to either OCS or OPS resulting in an efficient utilization of the network bandwidth. The different switching granularity enabled by this technique could also guarantee the quality of service (QoS) and latency requirements for various applications. In combination with a unified control plane, the DCN resources could be dynamically and flexibly previsioned in a SDN framework to overcome the deficiency of static configuration.

The data plane architecture and design scheme of each switching node have been reported in D3.1 [6]. This document mainly presents the implementation and system evaluation of LIGHTNESS data plane. To facilitate the communication with the SDN control plane, the implementation of such interface including agent operation mechanism and protocol extension are explained in detail.

1.2. Structure of the document

This document is structured as follows.

Chapter 2 provides an introduction on LIGHTNESS data plane architecture including both intra-cluster and inter-cluster scenarios. According to the previous design proposed in D3.1, the test bed used for evaluating the performance is introduced and the general description on the OpenFlow agents has been provided.

Chapter 3 describes the implementation of the hybrid NIC and four possible solutions for implementing the optical ToR switch. The operation has been explained and the benefits of each technique have been given. The investigations of sensitivity, filtering effect and crosstalk between adjacent channels have also been presented for the AoD.

Chapter 4 focuses on the implementation of modular distributed controlled OPS node and the analyses on the packet loss and latency have been provided for a 4x4 system. For the integration task of the prototype, the label processor and label generator have been designed and fabricated. The results of numerical and experiment evaluation have been illustrated. At the end of this chapter, the implementation of the OpenFlow agent for the OPS is presented.

2. LIGHTNESS data plane implementation

Future Data Centre Networks are required to provide high scalability and flexibility, besides of higher intra-Data-Centre bandwidth and cost/energy efficiency, while minimizing traffic latency. As analysed in D3.1 [6], both Optical Circuit Switching based DCNs and Optical Packet Switching based DCNs have their own difficulties to fulfil all these objectives. In our solution, OCS and OPS are synthesised into DCN design so that they can complement each other: while OCS-based DCN can accommodate long-lived high-capacity smooth data flows with very little latency, OPS-based DCN can offer flexible bandwidth capacity for each optical link when facing dynamic and unpredictable traffic demands with either short or long lived data flows.

As proposed in D3.1, the flexible and programmable DCN architecture with Architecture on Demand node [7] is illustrated in Fig. 2.1 and Fig. 2.2. For the intra-cluster DCN architecture, one cluster consists of a number of racks of servers (75 racks are shown in Fig. 2.1). A ToR switch is used for interconnection of the servers in one rack to the multiplexer (MUX)/de-multiplexer (DEMUX) cluster interface. Channels are provided by each ToR and each channel can support either OPS or OCS transmission which is controlled by the ToR switch. Those channels are combined by a MUX or a DEMUX, creating an optical fibre input and output ports of one ToR. As shown in Fig. 2.1, an AoD interconnects all the input and output ports of different ToRs though the OCS and OPS modules, and traffic from/to other clusters as well. On the other hand, for the inter-cluster configuration presented in Fig. 2.2, a group of clusters (64 clusters are shown in Fig. 2.2) are interconnected by an inter-cluster AoD. Each cluster has several single-core fibres (or one multi-core fibre) connected to the AoD via a MUX/DEMUX and an optical fibre cable. OPS and OCS signals traveling between any two clusters share the same path to the inter-cluster AoD. An Inter-cluster OPS module is connected to the inter-cluster AoD node for dynamic and unpredictable traffic demands with either short or long lived data flows between clusters. ToRs in different clusters communicate with each other through different channels.

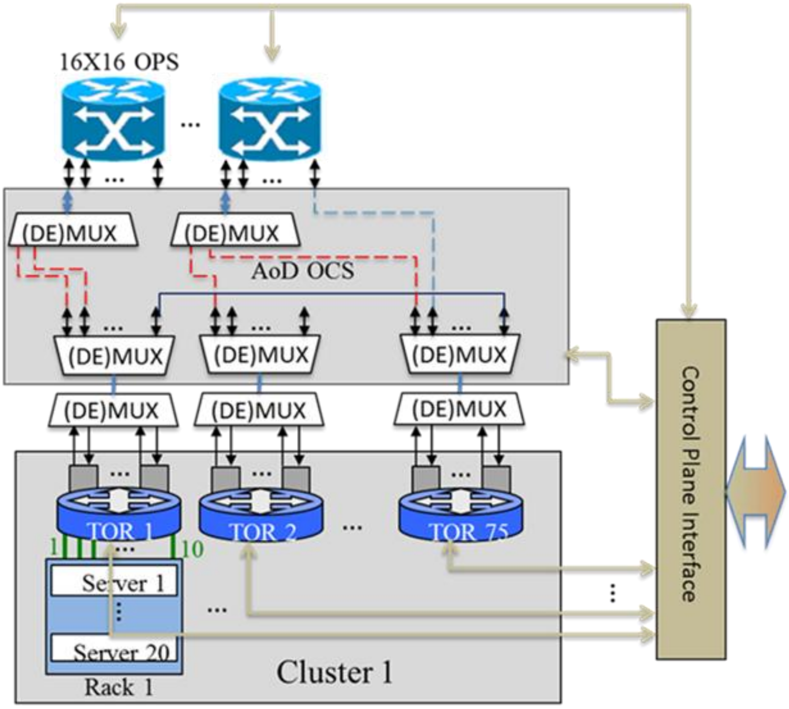

Figure 2.1: AoD-based intra-cluster DCN architecture

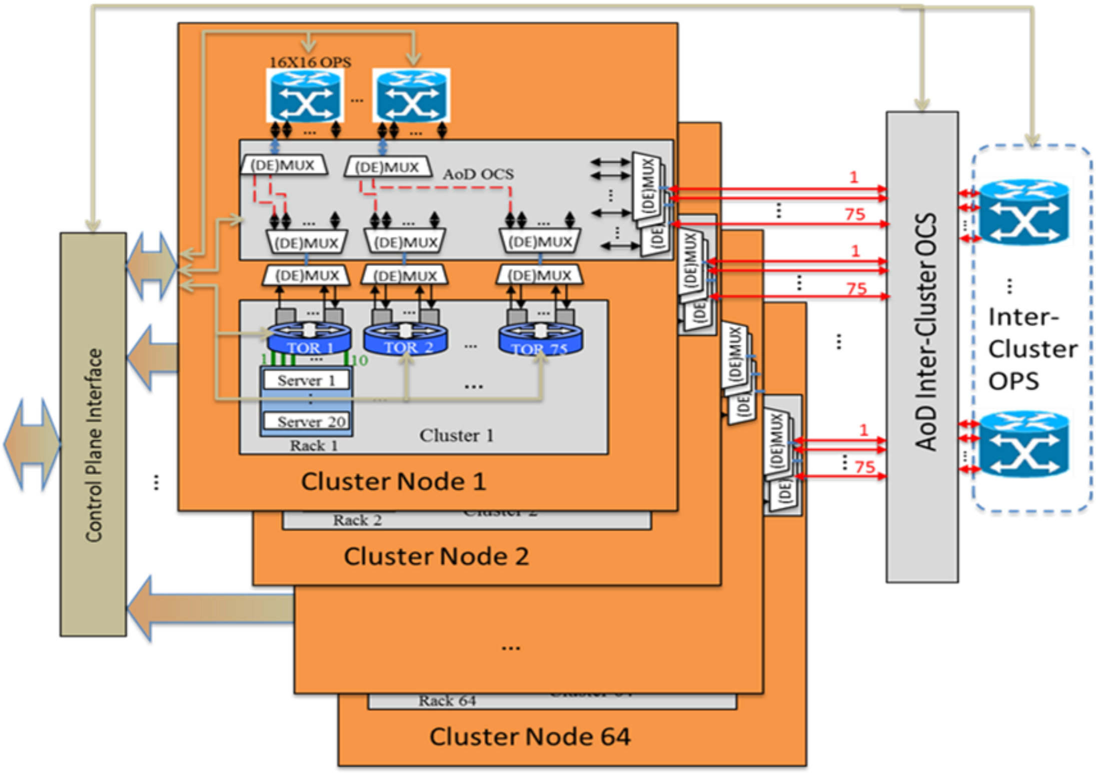

Figure 2.2: AoD-based inter-cluster DCN architecture

2.1. Test bed description of the hybrid OCS/OPS data plane

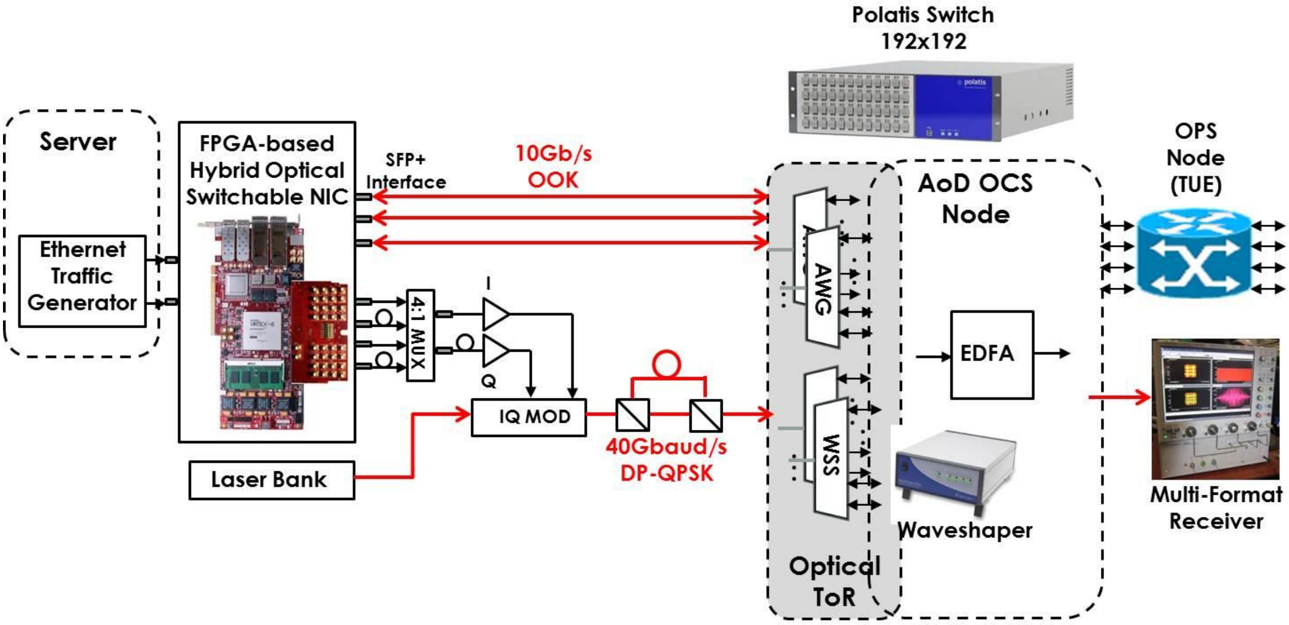

Figure 2.3: Test bed setup for AoD-based DCN

Figure 2.3 shows the experimental system setup of the AoD-based DCN. As proposed in D3.1, the hybrid optical NIC, plugged into each server, is implemented with a high performance FPGA-based platform (Xilinx Virtex 7 XC7VX690 FPGA). This FPGA-based line card is able to substitute the traditional NIC and communicate with the server through PCI express interface. FPGA Mezzanine Card (FMC) interface on this platform with 10x10Gbps channels can support 10x10G Wavelength Division Multiplexing (WDM) C Form-factor Pluggable (CFP) links or up to one channel of 100G for inter-rack communication. Two Avago MiniPods with 24x10Gbps links in total can support 24x10G vertical-cavity surface-emitting laser (VCSEL) for intra-rack communication. Besides, it is capable of supporting some other functions such as switching over between different on-chip network functions such as short lived packet flows (OPS) and long-lived data flows (OCS) on demand.

An Ethernet traffic generator is used to emulate Ethernet signals from server. The groomed and classified ingress traffic from multiple access links is then assigned to different egress ports of FMC interface. The Ethernet-to-NIC interface can handle variable size of Ethernet traffic from 64 Bytes up to 1500 Bytes and flew-in on different bit rates up to 10Gb/s to egress ports with differential SMA interface.

At the output stage of the FPGA-based hybrid optical NIC, the aggregated data streams are separated in groups to drive several optical transmitters for further aggregation for the optical DCN communication. Transmitters with different modulation formats or symbol rate are attached to different egress ports for further processing and transmission. Egress port assignment is applied according to total capacity requirements and link conditions. Two sets of transmitters are thereby setup for different application scenarios: 1) electrical signals with baud rate able to vary from 10G to 100G are generated and applied to an IQ modulator to modulate different optical channels to obtain high capacity QPSK signals. For example, as shown in Fig. 2.3, a 4:1 multiplexer is used to multiplex four data streams from four egress ports to 40Gbit/s electrical signals. Then the achieved 40Gbit/s electrical signals drive an IQ modulator to modulate 8 external cavity lasers (ECLs) to obtain 40Gbaud QPSK signals. The central wavelengths of 8 ECLs are tuned within a 100GHz grid for 40Gbaud signals. Then a polarization multiplex (PM) stage is used to achieve 40Gbaud PM-QPSK signals. In addition, OOK signals can be also generated by replacing the IQ modulator with an intensity modulator in this setup if required. 2) WDM channels of 10Gbaud OOK NRZ signals can be generated directly by enhanced small form-factor pluggable (SFP+) transceivers. These transmitters are very easy to be implemented and the channel numbers can be up to 10. As illustrated in Fig. 2.3, three channels of 10Gbaud OOK-NRZ signals are directly launched from SFP+ interface of the hybrid optical NIC.

Then, the generated WDM QPSK or OOK-NRZ signals are launched into AoD which consists of an optical backplane, e.g. Polatis beam-steering switch with a large port-count (192x192) [8], connected to several signal processing modules, such as Arrayed Waveguide Grating (AWG)-based MUX and DEMUX, wavelength selective switching (WSS) nodes implemented with Finisar WaveShaper [9], Erbium doped fibre amplifier (EDFA) etc.; and optical fibre inputs and outputs. A number of different AWGs with fixed channel spacing of 50GHz, 100GHz and 200GHz are setup for aggregating or distributing different channels from each NIC while the 1:4 WSS can switch programmable C-band spectrum slots, from 10-GHz up to 5-THz with a 1-GHz resolution, to any egress port.

Different arrangements of inputs, modules and outputs can be constructed by setting up appropriate cross-connections in the optical backplane [7]. Thus, arbitrary DCN architectures can be dynamically created involving only the required transmission and functionality by reconfiguration of cross-connections in the Polatis switch. For example, long-lived traffic generated from a NIC can be transported to another NIC through cascaded AWG or WSS channel in the AoD-OCS node so that servers in different racks can communicate with each other by OCS link; on the other hand, short-lived traffic can be sent to OPS.

The DP-QPSK signals is detected by a multi-format receiver while the OOK signals can be received directly by the SFP+ interface on FPGA board.

Implementation and evaluation of the sub-systems involved in the hybrid OCS/OPS data plane test bed are given in the chapters 3 and 4.

2.2. Implementation of the interface to control plane

The southbound Interface is adopted by the LIGHTNESS SDN control plane to interact with the underlying heterogeneous switches/devices (e.g. OPS, OCS and NIC) in the data plane. For the communications between the control plane and the data plane, a generic node-wide information model is defined to fully expose the features and capabilities of heterogeneous data plane devices to control plane, and open standard protocols (such as OpenFlow) are extended to carry these information. Specifically, the OpenFLow messages are exchanged between the control plane and the physical devices in the data plane through the southbound interface to implement and support different procedures (e.g. collection of features, statistics information and status monitoring in the bottom-up direction, configuration/modification in the top-down direction). This southbound interface is implemented, at the data plane side, by a set of OpenFlow agents which are software pieces that reside on top of the optical data plane devices and enable the communication with the control plane. It utilizes the network elements management interface (like Simple Network Management Protocol (SNMP), Transaction Language 1 (TL1) and Vendor application programming interface (API)) or Ethernet raw sockets to communicate with the data plane devices. Due to the heterogeneity of the data plane, a specific agent has to be designed and developed to control each type of optical device. In particular, the NIC, the OCS and the OPS switches need a dedicated agent.

Nonetheless, an architecture with a common structure has been designed to facilitate the development, maintenance and upgrading of the agents. This architecture is depicted in Fig. 2.4.

The lower block of the agent contains a hardware (or technology) specific component which takes the responsibility for communicating with the physical optical network device. The upper block of the agent is common to all the agents independently from the hardware they manage. Hence, the OF API implements a unified southbound communication protocol (based on an extended version of the OpenFlow protocol in our case), and the resource specification module contains the information model of the hardware to be controlled. Note that, although heterogeneous data plane devices need to store different information, the information model structure has been kept as much homogeneous as possible.

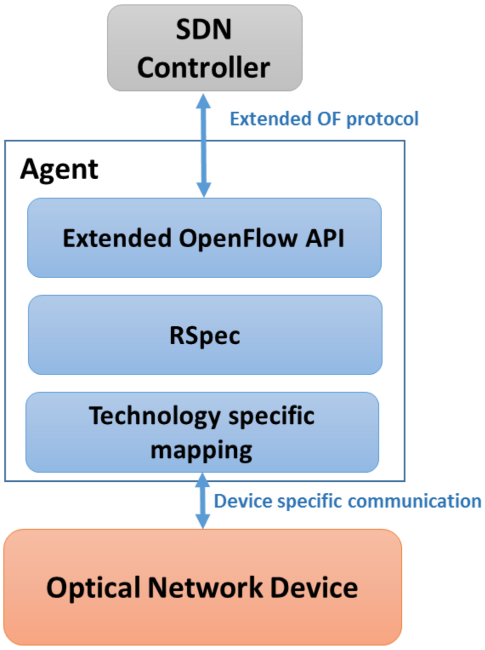

Figure 2.4: OF agent common architecture

Finally, the OpenFlow API component implements the communication with the SDN controller, thus enabling the configuration of the data plane from the control one. The functionalities and mechanisms associated to this (southbound) interface are detailed in [10] and the OpenFlow protocol extensions needed to support such functionalities have been defined in [11].

In the next sections more detailed information related to each particular device are presented.

3. Optical ToR, hybrid NIC, and OCS implementation

3.1. Optical ToR and hybrid NIC Implementation

In this section a more efficient optical ToR architecture and the hybrid optical NIC are presented. An overall picture of the DCN architecture with the optical ToR is illustrated in Fig. 3.1. The hybrid NIC is part of the server, which can aggregate the traffic from the server and sort out between the long and the short lived traffic and map them to different transponders. Thus, the hybrid OPS/OCS switchover functions can be implemented in the Hybrid NIC on the server. The ToR switch on each rack does not need to employ any electronic platform but contains only pure optical components with fixed interfaces connected to each server in the rack. Traffic between different racks is directed to optical switching nodes, e.g. OCS or OPS, through the optical ToR switch based AoD without any latency (apart from the latency introduced by the optical link). The hybrid NIC also includes a group of multi-channels based on VCSEL technology (e.g. Avago MiniPods) for intra-rack communication. Therefore all the servers in a rack are optically connected. With this approach, both the communication between servers within the same rack and the traffic exchange between servers in different racks do not need to pass the electronic ToR switch via optical-electro-optical conversion preventing additional latency. Therefore this approach introduces high efficiency with directly server-to-server links.

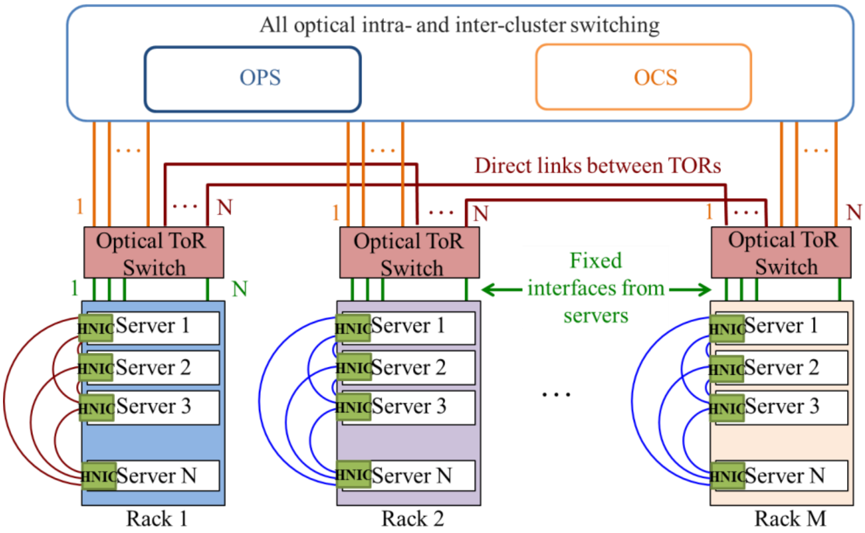
All optical intra- and inter-cluster switching
OPS
OCS
Direct links between TORs
1
N
Optical ToR Switch
Fixed interfaces from servers
HNIC
Server 1
Server 2
Server 3
Server N
Rack 1
Rack 2
Rack M
...

Figure 3.1: Intra-data centre network architecture with optical ToR

The FPGA-based hybrid NIC high-level design is shown in **Errore. L'origine riferimento non è stata trovata.**. It shows several interfaces. The OpenFlow based interface with the control plane is based on 10Gbps link. The interface with the server is based on PCI Express, through that the NIC receives/sends Internet Protocol (IP) packets from/to the server. To realize OPS transmission, a label interface is employed to send/receive labels and acknowledgement (ACK) signals to/from Switch Module FPGA. The inter-rack communication is through 10x10G Dense WDM (DWDM) CFP links while the intra-rack communication is through 24x10G VCSEL MiniPod interface.

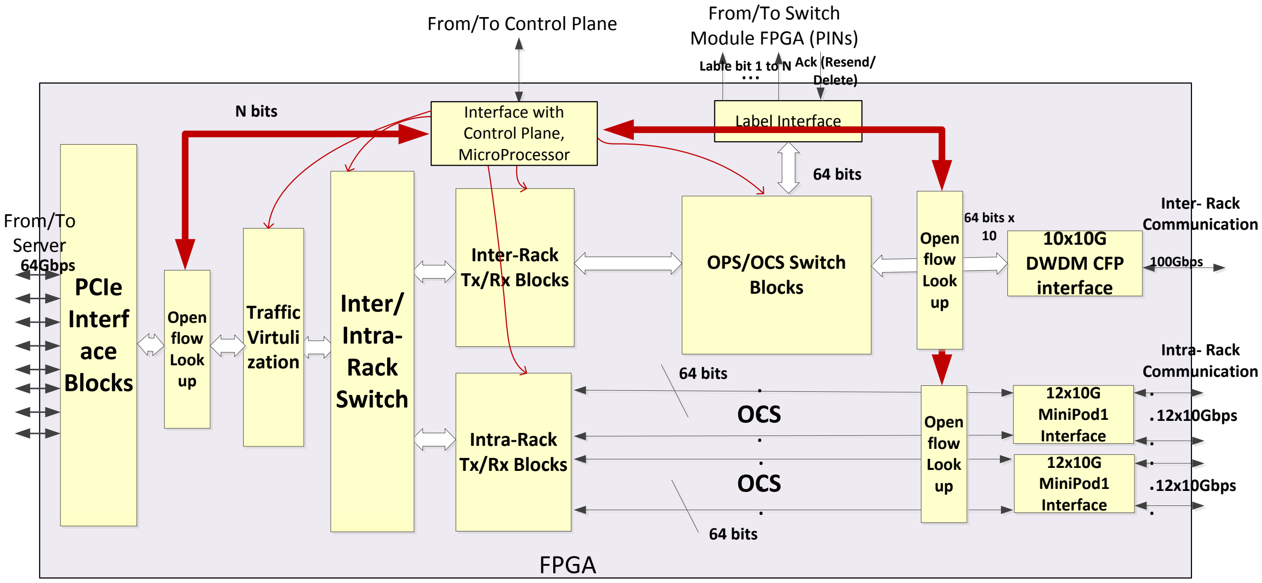

Figure 3.2: Functional Block Diagram of FPGA-Based Hybrid NIC

Four different architectures to implement the optical ToR are investigated. The first architecture is illustrated in Fig. 3.3. The servers with the plugged hybrid optical NIC are aggregated in a rack. Traffics generated from each server are loaded to different optical channels by the NIC. Traffic exchanges between servers in the same rack are realized directly through the 24x10Gbps intra-rack communication links (Avago MiniPods). The traffic from the servers in the same rack (with dedicated

WDM channel for each server) directed to the inter-racks can be simply assembled with AWG and transmitted through the the AoD. The AoD based on OCS and OPS modules forward the traffic to the requested destinations. On the receiving side, received traffic is firstly de-multiplexed by the AWG and each channel is received by corresponding receiver. The central wavelength of each channel is fixed and set according to the channel spacing of AWG used in optical ToR. The type of transmitter on each NIC can be either OOK or DP-QPSK and the baud rate can be chosen between 10Gb/s to 100Gb/s. For example, in Fig. 3.3, a 25Gbaud/s DP-QPSK transmitter is applied on the NIC so that an uplink from NIC with 100 Gb/s capacity is established.

The optical ToR architecture illustrated in Fig. 3.3 is cost-efficient. Only passive components (AWG) are used in this optical ToR, which leads to negligible latency. However, blocking might occurs when servers in different racks with different channel wavelengths need to communicate with each other.

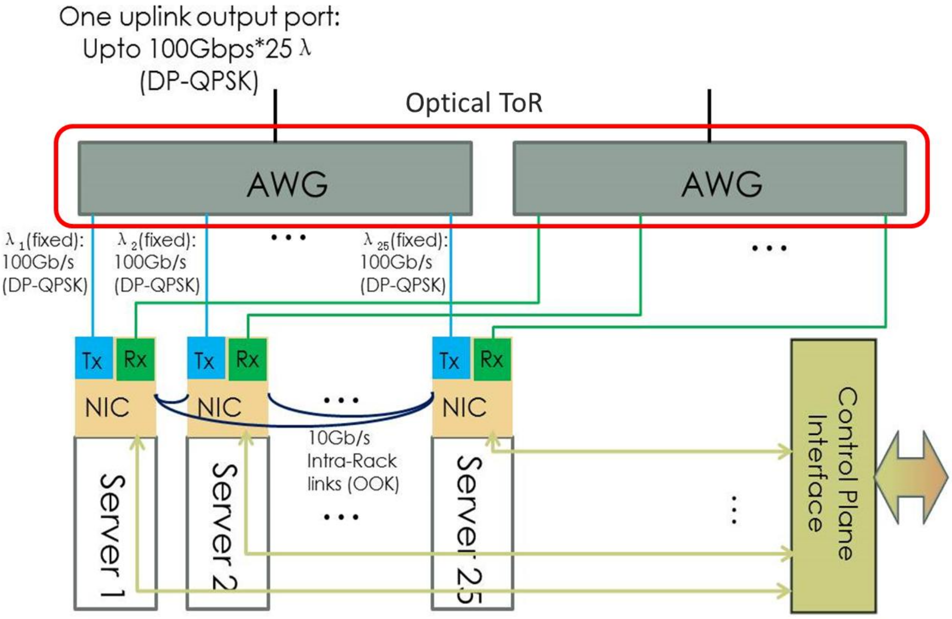

Figure 3.3: Optical ToR Scenario1: implemented with AWG

A possible improvement for this architecture is replacing the 1xN AWG with an NxN Reconfigurable AWG (R-AWG), as illustrated in Fig. 3.4. By introducing a tunable interface in the NIC, it is possible to tune the input wavelength for each channel according to the desired port so that channels from each transmitter can be switched to arbitrary output port of R-AWG. All the output channels from the R-AWG are connected to the AoD. Part of the channels can also be used for direct interconnection between different racks. Thus, connections between different racks or from racks to switching node are flexible and reconfigurable. In Fig. 3.4, a 25Gbaud/s DP-QPSK tunable transmitter is utilized on the NIC and a 25x25 R-AWG is employed for the optical ToR, so that 25 uplinks and each with 100 Gb/s capacity from the optical ToR are achieved.

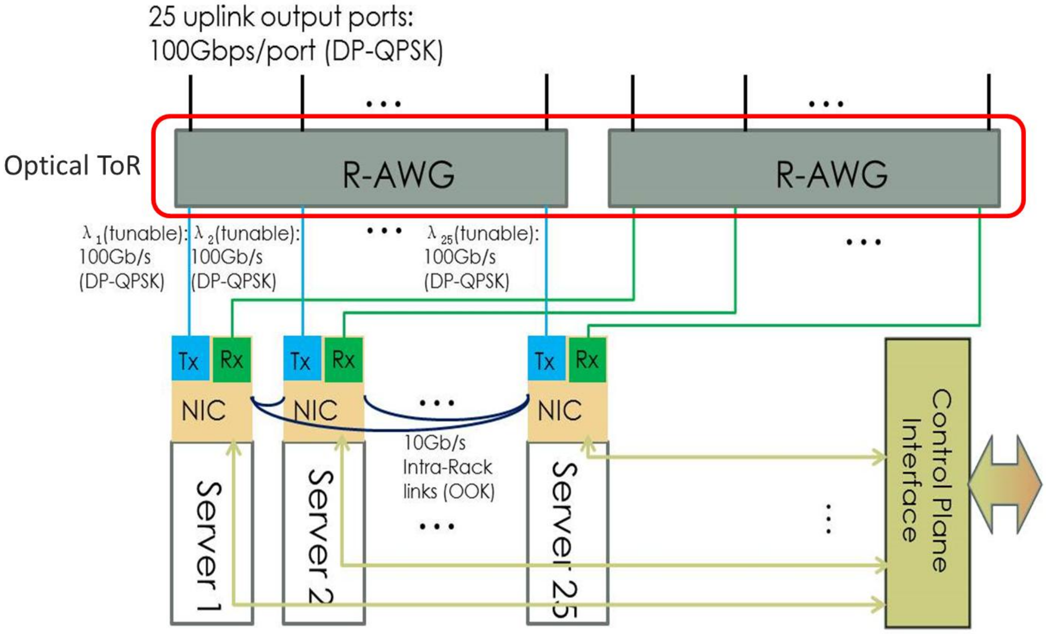

Figure 3.4: Optical ToR Scenario2: implemented with NxN R-AWG

Another possible architecture is illustrated in Fig. 3.5. The utilization of WSS switches in the optical ToR, instead of AWGs or R-AWGs, provides extra flexibility to the ToR. An NxM WSS switch with M input ports and N output ports can switch programmable C-band spectrum slots from 10-GHz up to 5-THz with a 1-GHz resolution between arbitrary ingress and egress ports of the WSS. Thus, apart from tuning the input wavelength for each channel, channels from each NIC can be selected at will and aggregated at arbitrary output port. The re-assembled traffic then is switched to OPS or OCS node accordingly and sent to the specific port of WSS on receiving side so that it can be received by the required server. In this scenario, N uplinks are connected from optical ToR to AoD and the capacity of a single unlink can be up to 2.5Tb/s depending on traffic demands. The definitions of WSS ports number N and M depend highly on which type of NxM WSS is commercially available. An alternative solution to realize NxM WSS could be cascading two 1xN WSSs as MUX and DEMUX respectively. The WSS switches are able to be controlled by open flow based SDN controller. The implementation of interface with control plane is described in Section 3.4.

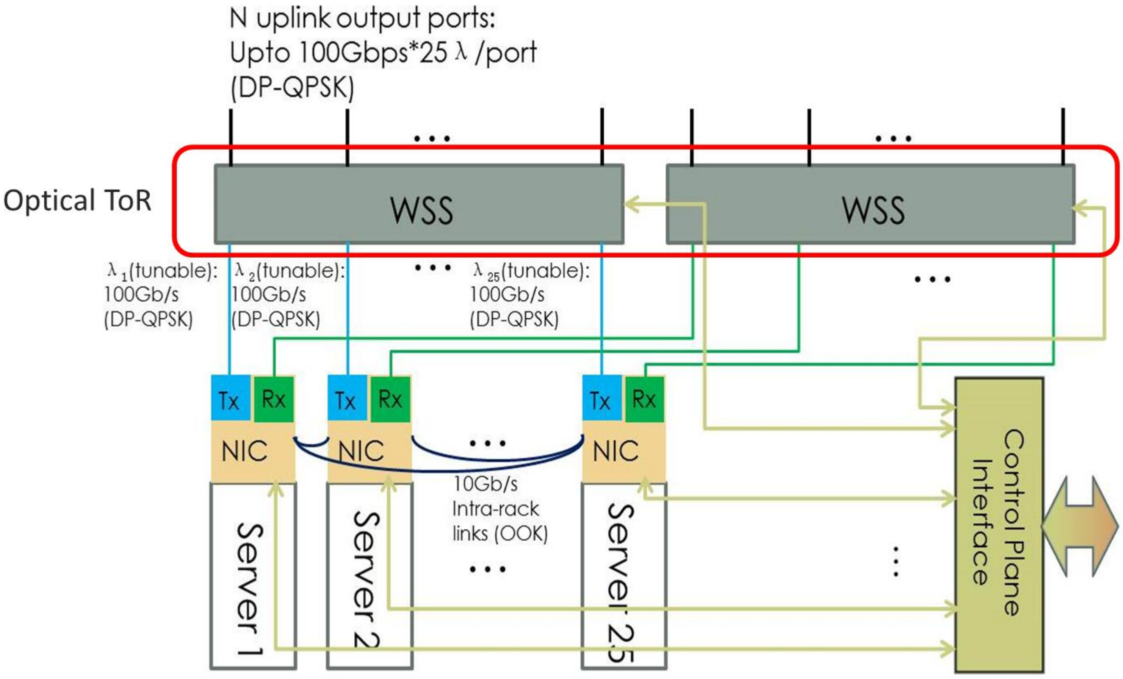

Figure 3.5: Optical ToR Scenario3: implemented with NxM WSS

As the WSS switch is able to program spectrum slots and slot shape for any internal link, there is another advantage we can take from it. The central wavelengths of each channel are not necessary to be tuned within WDM grid but can be more flexible. Especially for OOK signals, smaller channel spacing can be applied, which leads to more channels available for the optical ToR. For example, as illustrated in Fig. 3.6, four channels with 25 Gb/s OOK transmitters can be set up for each NIC and the spacing between adjacent channels can be lower than 50G. Thereby each server can communicated with 4 other servers at the same time and 100 channels in total are available in the optical ToR. The capacity of a single uplink from the ToR can be up to 2.5Tb/s. In addition, employing OOK signals for traffic exchanging in data centre also reduces the complexity and as well the cost of transponders and receivers. However, this scenario clearly requires a larger amount of input and output ports for the WSS device.

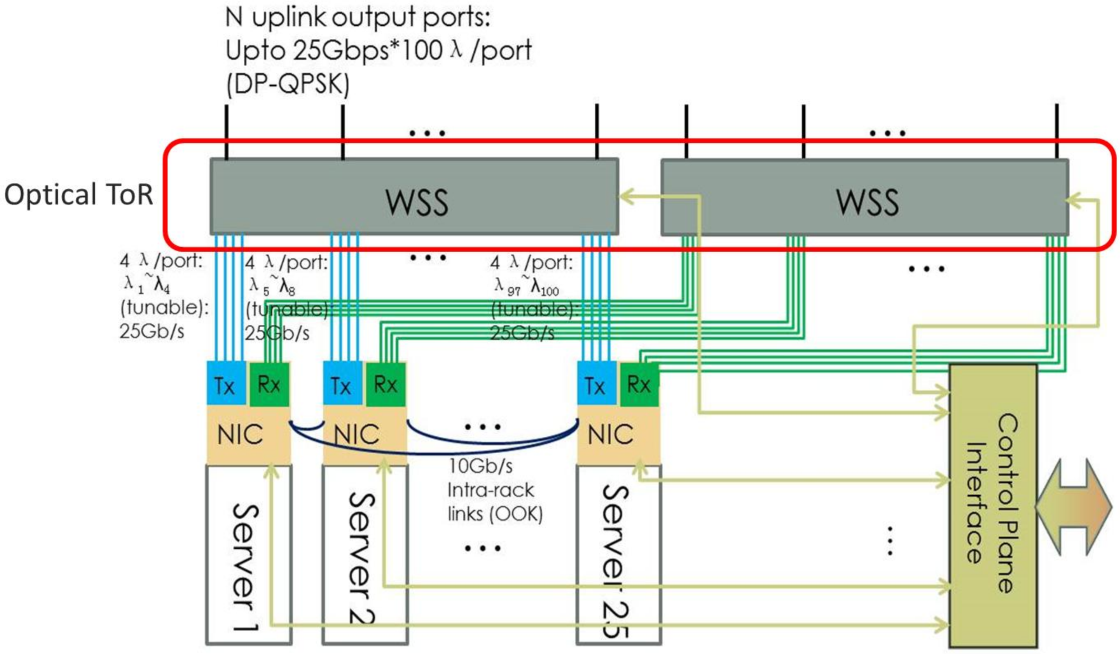
N uplink output ports:
Upto 25Gbps*100 λ /port
(DP-QPSK)
Optical ToR
WSS
WSS
4 λ /port: λ1~λ4 (tunable): 25Gb/s
4 λ /port: λ5~λ8 (tunable): 25Gb/s
4 λ /port: λ97~λ100 (tunable): 25Gb/s
Tx
Rx
NIC
10Gb/s Intra-rack links (OOK)
Server 1
Server 2
Server 25
Control Plane Interface

Figure 3.6: Optical ToR Scenario4: implemented with NxM WSS and multi-channel Txs

3.2. OCS Implementation

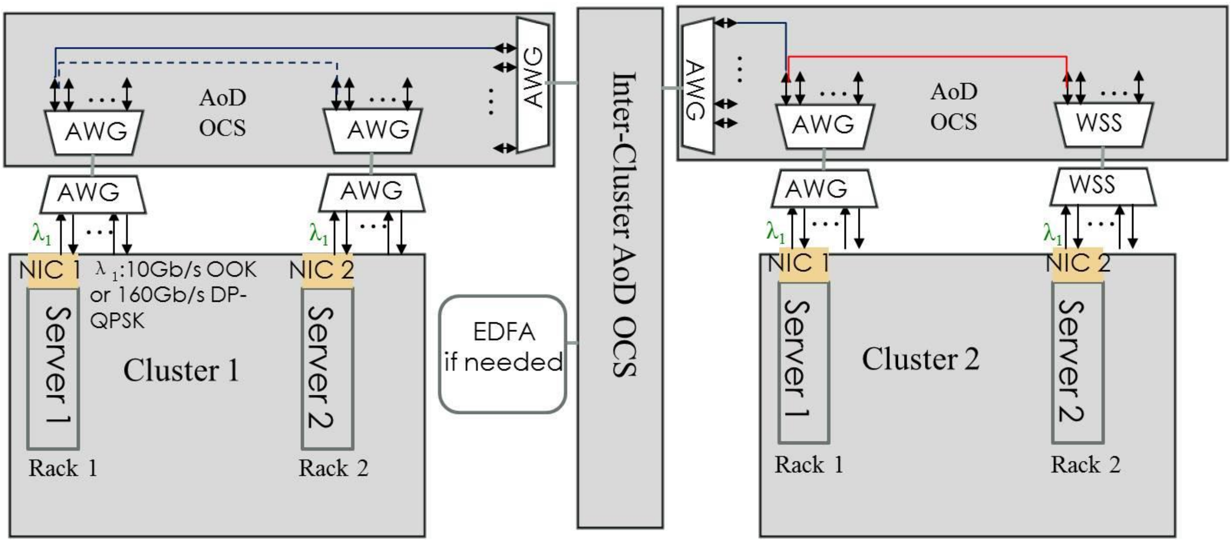

Figure 3.7: Configuration of AoD for intra- and inter-cluster DCN

The experimental setup to evaluate the OCS based AoD is illustrated in Fig. 3.7. Two FPGA-based optical NICs (NIC 1 and NIC 2) represent two servers in different racks within Cluster 1 of DC. Pseudo-random bit sequences (PRBS) code of length 2^{15}-1 is generated from FPGA to emulate traffic signals from Server 1 in Cluster 1, which is assigned to different egress ports of FMC interface on NIC 1. Two alternative transmitters are setup for different application scenarios: 1) a 4:1 multiplexer is used to multiplex four data streams from four egress ports to 40Gbit/s electrical signals. Then the achieved 40Gbit/s electrical signals drive an IQ modulator to modulate 8 ECLs to obtain 40Gbaud QPSK signals.

The central wavelengths of 8 ECLs are tuned within a 100GHz grid for 40Gbaud signals. Then a polarization multiplex stage is used to achieve 160Gb/s PM-QPSK signals for all 8 optical channels. 2) 10Gb/s OOK-NRZ signals are generated directly by WDM SFP+ transceivers on NIC 1. The central wavelengths of these optical channels are fixed and agreed with DWDM frequency grid. In addition, OOK signals with baud rate varying from 10G to 100G can also be generated by replacing the IQ modulator with an intensity modulator in the former transmitter setup, in which case the central wavelengths of OOK-NRZ channels are tunable.

Then, the generated QPSK or OOK-NRZ channel λ_1 is launched into an AWG (optical ToR Scenario 1) and switched to the intra-cluster AoD-OCS node, which consists of a large port-count optical backplane (192x192 Polatis beam-steering switch), and AWG-based MUXs and DEMUXs. Channel λ_1 can be selected and transported through another three cascaded AWGs by building required cross-connections in the optical backplane, and received by NIC 2 in Cluster 1 (the dash line in Fig. 3.7).

Another intra-cluster AoD-OCS node is setup as well. Similar to the configuration of Cluster 1, two FPGA-based optical NICs (NIC 1 and NIC 2) represent two servers in different racks within Cluster 2. Instead of AWGs, optical channels from NIC 2 are connected to a pair of cascaded 1x4 WSSs (Finisar Waveshaper) which emulate one 4x4 WSS module (optical ToR Scenario 3&4), before switched to the intra-cluster AoD-OCS node. Traffic generated by NIC 2 in Cluster 2 is loaded on channel λ_1 and transmitted through cascaded WSSs and AWGs, received by NIC1 in Cluster 2 eventually (red solid line in Fig. 3.7).

In order to emulate traffic exchanges between two servers in different clusters, an inter-cluster AoD-OCS node is setup to connect Cluster 1 and Cluster 2. In this case, traffic generated by NIC 1 in Cluster 1 is loaded on channel λ_1 and transmitted through six cascaded AWGs, received by NIC 1 in Cluster 2 by setting up required cross-connections in the optical backplane for both intra- and inter-cluster AoD-OCS nodes (blue solid line in Fig. 7). An EDFA node is connected to the inter-cluster AoD-OCS node and can be inserted into intra- or inter-cluster traffic link if required.

An off-line multi-format receiver is used in above setup to receive DP-QPSK signals. The 10Gb/s OOK-NRZ signal can be received directly by SFP+ transceiver on the optical NIC and the BER performance is measured by error counter in the FPGA.

3.3. System evaluation

3.3.1. Receiver Sensitivity of Optical ToR for OOK signal

In order to make optical DCN power-efficient and cost-efficient, we are trying to avoid the employment of high energy-consuming optical modules, e.g. EDFA and Wavelength Converter (WC), in the AoD-OCS node. On the other hand, thanks to the short distance of optical link in data centre, optical loss and dispersion during propagation in optical fiber are negligible. Signal quality is only affected by channel filtering and crosstalk from adjacent channels. Especially for receiving 10 Gb/s OOK-NRZ signals, a very low receiving power threshold can be supported thanks to the utilization of Avalanche Photodiode (APD) for the SFP+ interface on optical NIC.

Table 1 lists the measured average insertion loss of optical modules adopted in the OCS based AoD. AWGs with different channel spacing (50GHz, 100GHz and 200GHz) have very similar insertion loss when filtering the 10 Gb/s OOK-NRZ signal. Figure 3.8 shows the 10 Gb/s OOK-NRZ sensitivity versus filter bandwidth for WSS.

Optical Module	Polatis Beam-Steering Switch	AWG (100GHz spaced)	1x4 Waveshaper (50G bandpass filter)
Average Insertion Loss	-0.8dB	-3.67dB	-4.67dB

Table 3.1: Measured insertion loss for optical modules adopted in the OCS based AoD

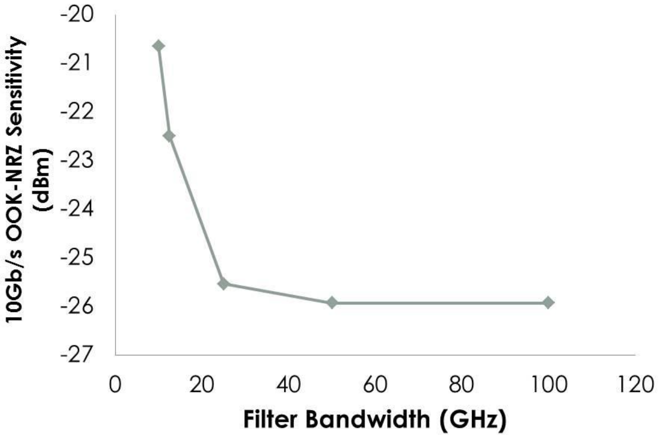

Figure 3.8: Filtering effect on 10 Gb/s OOK-NRZ signal for WSS detected by the SFP+ transceiver on optical NIC

In the AoD-based DCN architecture, cascaded WSS or AWG stages may result in a reduced end-to-end channel bandwidth, which may cause signal distortions with an associated power penalty [13]. On the other hand, if the selected bandwidth in WSS is too wide for the transported channel, spectral resources are wasted. In order to evaluate the effect of narrow filtering, a 10 Gb/s OOK-NRZ signal is transmitted through a WSS and received by the APD at SFP+ interface on optical NIC. The filter bandwidth is decreased from 100 GHz down to 10 GHz while the minimum power of the output signal can be detected by APD is observed. Results are presented in Fig. 3.8. As the filter bandwidth is reduced the edges of the signal spectrum are attenuated. This introduces an increasing power penalty. A 0.4-dB penalty is observed at 25 GHz filter bandwidth and it is increasing rapidly for narrower filter bandwidths.

Another consideration with WSS utilization is related optical ToR Scenario 4 (Fig. 3.6): How narrow the channel spacing while inter-channel crosstalk is suppressed by WSS filters? Figure 3.9 shows the performance of a 10G channel with an adjacent 10G channel at varying channel spacing de-multiplexed using the Waveshaper as a filter with 10-GHz bandwidth. There is a flat region where the SNR shows little variation, from 50 GHz down to around 25 GHz. The penalty at 20-GHz spacing is 1 dB and increases rapidly for narrower channel spacing. Furthermore, packing channels closer together also increases the interaction between them and may give rise to non-linear impairments such as XPM and FWM, which also constrain channel spacing.

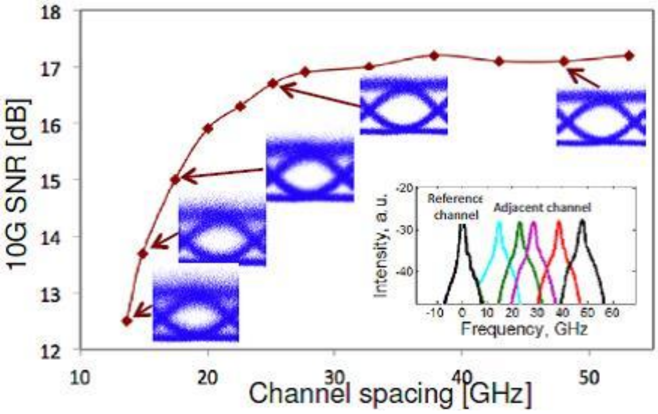

Figure 3.9: 10Gb/s SNR performance for varying channel spacing in WSS

BER performance of 10Gb/s OOK-NRZ, as illustrated in Fig. 3.10, is measured at the SFP+ transceiver on optical NIC after attenuation with following settings for the FPGA:

- Tx Differential Output Swing = 1200 mV;
- Tx Pre-Emphasis =1;
- Tx Post-Emphasis =1;
- Rx Equalization =0.15dB;

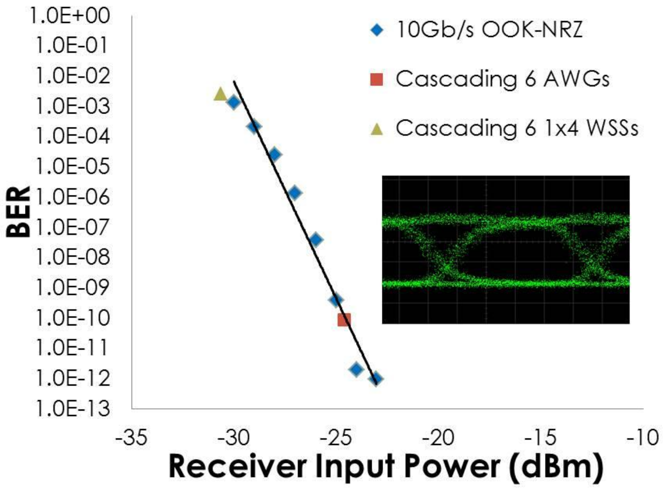

Figure 3.10: BER performance of 10Gb/s OOK-NRZ signal at the SFP+ transceiver on optical NIC

The output power from SFP+ transponder on the NIC is measured as 1.4dBm, and the receiving power threshold of APD on the NIC is measured as -30dBm. Error free performance is achieved when receiving power higher than -23dBm. In other words, error free performance can be achieved for the communication in the DCN if power loss of optical signals is lower than -24.4dB.

As shown in Fig. 3.7, for traffic exchange within the same cluster, optical signals need to travel through four AWGs or WSSs and three cross-connections inside optical backplane. The link loss during the path is -20.05dB in the case of cascading AWGs and -21.08dB in the case of cascading WSSs. Thus, error free communications can be supported preventing the use of EDFA for intra-cluster DCN architecture.

For communications between different clusters, optical signals need to travel through six AWGs or WSSs and five cross-connections inside optical backplane, as shown in Fig. 3.7. The BER performance in this case is shown in Fig. 3.10 as well. The link loss during the path is -30.48dB in the case of cascading AWGs and -32.02dB in the case of cascading WSSs. Thus, EDFA nodes are needed for inter-cluster DCN architecture.

3.3.2. Receiver Sensitivity of Optical ToR for QPSK signal

As described in Section 3.2, 8 channels of 40Gbaud DP-QPSK signals are generated and launched into cascaded 100GHz spaced AWGs or WSSs from optical ToR. Thus, 1.28Tb/s ToR-to-ToR traffic are achieved. A multi-format receiver is used for demonstration of processing received QPSK signals. All the signals have to be amplified before they are sent to this coherent receiver. Fig. 3.11 shows the spectrum of aggregated 8 channels with DP-QPSK signals. For all the channels, almost error-free transmission is achieved. The eye diagram and the recovered constellation diagram are shown as an insert in Fig. 3.11.

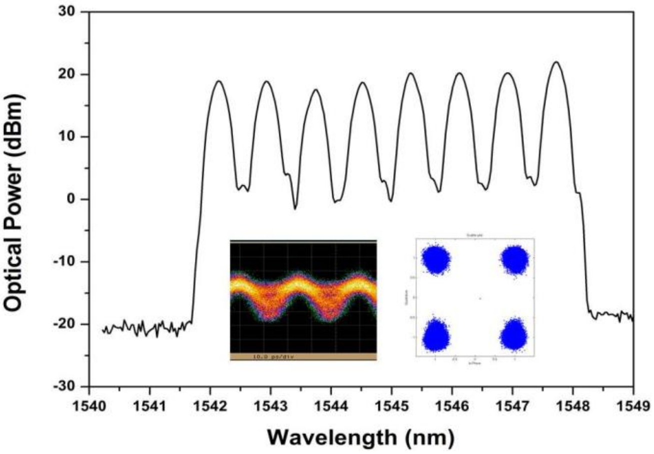

Figure 3.11: Optical spectrum of aggregated 8 channels of DP-QPSK signals in 100GHz Grid

3.4. Interface with the control plane

A brief summary of interface between LIGHTNESS control plane and data plane devices is presented below. Full OpenFlow extensions and control plane southbound interface design details (including OF agent design) are elaborated in D4.2 and D4.3, respectively.

- Interface between the Control Plane and the FPGA-based hybrid optical NIC: To properly forward the packets generated from server, a look-up-table (LUT) is needed for the NIC to direct packets/flows to output ports in a proper way. And, one of the unique features of LIGHTNESS is the flexibility to choose OPS or OCS connections for flows. Specifically for OPS connections, the control plane should identify the sending time slot and the label attached to the packet.
- Interface between the Control Plane and the optical ToR: An optical ToR solution (e.g., bandwidth variable WSS) is selected in LIGHTNESS to provide NIC-to-NIC all optical connection. The cross connection configuration of the ToR is needed from the control plane, especially supporting flex-grid technology.

- Interface between the Control Plane and the OCS: The OCS node is exposed to the control plane as one virtual switch entity with abstracted capabilities of the subsystems attached to the large port backplane. So, the control plane needs to be able to properly utilize the capability of the AoD and configure it (including back plane cross connection and associated subsystem configuration) according to the service requirements.

4. OPS implementation

LIGHTNESS has introduced an OPS architecture with highly distributed control for port-count independent reconfiguration time [14]. The modular structure and RF tones labelling technique adopted in the system enable the parallel processing for each input as well as for each label bits that greatly reduces the processing time. This is extremely important to ensure low latency when scaling to a large number of port-count to satisfy the increasing demand of huge amount of data traffic. As the lack of practical optical buffer and to avoid power-hungry optical-to-electrical-to-optical (O-E-O) conversions, the electrical buffers have been adopted and placed at transmitter side. An efficient optical flow control was introduced at optical level without occupying extra bandwidth and space resource which improves the performance of latency within a simple structure [15]. All these features allow the proposed architecture, applied in short links, for providing a flattened interconnected network with sub-μs latency and high throughput. Moreover, the optical packets could be switched transparently with multiple modulation formats and high data rate to deal with data-intensive applications [16].

Targeting at the prototype implementation required in WP3, the control signalling and the interface between different functional blocks should be well organized and optimized in size especially when electronics are involved. RF tones labelling technique values more on RF electronics and requires fine design for label processing as well as label generation. The label generation also serving as part of the interfaces between ToR and OPS node that plays an important role in the system implementation.

To fulfil the dynamic and on-demand allocation of DCN resources for intra-DC traffic, the switch controller should be able to communicate with the SDN-based control plane. Therefore, for each switch controller a dedicated interface should be employed to allow the control plane for monitoring and dynamically configuring the OPS. An Open Flow agent has been employed to translate the protocol running on the southbound interface into the commands compatible with the switch controller.

In this chapter, we first present the implementation of modular OPS node, the label processor and label generator, and then the evaluation results are given. In the last section, the communication interface with control plane is described in detail and experimental investigation is reported.

4.1. Implementation

4.1.1. Distributed controlled modular OPS

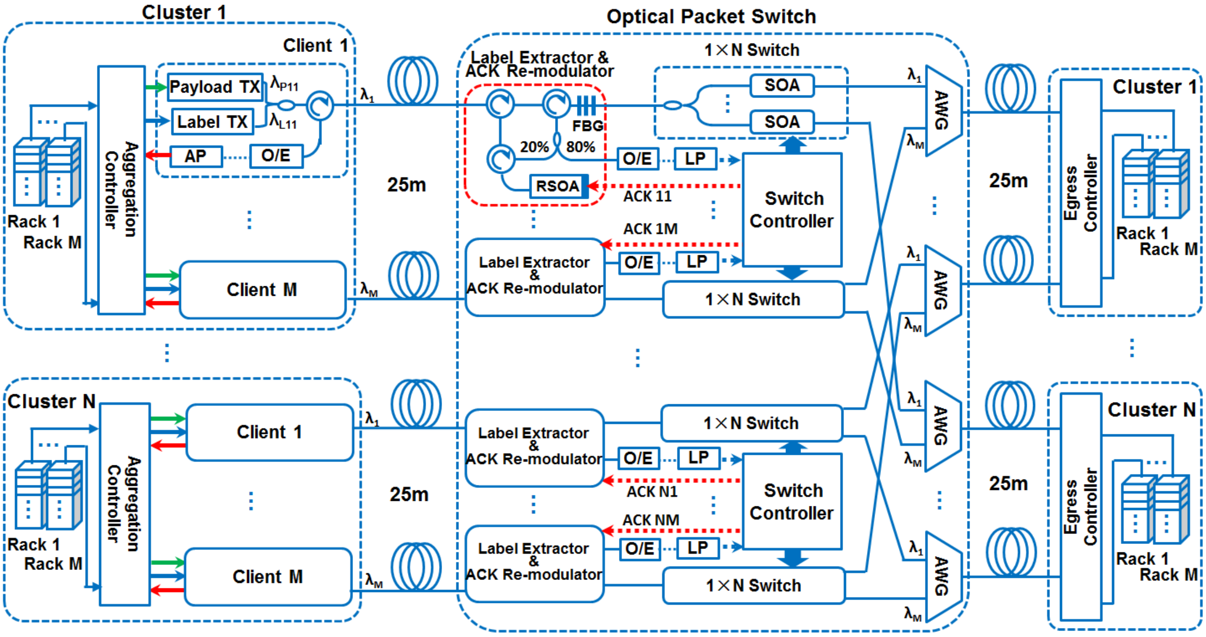

Figure 4.1: OPS architecture with distributed control

The schematic of highly distributed-controlled modular OPS architecture is shown in Fig. 4.1. Each cluster groups M ToRs and an aggregation controller is used for balancing the traffic load and aggregating the input data coming from different ToRs. Traffic flows will be assigned with different wavelength λ_1, λ_2 ..., λ_M and transmitted to OPS node. Switching is performed based on the in-band label information carried by each packet [17]. After the packet being sent out, aggregation controller will store the copy in a FIFO until receiving a positive acknowledgement that the packet has been transported to proper destination.

OPS node consists of N identical modules and each of them handles the packets from the corresponding cluster. Label extractor separates the optical label from the optical payload by using a fibre Bragg grating (FBG). The optical payload is then fed into the Semiconductor Optical Amplifier (SOA) based broadcast and select 1xN switch while the extracted label is split into two parts. One of them is detected and processed by the label processor (LP) after optical-to-electrical conversion (O/E). The switch controller retrieves the label bits, checks possible contentions and configures the 1xN switch to block the contended packets with low priority and to forward packets with high priority. Moreover, the switch controller generates the ACK used to inform the aggregation controller on the reception or re-transmission of the packets. The other part of label power is re-modulated in an reflective SOA (RSOA) driven with the base band ACK signal generated by the switch controller and sent back to cluster side within the same optical link [15]. This fulfils the efficient optical flow control in hardware which minimizing the latency and buffer size. Baseband ACK signal is easily extracted at the edge node by using a 50 MHz low pass filter, to remove the label information at RF frequencies. The adopted modular structure allows highly distributed control which makes the reconfiguration time of the overall switch port count independent. In addition, the N channels of each cluster could be processed in parallel, greatly minimizing processing time and thus the latency. Another advantage benefitted from the modular structure is that the overall performance of the OPS can be evaluated by testing a single module.

4.1.2. Label processor

In-band RF tones labelling technique has been deployed in the OPS system to reduce latency and improve the scalability. M binary coded RF tones are carried by each of the N wavelength which is inserted in-band with the payload's spectrum. Thus MxN label bits and 2^{MxN} ports can be addressed, providing a highly scalable method. Another key advantage is that the labels are processed in parallel, making the processing time independent of number of tones. The realized scheme is modular and easily replicable.

When the packets arrive to the OPS input, the label extractor detects the in-band label wavelength and sends it to the optical label processor, while the optical payload of the packets is forwarded to SOA-based switch matrix in the optical domain. The label extractor here mainly composes of a series of the passive narrow band pass optical filters (e.g. a series of cascading FBGs or integrated comb filters). After the optical to electrical conversion of the extracted label, the multiple RF tones signal is firstly divided into parallel paths and then sent to each channel of the RF tones processor. The obtained baseband label bits are finally sent to the FPGA-based controller that determines the packet destination and generate the switching control signal, and decides the setting of the optical switch matrix to forward the payload. Figure 4.2 reports the detailed implementation of the RF tones processor block for one of the label wavelength.

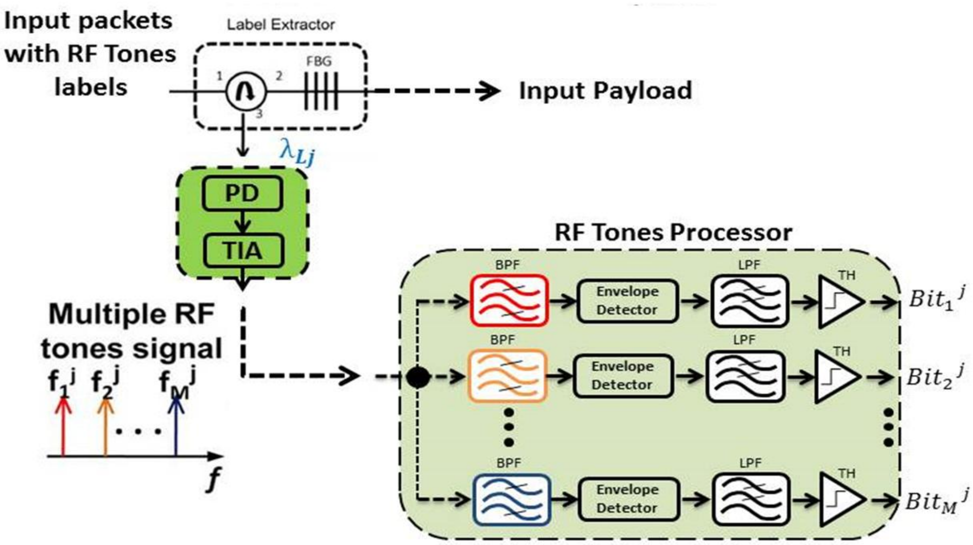

Figure 4.2: Implementation of the processor block for the RF tone in-band optical label

Each parallelized signal passes through the Band Pass Filter (BPF) with a central frequency of f_i^j, (i∈ [1, M], j∈ [1, N]), to select the corresponding RF tone. After the band pass filter, the envelope detector recovers the envelope and then the baseband label bits are shaped by the comparator after the Low

Pass Filter (LPF). Finally, the baseband label bits will be sent to the FPGA for the switching control generation. It should be noted that the RF tones processor block, process all the RF tones labels asynchronously and in parallel. Then the processing time is kept constant in spite of the number of label wavelengths and RF tones increases. With a limited increase in the latency and the complexity, an exponential increase of number of ports can be controlled by the OPS.

The design and realization of the LP should follow certain specifications to guarantee the quality of detected label bits as listed in Table 4.1. First, the time to extract the optical label, convert it from optical to electrical domain, process the multiple RF tones signal and elaborate the information bits must be in the order of few tens of nanoseconds. Each frequency, passing through the correspondent channel, must experience the shortest possible delay. Secondly, the synchronization between the parallel channels must be considered. At the output of the circuit, the M bits must be synchronized to allow an easier management by the switch controller. This means that the rise and fall time of each RF tone must be very similar. Thus, the selection of the BPF and LPF filters is critical because they are the main components that affect the delay and rise/fall time of the entire design. Moreover, the power consumption of the RF tones processor is also important, in particular when OPS with large number of ports is considered. One of the main reason of extra power consumption (and degradation of the signals) is the issue with impedance matching. Particularly, for the compact design on PCB minimizing the size and power consumption, the impedance matching should be well performed. In addition, the input power can be variable because of the different optical path attenuation and the dynamic number of RF tones simultaneously transmitted. As the number of tones increases, the input power for each RF tones decreases and vice versa. The sensitivity should be good enough to recover the label with low power and the distortion caused by the nonlinearity should also be avoided for the high power tones. The LP should have a large dynamic power range to handle the fluctuation of the input power. Finally, the noise mainly coming from the amplification and crosstalk would greatly affect the LP performance. This requires the BPF and amplifier should be carefully selected.

Specifications	Requirements	Attributions
Delay	Low latency: few tens of nanoseconds	The order of BPF and LPF
Rise/Fall time	Few nanoseconds, consistency for all tones	The selection of the BPF and LPF for each channel
Impedance matching	Well match with low reflections for RF components	Transmission line and matching network
Dynamic range	Handling the power fluctuation of different optical path and number of RF tones	Combination of amplifiers and attenuators
Noise and distortions	The crosstalk and distortion from nonlinearity should be controlled	IP3 and gain of the amplifier and characteristics of BPF

Table 4.1: Specifications for label processor

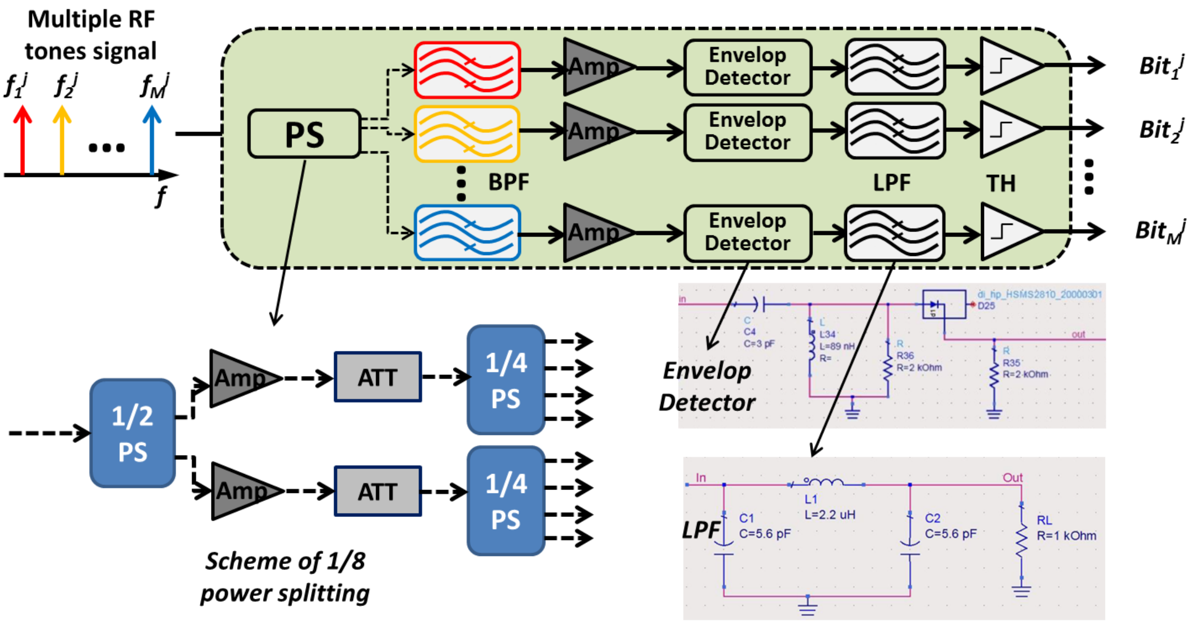

Figure 4.3: Implementation of power splitter, envelop detector and low pass filter

An 8-bit LP has been designed and implemented based on the functional requirements and specifications. The bandwidth of each tone depends on the baseband label signal. For label bits with 300ns duration and 30ns guard-time, the effective bandwidth of the modulated tone is around 100MHz. The frequencies of the tones are ranging from 100MHz to 1.5GHz with a spacing of ~150MHz between adjacent tones.

The first stage of LP is the power splitter (PS) to broadcast the RF tones signal to all the channels. As most of the components are frequency dependent, the higher frequencies will experience higher losses. It is therefore important to properly consider the method to perform the splitting avoiding possible degradation. A tree configuration as shown in Fig. 4.3 has been chosen for the splitter considering both the channel equalization and the scalability issues. Amplifiers are inserted into the paths to compensate the losses caused by the splitters.

The envelop detector utilizes a Schottky Diode based scheme to recover the baseband label signal. However, the impedance of the diode, which depends on the bias current, is very different from the conventional 50Ω impedance of the RF and microwave circuit. An impedance matching network is necessary to optimize the signal power delivered to the diode based envelop detection. The adopted scheme reported in Fig. 4.3 makes a good balance between the sensitivity and complexity of the impedance matching network, and at the meantime, is tolerant of the effect of the non-ideal components.

A LPF is needed to separate the base band envelop from other high frequency components. For the envelop detector, higher load impedance could increase the sensitivity while most conventional LPF has a 50Ω matched impedance. For this reason, the LPF has been designed with discrete components to obtain high impedance and low delay as well as fast response. A 3rd order Butterworth type LPF has been deployed and the scheme is given in Fig. 4.3. It introduces less delay and less rise/fall time than the Chebyshev and Elliptic type filters.

The scheme presented above has been realized with micro-strip transmission lines on a substrate capable to work with frequency of GHz, using PCB technology. The surface mount components have been soldered by reflow technique.

4.1.3. Label generator

Pairing with the LP in the OPS node, at the transmitter side (NIC), there should be a label generator module to convert the digital label bits into the desired optical RF tones in-band label as shown in Fig. 4.4. The label generator (LG) should be able to easily scale to support more destination ports and more functionality of controlling. The optical output wavelength should be centred at the pass band of the label extractor, which means the wavelength should be accurate and stable. The tone purity should be guaranteed at the best level: no crosstalk and low intermodulation products (occurring when the signal passes through an active device by the combination of two or more tones).

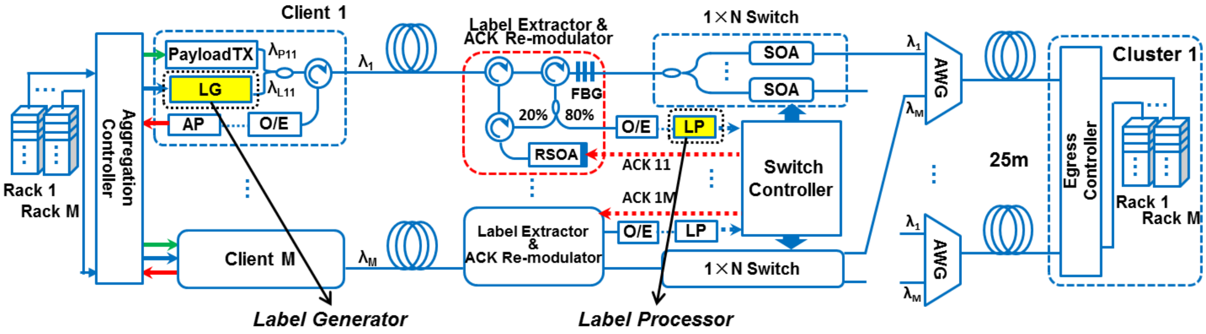

Figure 4.4: Label generator and corresponding label processor in the system

The Label Generator under design needs to work with 8 bits in parallel, on the following tones: 130MHz, 253MHz, 410MHz, 615MHz, 820MHz, 940MHz, 1189MHz, and 1400MHz, which are same with the working frequencies of the LP. The bandwidth given to each tone depends on the band-pass filter in the Label Processor and in general it is around 100MHz.

The schematic of the LG RF board and the laser driver board is reported in Fig. 4.5. The input signal consists of parallel baseband label bits which represents the destination of the packet. Main function of the LG RF board is to up-convert each bit from the baseband to RF frequency that could be carried by single optical wavelength. For this operation a mixer and an oscillator have be employed. The input power of the intermediate frequency (IF) input port has been controlled by the attenuator to equalize the power of different tones and more important, to avoid the distortion caused by the saturation. 8 modulated RF tones are then combined by a coupler (first stage: two 4/1 combiners, second stage: one 2/1 combiner). Good tone purity is especially required here that higher order components will act as noise to other tone channels. A BPF or LPF placed after the mixer may help improve the SNR. At the output of the LG RF board, the power loss are compensated with a RF amplifier that the output could reach the level required by the laser driver board. While the risk of intermodulation exists when multiple frequencies are present in the active device, it needs careful selection of the amplifier used here.

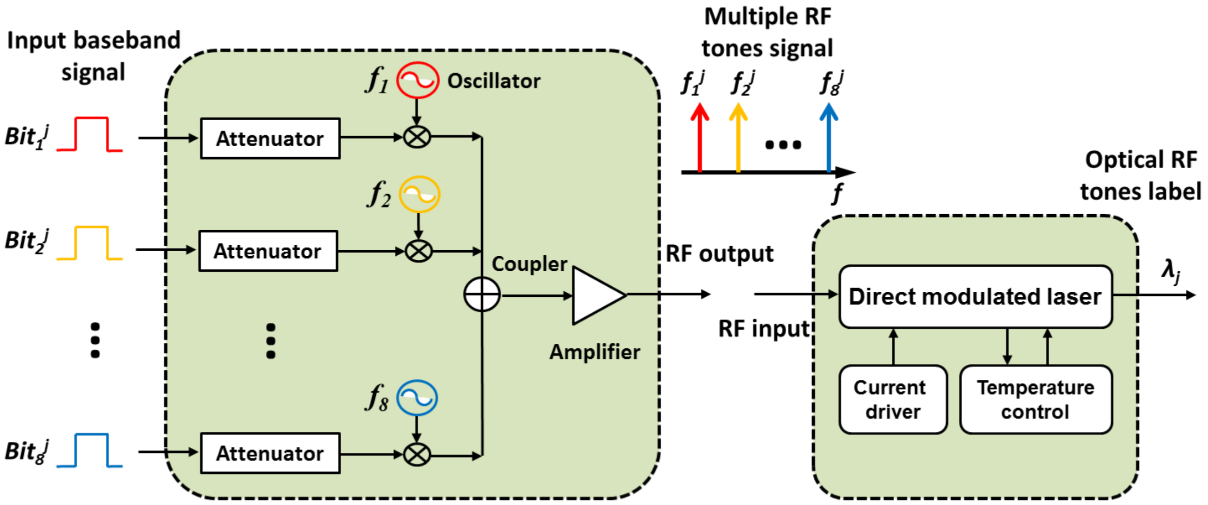

Figure 4.5: Schematic of the label generator RF board and laser driver board

The RF tones signal generated by the LG drives a direct modulated laser to finally generate the optical label signal. A driver board has been designed and implemented as shown in Fig. 4.5. Current driver supplies a DC bias to the laser while the RF tones signal is applied on the 50Ω matched input port to modulate the laser. The temperature is controlled by a specific module to stabilize the wavelength and on the other hand, to finely adjust the central wavelength to match with the optical FBG-based label extractor. Finally, the optical RF tone label signal is coupled to the optical payload and transmitted to the OPS node.

4.2. System evaluation

4.2.1. 4x4 OPS system

A novel flat DCN architecture for ToR interconnection based on a scalable optical switch system with hardware flow control for low latency operation has been proposed and a 4x4 system has been demonstrated [18]. The dynamic flow control and payload switching operation of the system have already been reported in D3.1 [6]. In this deliverable, further investigation results on packet loss, latency and scalability issues are given in detail.

The label that represents the packet's final destination is generated by the aggregation controller and stored in the finite-size FIFO queue as illustrated in Fig. 4.6. It will be released from the FIFO once the packet has been successfully forwarded. In this case the aggregation controller will receive a positive flow control signal. Otherwise, the packet will be retransmitted. However, if the FIFO is already fully occupied and there is a new packet to be served at the next time slot, this packet will be instantly dropped and considered lost due to buffer overflowing. The packet loss is then calculated as the ratio of the number of lost packets to total number of generated packets.

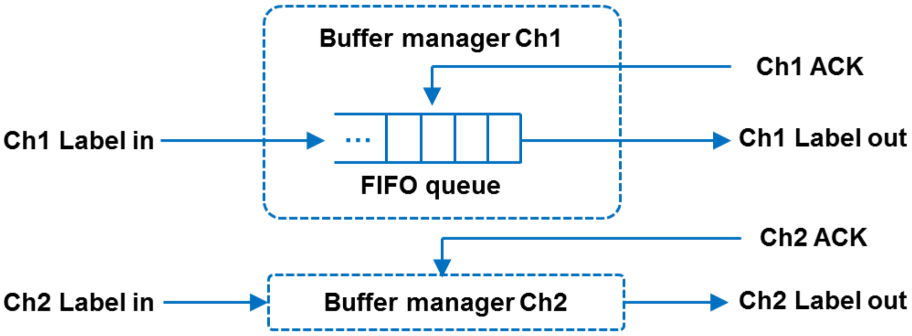

Figure 4.6: Buffer management in aggregation controller

At each time slot, the aggregation controller will generate a packet for each different client with the same average traffic load. The destinations decided by the label pattern are chosen randomly between the two possible outputs according to a uniform distribution. Based on the label information, the switch controller forwards the packets to the right output and if a contention occurs, only the packet with higher priority will be properly delivered. Instead of using a fixed priority for the contention resolution algorithm, a round robin scheme is employed as priority policy to efficiently balance the utilization of the buffer and the latency between the two clients. This means that the priority will be assigned slot by slot. As a result, a packet in the FIFO will be definitely sent to the proper destination within two time slots, and the respective buffer cell released.

Figure 4.7 (a) shows the packet loss for different input loads and buffer sizes. The total amount of time considered is 2×10^{10} time slots. As expected the packet loss increases with the input load. Larger buffer size could improve the packet loss performance for input loads smaller than 0.7. It does not bring significant improvement when the load ≥0.7 because the buffer is always full and overflowing causing high packet loss. Figure 4.7 (b) presents the buffer occupancies when traffic load equals to 0.5, 0.6, 0.7 and 1, respectively. For the first 200 time slots, it is clear that for load of 1, the 16-packet buffer is rapidly filled up and for load of 0.7 the buffer is fully occupied most of the time which will cause the buffer overflowing. Average end-to-end latency for the system with a buffer size of 16 packets is reported in Fig. 4.7 (c). The number of packets that has been successfully forwarded without retransmission and the one that has been retransmitted once are recorded and employed to calculate the average latency. The lost packets are not considered in the latency calculation. Similarly to the packet loss curves, the average latency increases approximately linearly for input loads up to 0.7. As the traffic becomes heavier, the possibilities of contention also increase which results in more retransmissions, and thus larger latencies. However, when the load is higher than 0.7, the buffer is always full but the average latency remains constant since the round robin policy and the lost packets are not considered in the latency calculation. Indeed, due to round robin policy, every packet having entered in the buffer queue will finally win the contention within two time slots. This explains the saturation of the latency curve at 645 ns which includes 250 ns off-set latency caused by the 25m transmission link. Figure 4.7 (d) shows the average retransmission rate which represents the contention probability as a function of the input load. It is calculated as the ratio of retransmissions to the total number of transmitted packets. The retransmission rate curve keeps the same shape as the latency one and saturates when the input traffic load exceeds 0.7 in which case the actual traffic inside the switch is reaching the maximum due to the retransmissions. From Fig. 4.7 it can be concluded that the system could handle an input load up to 0.5 providing a packet loss lower than 10^{-5} and an average end-to-end latency lower than 500 ns.

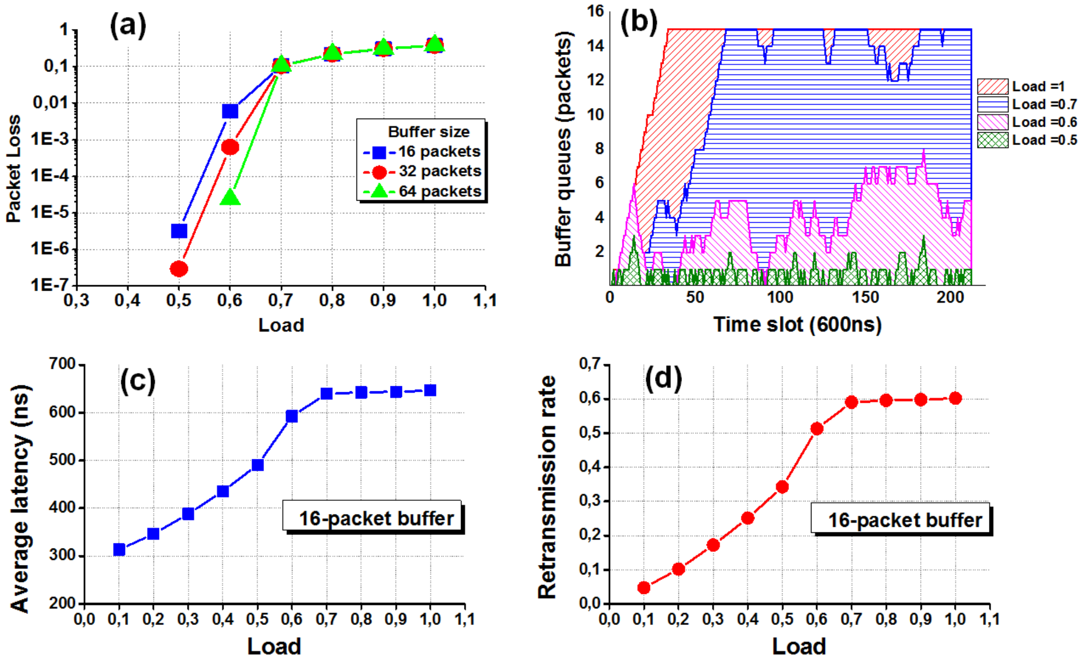

Figure 4.7: (a) Packet loss vs. load with different buffer size. (b) Buffer queue occupancy for different input load. (c) Average latency with buffer size of 16 packets. (d) Retransmission rate with buffer size of 16 packets.

We further investigate the system scalability in consideration of supporting a large port count. The total number of ports supported by the OPS is given by N×M because of the presence of N modules and M clients in each module. The performance of the overall system could be translated into the performance of 1×N optical switch due to the identical structure of N modules. In this scenario, the main limiting factor for scaling the OPS is the splitting loss experienced by the payload caused by the 1×N broadcast and select stage. Therefore we employed a variable optical attenuator (VOA) to emulate the splitting losses, as schematically reported in Fig. 4.8 (a). At the output of the SOA switch, the BER and the OSNR are measured to evaluate the payload quality.

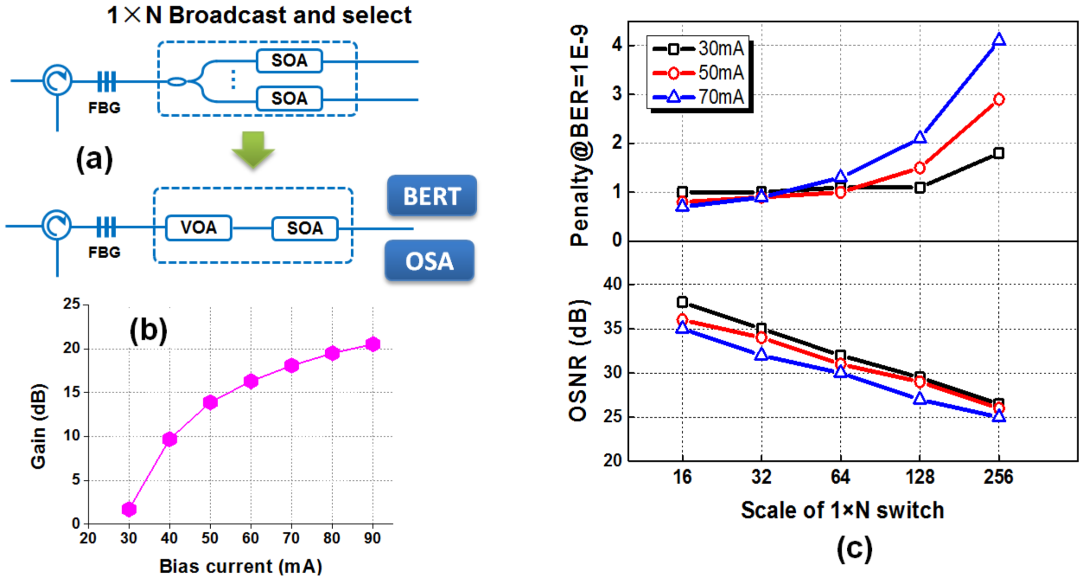

Figure 4.8: (a) Set-up for scalability investigation. (b) Gain characteristic with bias current of the SOA switch (c) Penalty and OSNR vs. scale of 1×N switch.

The input optical power of the 1×N optical switch is 0dBm and the attenuation caused by the VOA is set to be 3dB×log$_2$N. The SOA will be switched on to forward the packet, and at the meantime amplifies the signal. Fig. 4.8 (b) gives the gain characteristic versus bias current of the SOA from which we could see that the SOA operates transparently at 30mA and 18dB amplification could be supplied when biased at 70mA. Considering the splitting loss, the SOA could compensate the 18dB loss caused by the 1×64 broadcast stage resulting in a lossless 1×64 optical switch. Fig. 4.8 (c) shows the power penalty (measured at BER=1E-9), and the OSNR of the switched output as a function of N for different SOA bias currents. A penalty of < 1.5 dB for N up to 64 is measured regardless of the bias current of the SOA. For N > 64 the penalty increases mainly caused by the deterioration of OSNR as a result of splitting loss. The BER performance gets worse when biasing at a higher current due to noise that becomes more prominent. The results clearly shows that the OPS under investigation could be potentially scaled up to a large number of ports at the expense of limited extra penalty and a lossless system without extra amplification could be achieved with the bias current of 70mA.

4.2.2. Label processor

Numerical investigation of the LP has been conducted by using sophisticated circuit simulator Advanced Design System (ADS). To have a realistic perspective of the real behaviour of the RF tones processor, the generated 8 RF tones are sent through the optical link and saved after the O/E convertor in a .csv file. Then it is loaded into ADS simulator which contains the implemented circuit with commercial component model and the matched transmission lines, to simulate the performance of the LP.

First, the dynamic range for each tone channel is investigated with various input power. Eight different modulated tones with the frequencies of 130, 253, 410, 615, 820, 940, 1189 and 1400MHz are generated with amplitude changing from 6mV to 200mV and sent into the respective channel. Figure 4.9 (a) shows the recorded output level for 8 channels with different input power. It could be observed that input signals smaller than -20dBm produce a very small output for last 4 tones. From -16dBm the curves are quite linear and the differences among the 8 channels are quite small. Figure 4.9 (b) shows the range in which the threshold could work without making errors, named with detectable range. For input power less than -20dBm the output of the last four channel is unreadable (<-10dBm) and at least -16dBm of input power is needed to have a readable output signal.

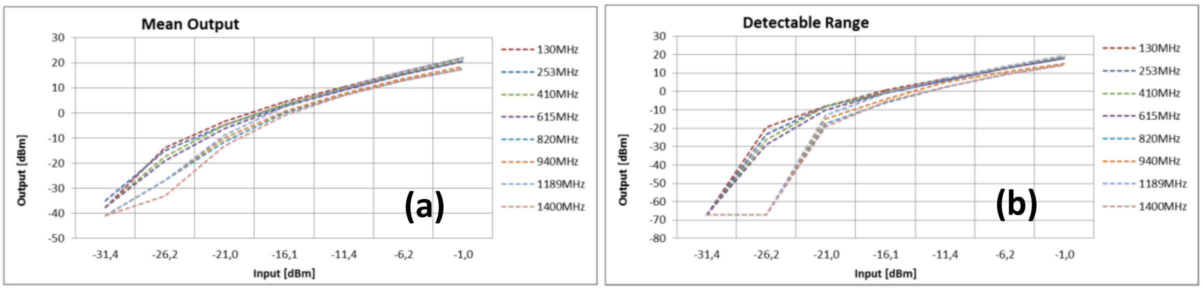

Figure 4.9: Output power (a) and detectable range (b) for each channel

Then, the simulation on the overall system performance has been investigated, by varying the amplitude of the combined 8 tones signal from 110mV to 400mV. To have more accurate results, we use the signal saved in a real case and inserted it in a special source block in ADS. As shown in Fig. 4.10, The multiple RF tones signal, generated by the AWG, is firstly amplified and then employed to modulate a CW laser at 1550 nm by using a 10 Gb/s Mach-Zehnder optical modulator. The optical output of the modulator (the optical label) is transmitted within the optical payload to reach the OPS node. At the receiver side, the optical label is converted to an electrical signal by using a photo detector. The detected electrical signal, that includes the RF tones, is recorded by an oscilloscope (in CSV format). It will be used as an input of the simulation set-up in ADS to simulate the performance of LP including the distortion and interference effect caused by the E/O and O/E conversion. The amplitude of the recorded signal as the input of the simulation is 500mVpp.

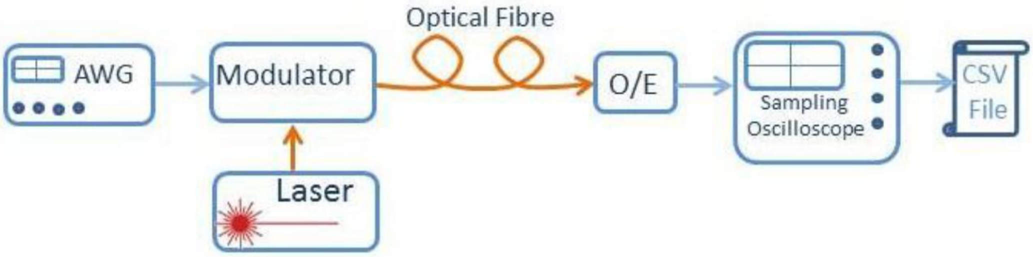

Figure 4.10: Experimental set-up used to generate the input RF tones signal for the simulation

To analyse the crosstalk between the channels, sixteen combinations of tones were used. In this way the first channel has the same pattern of the fifth, the second channel has the same pattern of the sixth, and so on. It was very difficult generate all possible combinations for the eight tones, mainly for the simulation time required. The amplitude of the input signal to the label processor is, both in the electrical set-up and optical one, 500mVpp.

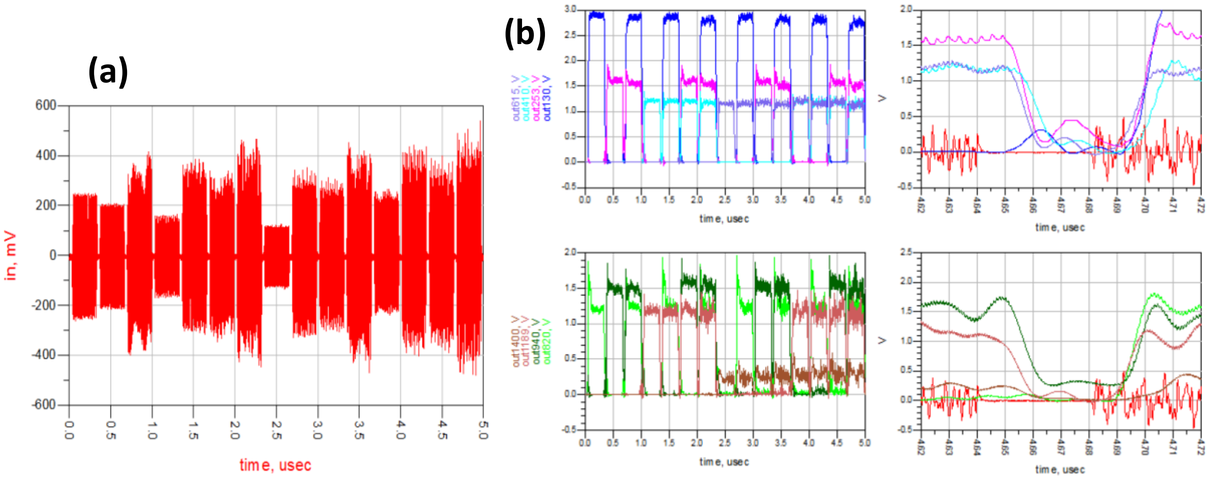

Figure 4.11: (a) Input 8 tones label after optical link; (b) 8 output signals

Figure 4.11 (a) presents the waveform of input 8 tones label signal. The output of each channel and a zoom-in showing the rise/fall time and delay are given in Fig. 4.11 (b). We can observe more crosstalk between the channels and more ripples manly for the second group of the four channels. The eighth channel has very small amplitude making it quite unreadable. This is mainly due to the simulation model of the second 1x4 splitter is unavailable that the In the simulation we have used the model of the 1x4 splitter with bandwidth ranging from 10MHz to 1GHz. The delay is around 20ns and the rise/fall time is less than 15ns.

After the simulation work, the design and the fabrication of the PCB of the LP is conducted. The PCB layout designed with ADS is then turned into an industry standard Gerber file which is sent to a PCB production company. The PCB layout and the prototype board are shown in Fig. 4.12. Considering ten amplifier consuming 52.8mW (16mA at 3.3V) and the power dissipation of the resistors in the polarization network for each amplifier, the total power consumption is around 2W, given that the board absorbs 170mA at 12V.

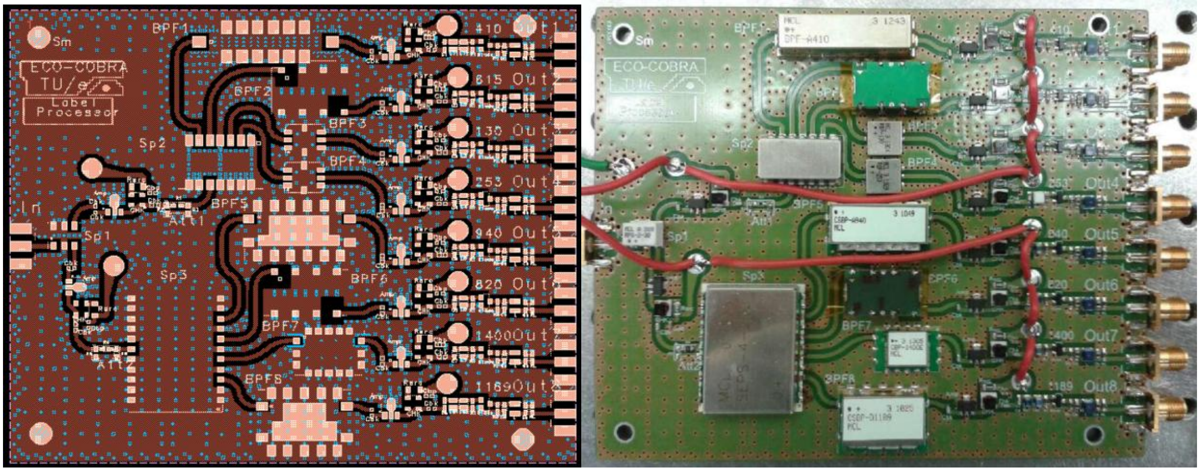

Figure 4.12: PCB layout and the prototype of the LP

Similar with the simulation work, the PCB board is tested with RF tones label signal and the results are compared with the ones achieved in simulation. The AWG generates the multiple eight tones signal with the input power ranging from 1dBm to 19dBm. The patterns of the tones and the outputs for 7dBm total input power are reported in Fig. 4.13. To improve the sensitivity and the amplitude dynamic range, we changed the first two amplifiers Gali-S66+ with Gali-49+. The two amplifiers have the same footprint but the later one has higher 1dB compression point. In the simulation Gali-S66+ works fine because the model provided by the constructor only includes linear effect and actually, it generates more noises due to the easy saturation and output power is limited. The Gali-49+ has a lower gain and for this reason we have observed some differences between the output signal level of the simulated circuit and the measured one. Moreover, the impedance difference of the oscilloscope (50Ω) and the one used in the simulation (8KΩ) also explains the lower sensitivity of the PCB test.

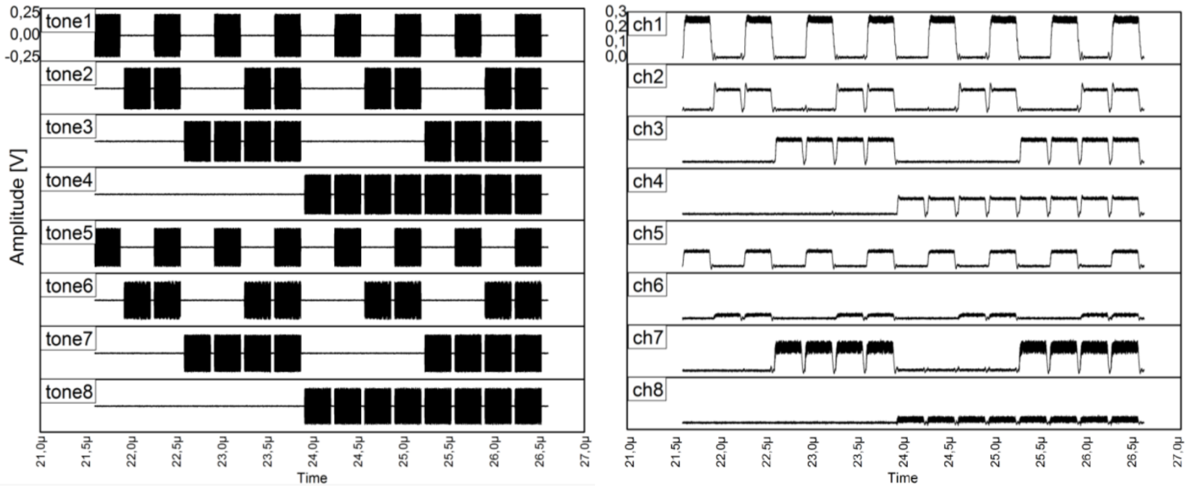

Figure 4.13: Pattern of the input signals for each tone and the output for 7dBm input power

The detectable range with different input power is given in Fig. 4.14. Sixth and eighth channels have lower output than others mainly due to the mismatch of the impedance. With a total input power of 7dBm, 6 channels have a detectable range larger than 100mV. Up to 13dBm, all channels experience a decrease of the detectable range caused by the saturation of the amplifiers.

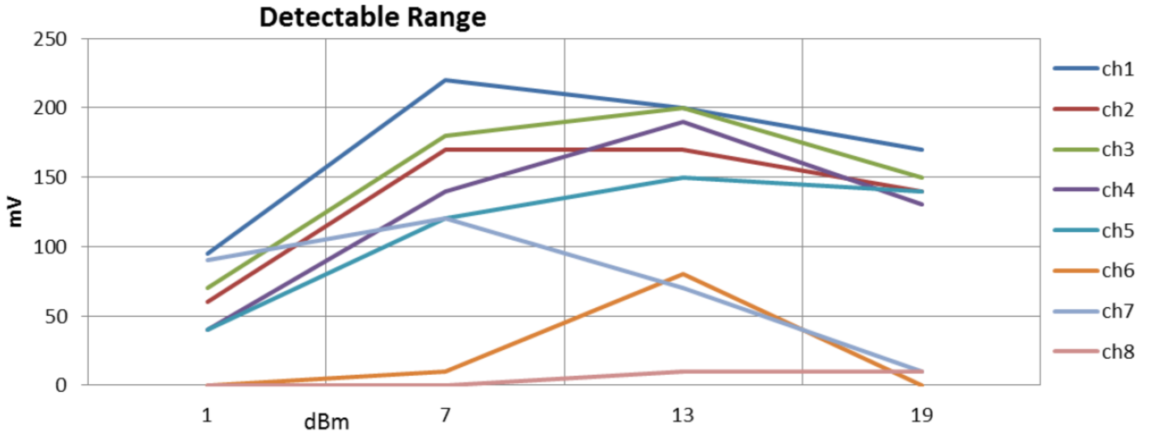

Figure 4.14: Detectable range in millivolts for multiple eight signals

As the most important parameters, the delay and rise/fall time are perfectly in line with the simulations. The delay is around 20ns, the rise time ranges from 10ns to 15ns and the fall time ranges from 12 to 16ns. Being the first prototype of the LP, the board works well with six channels. Note that the issues with the two worst channels are independent of number of tones because it also occurs when the single tones are tested. Possible solution is to improve the impedance matching network design. The dynamic range and detectable range are in the order of tens decibels and hundreds of millivolts, respectively. If we do not consider the extra power required to make the two worst channels detectable, we measured around 7 dB of difference with the dynamic range observed in the simulations. This is mostly due to the lower gain of the amplifier employed in the experiments.

4.2.3. Label generator

The baseband label signal is up-converted to a certain RF tone and then combined together to form the RF tones label. As the key parameter of evaluating the quality of the RF tones label, the tone purity will be greatly affected by the crosstalk generated by the mixer. Two types of crosstalk may exist. First one comes from the nonlinear effect that the IF port input power should be lower than the 1dB compression point. A variable attenuator has been exploited to adjust the power of the baseband signal to guarantee the low distortion. The other one is overlapping of sideband signal. As the nonlinear model of mixer is not available and input power could be controlled by the attenuator, here we mainly show the simulation results on the effect caused by sideband overlapping. Figure 4.15 shows the spectrum of the generated RF tones label by using ADS simulation.

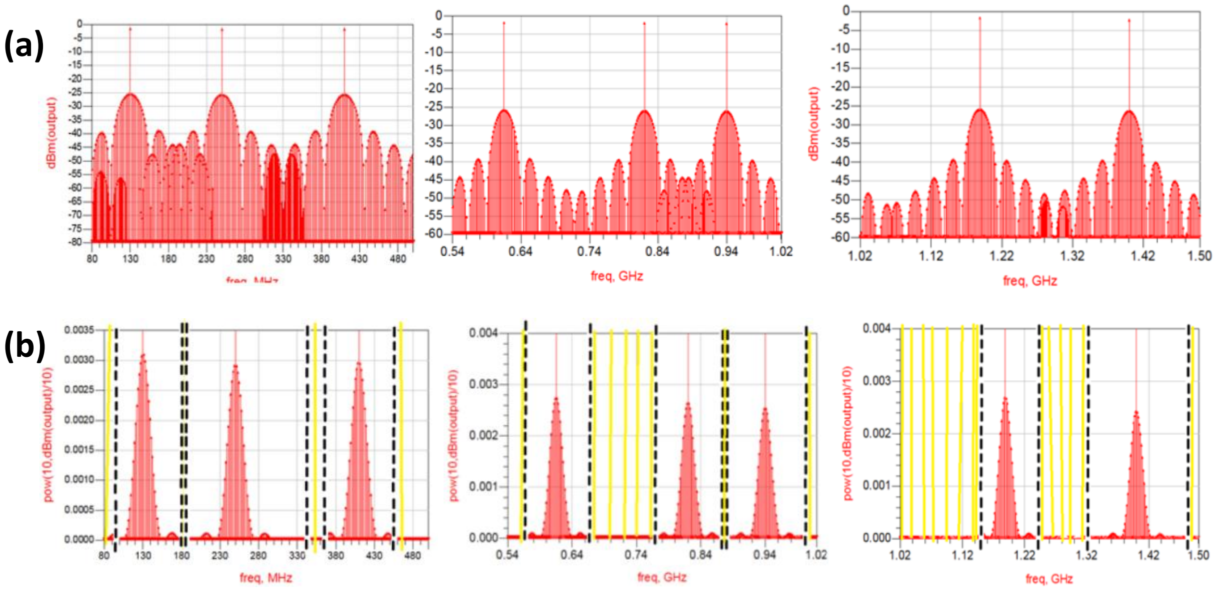

Figure 4.15: Spectra of the 8 RF tones obtained by Harmonic simulations: (a) value in dBm; (b) value in linear

In this case, the duration of the label signal is 600ns and the guard time is 40ns. The output signal-to-crosstalk-ratio is at least 29dB for each tone. The suppression of the noise is good enough even though the decrease of the duration will slightly deteriorate the performance. To obtain a better signal-to-crosstalk-ratio (lower crosstalk power), it would be necessary to use some band pass filters (i.e. the same used in the Label Processor for each tone), placed between the mixers and the power combiner to filter out the sideband noise and the high-order components generated by the mixer. But it would result in an increase of the delay and rise/fall time, so for the first prototype of LG, the BPF is only applied to one channel to facilitate the further evaluation.

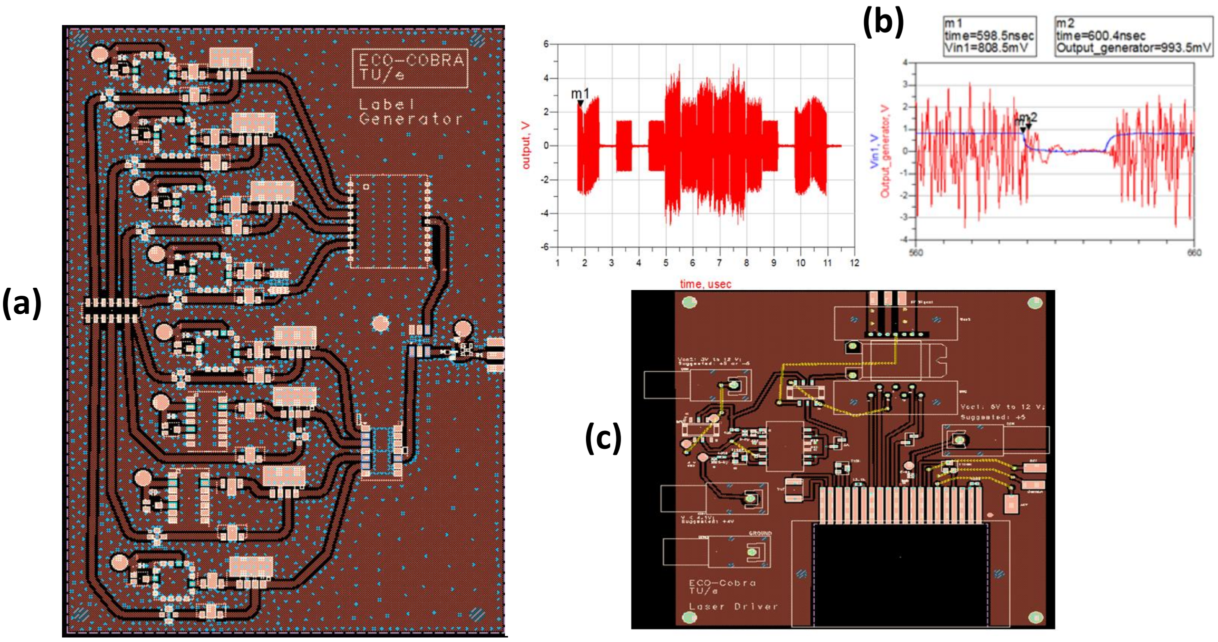

Figure 4.16: (a) PCB layout of label generator; (b) time traces of LG output in ADS simulation; (c) PCB layout of laser driver

The final PCB layout has been given in Fig. 4.16 (a). Figure 4.16 (b) shows the time traces of the output simulated in ADS. The blue curve in the bottom figure is the one of the input baseband signals. It could be seen that the guard time is still well defined in the output traces. The delay is lower than 5ns since no filters have been deployed in the design. The layout of the laser driver board is given in Fig. 4.16 (c). It would receive the output of the LG and modulate it on the output of the laser which will be placed on the board. A DC bias will be provided and the temperature will be precisely controlled. Both LG and laser driver boards are yet under test and the evaluation results of the PCB will be reported in next deliverable.

4.3. Interface with the control plane

The implementation scheme of the interface with the control plane is proposed in D3.1 [6]. There, an agent lying on top of OPS node bridges the control plane (i.e. the SDN controller) and the switch controller. Along with the southbound interface between the control plane and the control agent, which uses OpenFlow protocol for communication, the interface implementation between the agent and the switch controller is the key enabler to fulfil the SDN enabled control framework. In this way, the agent configures the OPS node by translating the OpenFlow messages coming from the SDN controller to the appropriated (and proprietary) set of commands of the agent-to-switch controller interface.

In light of the above, Fig. 4.17 depicts the general architecture of the OPS agent and its connectivity to both the data and the control planes. As said, the agent communicates to the OPS node via a USB interface using a proprietary protocol. On the other side, it communicates to the SDN controller by means of an extended version of the OpenFlow protocol [11]. Besides, the agent keeps an information model of the OPS switch to facilitate its management.

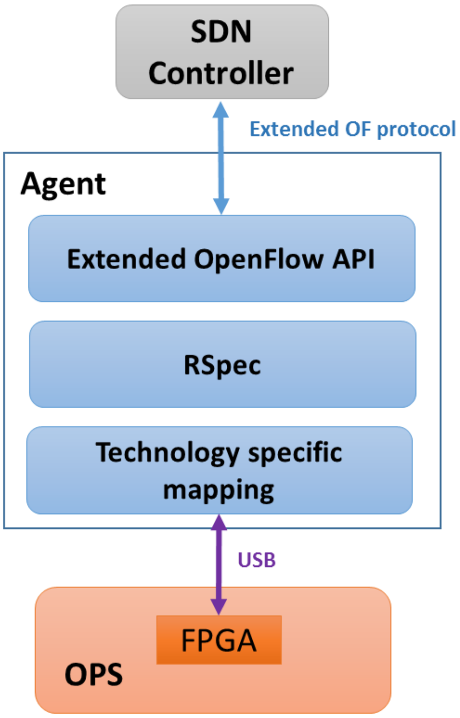

Figure 4.17: OPS switch OF agent

The switch controller implemented in the FPGA supports data communication with a host PC (running the OF-based agent) through a USB or PCI port. Although the use of the PCI interface is being studied, the current solution to communicate the switch controller and the agent is implemented over the USB interface. A proprietary protocol has been designed to this end. This protocol implements the functions needed by the agent to write and read the LUT residing inside the OPS switch controller and, thus, to manage OPS-based data flows. Moreover, statistics such as average load and contention rate can be collected through this interface and sent to the SDN controller. Table 4.2 details the functions that implement the USB communication interface.

ID	Function	Description
1	addLUTEntry	Add a new entry to the LUT. The input module, the label are specified. This function is called previous to addFlow to configure a new OPS flow.
2	deleteLUTEntry	Removes an existing entry of the LUT.
3	clearLUT	Removes all the existing entries of a given input module.
4	clearAllLUT	Removes all the entries in the LUT.
5	addFlow	Add a new flow associated to an existing entry. The Load of the flow and wavelength to be used are specified.

6	deleteFlow	Removes an existing flow.
7	modifyFlow	Modifies the parameters of an existing flow.
8	clearPortFlows	Removes all the flows associated to a given input port
9	clearModuleFlows	Removes all the flows associated to a given input module.
10	clearAllFlows	Removes all the flows in the OPS node.
11	getPortCounters	Collects the statistical information of a given input port.
12	getModuleCounters	Collect the statistical information of a given module.

Table 4.2: OF protocol extensions to support extended OPS switch monitoring

In summary, the OF agent implements the interface between the OPS node and the control plane. This agent is the responsible entity for configuring the OPS node by translating the OpenFlow messages coming from the SDN controller through the southbound interface to the appropriated set of commands of the proprietary protocol implemented between the agent and the physical OPS node.

5. Conclusions

This document presents the detailed implementation scheme and evaluation results of LIGHTNESS data plane. Stick to the design proposed in previous deliverable D3.1, the switching nodes including hybrid NIC, optical ToR, OCS as well as modular OPS are further investigated and developed. The interfaces consisting of optoelectronic circuitries as well as electronic control interfaces are investigated for realizing fast and accurate control mechanisms in up-coming tasks in WP3. In addition, to satisfy the requirements of SDN-based control plane framework, the interfaces used for communications with control plane have been developed for each switch fabric type.

The Hybrid NICs aggregate the traffic and intelligently perform the switching over function between the OCS and OPS. Interfacing with advanced optical transponders, the NIC and optical ToR have various connecting scenarios which have been discussed. The AoD architecture provides the interconnectivity among ToRs. Moreover, it would be an appropriate solution for the inter-cluster communication. The power budget and crosstalk have been evaluated in consideration of different modulation formats. Label processor and label generator in OPS node have been implemented with PCB design and the performance has been numerically and experimentally evaluated. The OpenFlow agents implementing protocol extension are also presented to demonstrate the communication and interaction with the control plane. The results on data plane implementation contained in this deliverable will greatly support the development of future research activities in WP3 and WP4.

6. References

[1] T. Benson, A. Akella, D. A. Maltz, “Network traffic characteristics of data centers in the wild,” in Proceedings of 10th annual conference on Internet measurement (IMC). (ACM, New York, 2010), pp. 267-280.

[2] S. Sakr, A. Liu, D. Batista, and M. Alomari, “A survey on large scale data management approaches in cloud environments,” IEEE Commun. Surveys & Tutorials 3(13), 311–336 (2011).

[3] Cisco Data Center Interconnect Design and Deployment Guide. Cisco Press, 2010.

[4] L. A. Barroso and U. Hölze, “The datacenter as a computer: an introduction to the design of warehouse-scale machines,” Morgan and Claypool Publishers, Los Angeles, 2009.

[5] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. S. E. Ng, M. Kozuch, and M. Ryan, "c-Through: Part-time Optics in Data Centres," in ACM SIGCOMM'10, 2010.

[6] LIGHTNESS Deliverable D3.1 “Release of the design and early evaluation results of OPS switch, OCS switch, and TOR switch”.

[7] Amaya, N., G. Zervas, and D. Simeonidou. 2011. “Architecture on Demand for Transparent Optical Networks.” In 2011 13th International Conference on Transparent Optical Networks (ICTON), 1-4. doi:10.1109/ICTON.2011.5970836.

[8] http://www.polatis.com/polatis-series-6000-optical-matrix-switch-192x192-sdn-enabled-industry-leading-performace-lowest-loss-switches.asp.

[9] http://www.finisar.com/products/optical-instrumentation/WaveShaper16000S.

[10] LIGHTNESS Deliverable D4.3 “The LIGHTNESS network control plane interfaces and procedures”.

[11] LIGHTNESS Deliverable D4.2 “The LIGHTNESS network control plane protocol extensions”.

[12] T. Benson et al.,“Understanding Data Center Traffic Characteristics,” ACM SIGCOMM Workshop: Research on Enterprise Networking (2009).

[13] S. Tibuleac and M. Filer, “Transmission impairments in DWDM networks with reconfigurable optical add-drop multiplexers,” J. Lightwave Technol. 28(4), 557–598 (2010).

[14] J. Luo, S. Di Lucente, J. Ramirez, H. J. S. Dorren, and N. Calabretta, “Low latency and large port count optical packet switch with highly distributed control,” in Optical Fiber Communication Conference, Technical Digest (CD) (Optical Society of America, 2012), paper OW3J.2.

[15] W. Miao, S. Di Lucente, J. Luo, H. Dorren, and N. Calabretta, “Low latency and efficient optical flow control for intra data centre networks,” Optics Express, Vol. 22, no. 1, p. 427-434 (2014).

[16] N. Calabretta, W. Wang, T. Ditewig, O. Raz, F. Gomez Agis, S. Zhang, H.de Waardt and H. Dorren, “Scalable optical packet switches for multiple data formats and data rates packets,” IEEE Photonics Technology Letters, 22(7), 483-485 (2010).

[17] J. Luo, H. J. S. Dorren and N. Calabretta, “Optical RF tone in-band labeling for large-scale and low-latency optical packet switches,” J. Lightw. Technol., 30(16), 2637-2645 (2012).

[18] W. Miao, J. Luo, S. Di Lucente, H. Dorren, and N. Calabretta, “Novel flat datacenter network architecture based on scalable and flow-controller optical switch system,” Optics Express, Vol. 22, no. 3, p. 2465-2472 (2014).

7. Acronyms

ADS	Advanced Design System
AoD	Architecture on Demand
APD	Avalanche Photodiode
API	Application Programming Interface
AWG	Arrayed Waveguide Grating
BPF	Band Pass Filter
CFP	C Form-factor Pluggable
DC	Data Centre
DCN	Data Centre Network
DEMUX	De-multiplexing
DWDM	Dense Wavelength Division Multiplexing
ECL	External cavity laser
EDFA	Erbium doped fibre amplifier
FBG	Fiber Bragg Grating
FMC	FPGA Mezzanine Card
FPGA	Field-Programmable Gate Array
IF	Intermediate Frequency
LG	Label Generator
LP	Label Processor
LPF	Low Pass Filter
LUT	Look-Up Table

MUX	Multiplexing
NIC	Network Interface Card
OCS	Optical Circuit Switching
OE	Optical-to-Electrical
O-E-O	Optical-to-electrical-to-optical
OF	OpenFlow
OPS	Optical Packet Switching
PCB	Printed Circuit Board
PM	Polarization multiplex
PS	Power splitter
QoS	Quality of Service
RSOA	Reflective Semiconductor Optical Amplifier
SDN	Software Defined Networking
SFP	Small form-factor pluggable
SNMP	Simple Network Management Protocol
SOA	Semiconductor Optical Amplifier
TL1	Transaction Language 1
ToR	Top of Rack
VCSEL	Vertical-cavity surface-emitting laser
VOA	Variable optical attenuator
WC	Wavelength Converter
WDM	Wavelength Division Multiplexing
WSS	Wavelength selective switching