Grant Agreement No. 619572

COSIGN

Combining Optics and SDN In next Generation data centre Networks

Programme: Information and Communication Technologies

Funding scheme: Collaborative Project – Large-Scale Integrating Project

Deliverable D2.2

1st generation TOR switch operation based on 10 Gbps base rate components

Due date of deliverable: June 30th, 2015
Actual submission date: June 30th, 2015

Start date of project: January 1, 2014 Duration: 36 months

Lead contractor for this deliverable: TU/e

Project co-funded by the European Commission within the Seventh Framework Programme		
Dissemination Level		
PU	Public	X
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)	

Executive Summary

This document describes the new type of Top-Of-Rack (TOR) switch developed and built by the TU/e which uses mid-board optics instead of edge of card pluggable modules. In this document we focus on the design and implementation stages of the HW & SW. The TOR has been designed and built based on the Trident I switch ASIC and supports the connection of up to sixty four 10Gbps Ethernet links. The fabricated prototype has been tested and demonstrates very good signal integrity on the optical outputs as well as radical drop in energy use (>50% savings) compared to standard designs based on the same chip set, employing edge of card interfaces. The TOR also supports OpenFlow 1.3 agent which allows the switch to communicate with OpenDayLight controller SW which is the COSIGN selected controller software platform.

In this report we include results of the design process indicating the low power consumption and the obtained signal integrity obtained on a low cost PCB material FR408. The layer 2 connectivity was tested for both 10 and 40Gbps aggregate bit rates with no packet loss with the use of dedicated packet analysis test equipment.

The conclusion from the 1st generation TOR is that mid board optics offer substantial benefits in all major aspects of switch design: cost, size and energy consumption. The results of this work will be used in the design and fabrication of the 2nd generation TOR which is due to be ready for testing by M24. This version will offer double the bandwidth of the 1st generation TOR in a reduced foot print. We consider further shrinking the foot print with the aid of mid-board optics a major enabler for making more compact and lower total cost of ownership (TCO). The major inhibiting factor in the adoption of optically enabled TORs with mid-board optics, rather than edge of card opto-electronic conversion, is still the assumed lower reliability of opto-electronic components. It is the role of the opto-electronics industry to change this notion. Initial steps to create a multi-source agreement to create standardization of mid-board optics have been recently launched by a large consortium (http://cobo.azurewebsites.net/about.html#). It is hoped that this initiative will support the trend to shift to mid-board optics for data center switches.

Document Information

Status and Version:	1.0	
Date of Issue:	30/06/2015	
Dissemination level:	Public	
Author(s):	**Name**	**Partner**
	Oded Raz	TUe
Edited by:	Oded Raz	TUe
Checked by :	Sarah Ruepp	DTU
	Georgios Zervas	UNIVBRIS
	Bingli Guo	UNIVBRIS
	Amaia Legarrea	I2CAT

Table of Contents

Executive Summary **2**

Table of Contents **4**

1 Introduction **5**

1.1 Reference Material 5
1.1.1 Reference Documents 5
1.1.2 Acronyms and Abbreviations 5

1.2 Document History 5

2 Switch HW design **6**

3 Switch OpenFlow integration **10**

4 Conclusions **11**

1 Introduction

Traditional use of edge of card transceivers for electronic to optical conversion is a limiting factor in the scaling of data centre networks and subsequently, optical interconnections are being introduced to provide increased bandwidth, flexibility, low latency and scalability in DC networks. In the COSIGN project the world's first high-end electronic switch engine with mid board optical modules placed within few centimetres from the switching ASIC has been designed and built. Several attempts to bring high speed switching and optical interconnects together have been reported in the industry (most noticeably by http://www.compassnetworks.com/technology/). In this demonstration proprietary silicon with unique 3D stacked opto-electronic was useds. A prototype based solely on commercially available components is presented by COSIGN, with which a 50% reduction in power consumption and 75% foot print reduction have been achieved compared to a reference design.

This document outlines the main results of the research effort executed by COSIGN consortium in terms of both HW and SW integration.

1.1 Reference Material

1.1.1 Reference Documents

[1]	COSIGN FP7 Collaborative Project Grant Agreement Annex I - "Description of Work"

1.1.2 Acronyms and Abbreviations

Most frequently used acronyms in the Deliverable are listed below. Additional acronyms can be specified and used throughout the text.

ASIC	Application Specific Integrated Circuit
BER	Bit Error Rate
BOM	Bill of Materials
CXP	100Gbps 10x10Gbps pluggable optical module
EOL	End of Line
FR-4	Flame Retardant
HW	Hard Ware
PCB	Printed Circuit Board
PHY	ICs (usually ASICs) used to re-shape, re-time or regenerate high speed signals on PCB
PRBS	Pseudo Random Bit Sequence
QSFP	40Gbps 4x10Gbps pluggable optical module
RU	Rack Unit
SerDes	Serializer Deserializer
SW	Software
TOR	Top of Rack

1.2 Document History

Version	Date	Authors	Comment
00	20/06/2015	See the list of authors	first draft integrated version ready

2 Switch HW design

High end top of the rack (TOR) data center switches must be able to switch large number of high speed lanes and offer good connectivity and performance. Optical technologies already offer the most suitable cabling technologies for bit rates of 10Gbps over several meters. Commercially available switching solutions rely on 19" wide and 1RU switching systems with several tens of different pluggable transceivers on their front panel. Using the current generation of optical transceiver modules, the front panel of a high end switch is fully occupied as can be seen in figure 2-1.

Figure 2-1: View of interface masters TOR switch front panel with sockets for 60 pluggable transceivers

The use of pluggable front panel transceivers leads to increase cost and power consumption and ultimately limits the amount of bandwidth which is accessible on the front panel. The main logic behind the design effort of the new TOR is to solve the limitation mentioned above. The proposed solution aims to move on from the usage of front panel optical transceivers focusing on the use of optical module integrated at a minimal distance from the switching ASIC. The result of this design choice has clear effect on the resulting TOR switch as can be seen from figure 2-2. In the figure, the newly fabricated switch and the reference design switch are shown side by side. Table 1 summarizes the main features of the TOR prototype developed by TU/e

Table 2-1: Main features of 1st generation TU/e TOR

	TU/e 1st generation TOR
Size	22x28cm
Power consumption	At maximum load 95 Watts (75 w/o optical interfaces)
Number of optical interfaces	6 (each one including 12 Tx and 12 Rx lanes)
Cost	confidential
SW features	Supports OF 1.3.4

Figure 2-2: Reference design compared to TU/e TOR design

Cost, power consumption and size are the three driving considerations behind the choice of TOR technology and its design. These three parameters played a key role in the implementation.

- **Cost:** the board has been designed to be ~75% smaller than the original reference design. This saves on the cost of the PCB. On top of the basic savings in PCB material due to the total area of the PCB, the placement of the optical transceivers in close proximity to the switch ASIC allowed the use of low cost FR408 PCB material instead of more expensive NELCO N4000-13 PCB material. In fact it is estimated that due to the very short distance even standard FR4 (which is much cheaper) could be used. A low cost CPU controller was introduced in the design and the number of optical modules has been reduced from 18 to 6 by moving to a denser optical transceiver form factor (CXP to replace QSFP). The implementation of these cost savings results in a 50% decrease in cost of BOM.
- **Power:** the board's smaller size allows for drastic shortening of the transmission lines carrying signals at 10Gbps from the SerDes-es in the ASIC to the input pins of the transceivers. Consequently, it is possible to eliminate the PHY's inserted on the transmission lines between the ASIC and the transceivers resulting in a 50% reduction of the power consumption of the new TOR switch. This is mostly due to the removal of the 18 PHYs each using 3W (54W in total) and the lowering of signalling levels on the transceivers by 50% (~1W power saving per 4x10Gbps transceiver ports and SerDes warpcore) for 18 warp cores (18Watts in total)
- **Size:** as data centers grow in size and number of servers, the data center network also has to grow. This implies that there is a need to increase the number of switches required to increase the capacity of the network. In order to simplify cabling of data center networks, often, many of the EOL (End of Line) and cluster switches are co-located. In that respect, the size of a single switch determines the eventual floor space needed to accommodate the switching infrastructure. The TOR switch developed in COSIGN is unique in the sense that it measures roughly a quarter of the size of a standard 19" 1RU switch. **This means that, in principle, one can save 75% of the floor space intended for the switching fabric!** In addition, reducing the size of the switches allows to pack them more densely. In that case, using shorter electronic cables for interconnecting switches reduces cabling complexity and possible cost while improving performance. An image showing two TOR switch side by side in a single 1RU 19" box is given in figure 2-3.

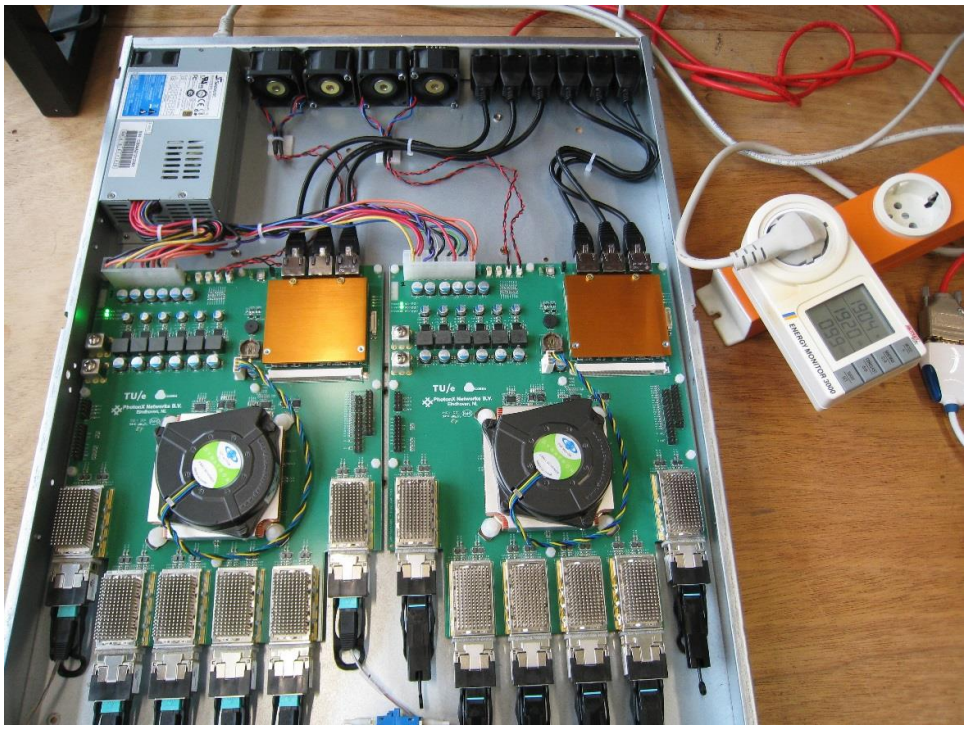

Figure 2-3: Two TOR boards size by side in a 1RU 19" box. Using true mid-board optical modules (instead of CSP) and placing the power supply outside the box should allow to place up to 4 TOR switches in one 1RU box

The board was tested during the design and prototype production stages. The first stage aimed to ensure that the power supply operates as specified by the ASIC manufacturer and provides the right currents for the right voltages in the correct sequence. A partially assembled board with only the power supply connected is shown in Figure 2-4 (Left).

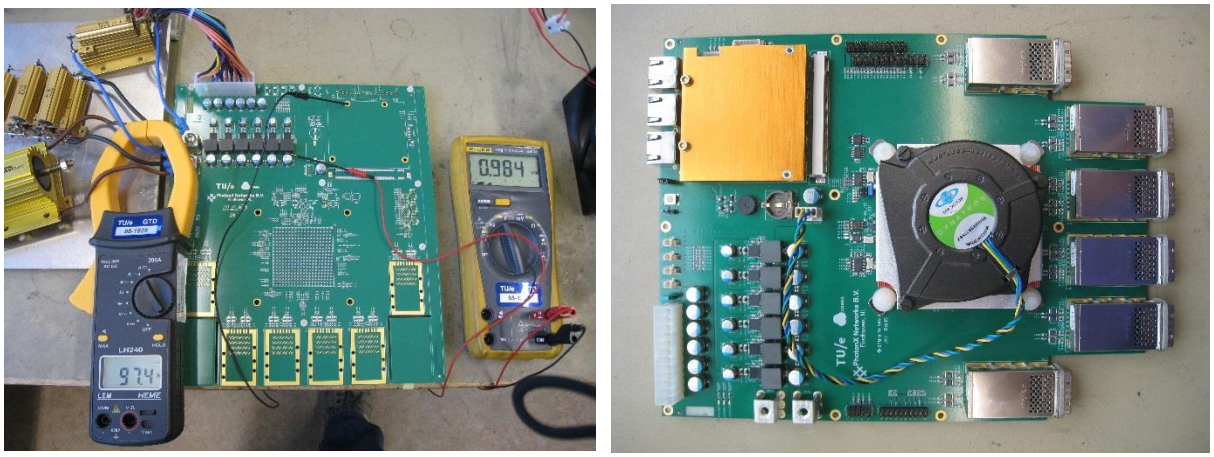

Figure 2-4: (Left)TOR power supply circuit under test;(Right) fully assembled

Two power meters are shown in the figure. The power meter on the right is measuring the output voltage of the 1V power supply. The power meter on the left shows the amount of current being drawn from this voltage source. Even when the supply circuit is providing 97.4 Amps the voltage drops by only 16mVolts (which is within spec). Once it was established that the power supply was operating as expected a 2nd board was fully assembled (see figure 2-4, right). Beside the SW challenges, one of the main points of concern in the design was signal integrity since the switch made use of FR408 and not Nelco PCB material as requested by the manufacturer. Once the SW was up and running all the ASICs output ports after the electronic to optical conversion (the optical outputs) have been tested. Figure 2-5 shows the eye patterns of all 64 channels at the output of the optical modules. In addition, optical loop–backs were used on the output of the CXPs using internal PRBS generators inside the ASIC to estimate eye margins. A typical result is given in table 2-2 below:

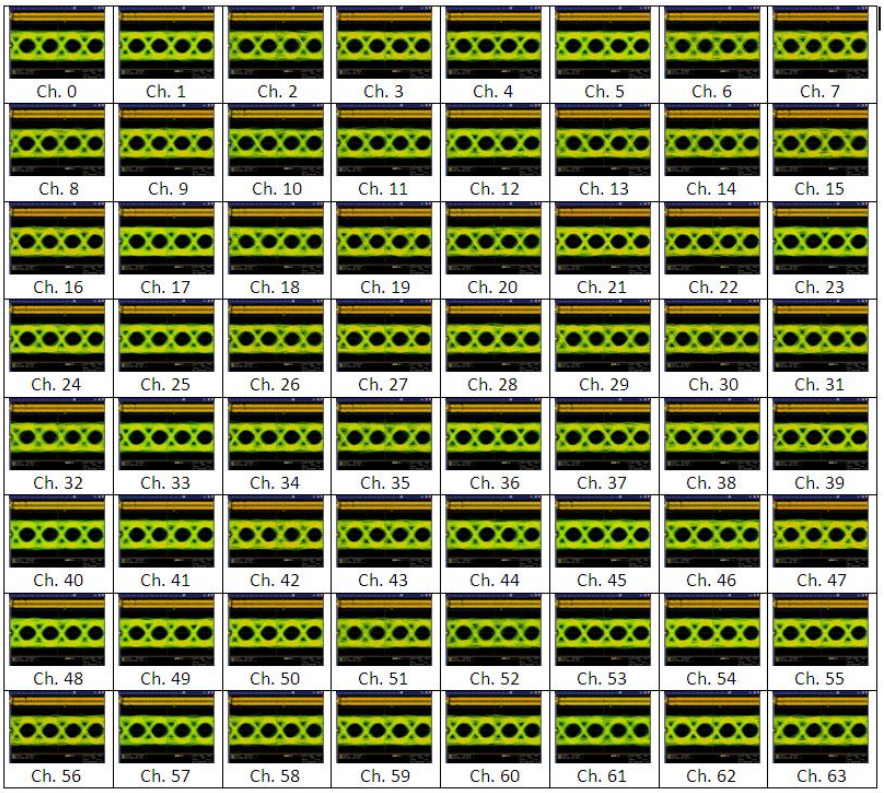

Figure 2-5: Eye patterns from all 64 channels on the TOR

Table 2-2: Estimated BER and margins of loop backed optical signal as measured by ASIC

```
phy diag xe6 heye_l

Port 7 : Start BER extrapolation for Left Eye

BER(extrapolated) = 1e-37
Margin @ 1e-12    is *better* than 60.700000
Margin @ 1e-15    is *better* than 57.500000
Margin @ 1e-18    is *better* than 54.500000

phy diag xe6 heye_r

Port 7 : Start BER extrapolation for Right Eye
BER(extrapolated) is *better* than 1e-34.000000
Margin @ 1e-12    is *better* than 48.900000
Margin @ 1e-15    is *better* than 40.500000
Margin @ 1e-18    is *better* than 32.800000
```

3 Switch OpenFlow integration

The TOR switch built by TU/e in collaboration with PhotonX is using a layer 2/3 Ethernet switch/router provided by Broadcom. Broadcom supports the use of OpenFlow through a dedicated software stack as described in figure 3-1 below:

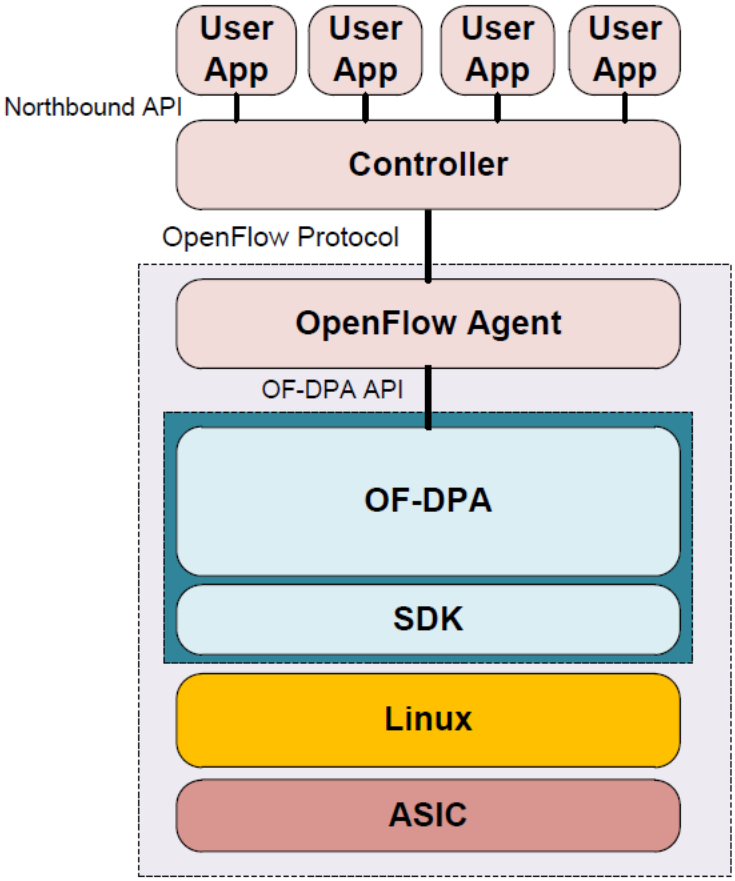

Figure 3-1 – Broadcom OFDPA component layering

The OF-DPA API, as defined in the Open Flow Data Plane Abstraction (OF-DPA) API Guide and Reference Manual, presents a specialized hardware abstraction layer (HAL) that allows programming Broadcom ASICs using OpenFlow abstractions. However, it doesn't yet process OpenFlow protocol messages. An OpenFlow agent is required to create a complete OpenFlow switch using OF-DPA. In addition, an OpenFlow Controller is required to field an OpenFlow network deployment using OF-DPA-enabled switches. Figure 3-1 illustrates the relationship of OF-DPA with the other OpenFlow system components.

In the COSIGN project the agent implemented on top of the OFDPA API is based on open source project Indigo. For more details about indigo visit **https://github.com/floodlight/indigo.**

The OF-DPA agent for the Broadcom ASIC has been compiled with the operating system and has been demonstrated to communicate with the ODL controller.

4 Conclusions

In this document we have detailed the progress followed in the design fabrication and testing of the 1st generation TOR switch using mid-board optics. We have obtained a switch that consumes less than 50% of a reference design switch, is 75% smaller and has a significant reduction in cost. The proposed switch by COSIGN can replace the reference design as it encompasses the same functions. The move to mid-board optical modules allows switches to be much more compact so that more of them can be packaged into available racks within data center network installations.

We have integrated the switch HW with the switch SW demonstrating that the switch supports an OpenFlow 1.3 northbound interface which allows it to connect the OpenDayLight controller platform to provide flexibility and ease the introduction of new functionalities.

The results of the work on the 1st generation prototype have been taken into account in the design process of the 2nd generation TOR that will be due to be ready for testing by the end of M24. This 2nd generation TOR will support double the switching bandwidth (up to 1.28Tbps) with more densely integrated mid-board optical modules. The 2nd generation TOR is designed to be 10% smaller than the 1st generation TOR. It is expected that it will run the same software package as the 1st generation TOR. This will facilitate the effort on SW integration to start with the 1st generation TOR as soon as the work in WP3 and WP4 gets underway, not waiting until the 2nd generation TOR has been tested fully.