UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Cost evaluation of intra-data center networks based on hybrid optical switching technologies

A Degree Thesis

Submitted to the Faculty of the

Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona

Universitat Politècnica de Catalunya

by

Isaac Queralt Garriga

In partial fulfilment

of the requirements for the degree in

Telecommunication systems ENGINEERING

Advisor: Salvatore Spadero

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

telecom BCN

Barcelona, May 2016

Abstract

The project is carried out at the Optical Communications Group (GCO), which is part of the Signal Theory and Communication Department (TSC) in the Polytechnic University of Catalonia.

The purpose of this project is to evaluate the Lightness optical network architecture comparing it against other intra-data center optical structure implementations.

The project can be divided into three major blocks. The first one is to study the current network technologies used in data centers and to see how the implementation of hybrid optical switching technologies can bring improvements to the table.

Once these infrastructures are well studied and understood, the second block of the project consists in designing a virtual simulator in order to compare three of the possible optical/hybrid network architecture implementations.

The way in which we want to compare the results is to analyze the cost to performance ratio given a certain structure for a data center and the specific service of virtualization.

Finally, the third block includes the analysis of the results and the comparison among the studied infrastructures.

Resum

El projecte es du a terme al Grup de Comunicacions Òptiques (GCO), que és part del departament de Teoria del Senyal i les Comunicacions (TSC) de la Universitat Politècnica de Catalunya.

La meta del projecte és l'estudi de l'arquitectura de xarxes òptica Lightness, així com la comparació d'aquesta amb altres estructures de connectivitat òptica dins de centres de dades.

El projecte es pot dividir en tres grans blocs. El primer és l'estudi de tecnologies actuals d'intercomunicació utilitzades en centres de dades i veure com la implementació de arquitectures híbrides que combinen part elèctrica i part òptica pot aportar millores al marc actual.

Una vegada les infraestructures s'hagin definit i entès clarament, el segon bloc consisteix en el disseny d'un simulador virtual amb l'objectiu de comparar la implementació de tres arquitectures de xarxes òptiques/híbrides.

La forma de comparar els resultats és a través de l'anàlisi de la relació cost/rendiment donada una estructura concreta de centre de dades per l'específic servei de virtualització.

El tercer bloc inclou l'estudi dels resultats obtinguts i la pertinent comparativa de les tres arquitectures triades.

Resumen

El proyecto se lleva a cabo en el Grupo de Comunicaciones Ópticas (GCO), que es parte del departamento de Teoría del Señal y Comunicaciones (TSC) de la Universidad Politécnica de Catalunya.

La meta del proyecto es el estudio de la arquitectura de redes ópticas Lightness, así como la comparación de esta con otras estructuras de conectividad óptica dentro de los centros de datos.

El proyecto se puede dividir en tres grandes bloques. El primero es el estudio de tecnologías actuales de intercomunicación utilizadas en centros de datos y ver como la implementación de arquitecturas híbridas utilizadas en centros de datos que incorporan una parte óptica puede aportar mejoras al marco actual.

Una vez las infraestructuras se hayan definido y entendido claramente, el segundo bloque consiste en el diseño de un simulador virtual con el objetivo de comparar la implementación de tres arquitecturas de redes ópticas/híbridas.

La forma de comparar los resultados es a través del análisis de la relación coste/rendimiento dada una estructura concreta de centro de datos específicamente para el servicio de virtualización.

El tercer bloque incluye el estudio de los resultados obtenidos y la pertinente comparativa de las tres estructures escogidas.

Revision history and approval record

Revision	Date	Purpose
0	30/04/2016	Document creation
1		Document revision

DOCUMENT DISTRIBUTION LIST

Name	e-mail
Isaac Queralt Garriga	Isaac_qg@hotmail.com
Salvatore Spadero	spadero@tsc.upc.edu
Albert Pagès Cruz	albertpages@tsc.upc.edu

Written by:		Reviewed and approved by:	
Date	30/04/2016	Date	
Name	Isaac Queralt	Name	Salvatore Spadero
Position	Project Author	Position	Project Supervisor

Table of contents

List of Figures

List of Tables:

1. Introduction

1.1. Statement of purpose

Internet has stablished itself as a powerful platform, impacting the everyday life of many people all over the world. As such, its raise in popularity and demand has been steadily increasing. New services and applications are appearing every day, being online media streaming, gaming or social networks some of the more prominent ones. Forecasts estimating the growth of traffic entering and exiting data centers conclude that, during the period of 2014 to 2019, there will be a increment of up to three times its size (from 3.4 zettabytes annually to 10.4 zettabytes).

Specifically, cloud computing has been adopted as a crucial part in this telecommunication industry. The raise in the aforementioned type of service has to do with the ability of cloud Data Centers to handle significantly higher traffic loads. Cloud Data Centers support increased virtualization, standardization, and automation. These factors lead to better performance as well as higher capacity and throughput.

Traditionally, one server carried one workload. However, with increasing server computing capacity and virtualization, multiple workloads per physical server are common in cloud architectures. Cloud economics, including server cost, resiliency, scalability, and product lifespan, along with enhancements in cloud security, are promoting migration of workloads across servers, both inside the data center and across data centers (even data centers in different geographic areas). Often an end-user application can be supported by several workloads distributed across servers. This approach can generate multiple streams of traffic within and between data centers, in addition to traffic to and from the end user.

Currently, DCs are built around commodity servers organized in racks, connected among them and with servers from other racks through Top-of-the-Rack (ToR) switches. ToRs are the first step in a hierarchical structure that manages the whole traffic in and out of the DC.

The main challenge that arises at this point is the difficulty to properly scale these infrastructures: data centers are required to scale up to hundreds of thousands of nodes. Many of the current solutions use network topologies that will not be able to keep up with the demand increase of such services at reasonable costs. Some fail to achieve cross-section bandwidth at the same time as keeping low oversubscription ratio near the root. Others offer immense scalability but deliver poor performance under heavy network load.

There is, consequently, the necessity to develop new infrastructures that can keep up to this growth.

Three technology trends are emerging as possible solutions to the new data center: leaf-spine architectures (which flatten the tiered architecture of the data center), software-defined networks (SDNs, which separate the control and forwarding of data center traffic), and network function virtualization (NFV, which virtualizes a variety of network elements).

The scope of this project is to analyze one of the proposed solutions to face the demands that future data centers will impose. Specifically, the work will be focused on the implementation of optical systems in data center networks and the comparison among three solutions, two of which are hybrid electrical/optical architectures (c-Through and Helios) and a completely optical solution (Lightness).

1.2. Requirements and specifications.

Project requirements

- Basic optical networking knowledge: required to understand the different studied architectures and to development of the simulator coherently, as well as to understand the results obtained after the data processing.
- Ability to develop a network simulator.
- Self-learning ability, as well as being capable of identifying solutions to engineering problems.
- Basic statistical methods knowledge: To analyze and interpret the results obtained with the simulator that will be designed.

Project specifications:

- Identification of the elements that conform the architectures studied.
- To obtain the implementation and maintenance cost of the optical infrastructures required to design a fully functional optical packet switch network.
- Functioning simulator with which we can obtain data for further analysis.
- Capability/performance analysis for the studied architectures to absorb VDCs (as well as an availability study if possible).

1.3. Methods and procedures

This project is a small appendix of a big European project carried out among many research centers. UPC researchers work in this project, which is what motivated the proposal to study the matter at hand.

As such, it is completely an annex and is not the continuation to any previous study.

1.4. Work plan with tasks

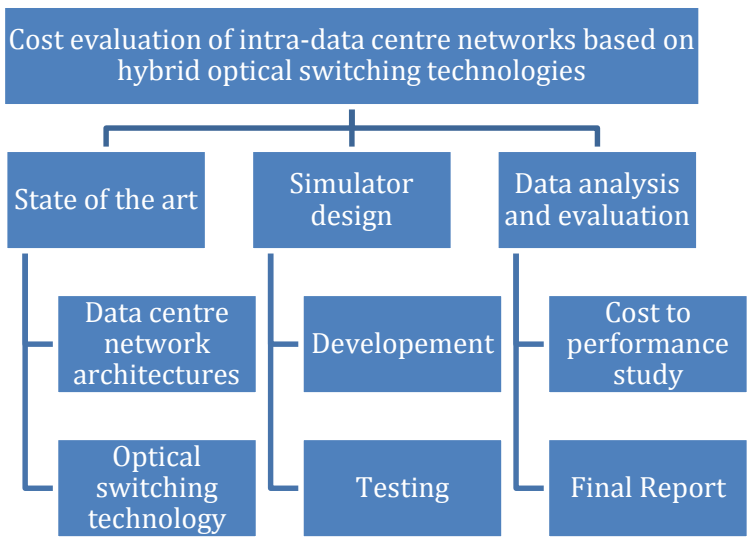


Fig. 1 Workplan diagram

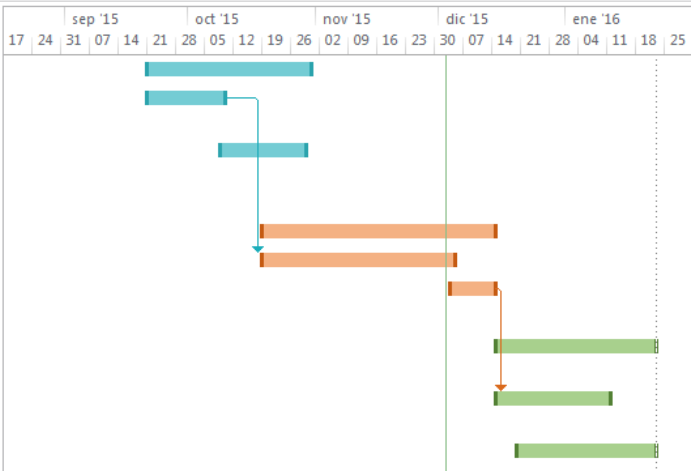
The work plan shows three major blocs in the development of this project. First of all and weighting an important part in the overall scheme of the project, there is the research of information around the topic. This implies, on one side, the understanding of both how each of the hybrid/optical architectures work as well as the optical switching technology used in these systems.

The second part, the one that has taken the longest and represented the biggest workload is the development of the simulation software. Its progress has been much slow than expected, as some parts of the design were probably not quite well defined and some tweaks and implementations have been made as doubts raised during the testing process.

Finally, the cost to performance study and the final report, in which the results for each simulation are analyzed, compared and commented.

1.5. Gantt diagram

Nombre de tarea	Duración	Comienzo	Fin
State of the art	30 días	lun 21/09/15	vie 30/10/15
Data center network architectures	15 días	lun 21/09/15	vie 09/10/15
Optical switching technologies	15 días	vie 09/10/15	jue 29/10/15
Simulator design	41 días	lun 19/10/15	lun 14/12/15
Development	35 días	lun 19/10/15	vie 04/12/15
Testing	7 días	vie 04/12/15	lun 14/12/15
Data analysis and evaluation	29 días	mar 15/12/15	vie 22/01/16
Cost to performance study	20 días	mar 15/12/15	lun 11/01/16
Final report	26 días	dom 20/12/15	vie 22/01/16



The altered Gantt diagram could not be achieved as software development took much longer than expected, delaying for a long time the completion of the project.

The project delivery period ended up being extended to allow for the required modifications to the software that were needed, as some of the previous iterations of it failed to model accurately the proposed schema.

1.6. Description of the deviations from the initial plan

The major deviation from the initial plan has been the time spent in software development and testing which, instead of lasting one month ended up taking over four months. The causes for this to happen were, on one hand, the lack of precision with which the program was defined and, on the other hand, the difficulty it entailed, which was vastly underestimated.

There have been a few iterations of the program, correcting problems as they were appearing, redirecting aspects that were dubious or that simply presented configuration issues and where not coherent with what was being expected.

Additionally, the lack of communication from the student to the project supervisors ended up damaging the development of the project.

2. State of the art of the technology used or applied in this thesis:

2.1. Optical solutions: where they stand

Before jumping into the analysis of these the three solutions some context of current optical technology and its possibility to be implemented.

Optical circuit switching technology is rapidly developing and achieving better results. Nowadays, the fastest electrical switches are limited to about 40Gbit/s per port, whereas optical links are already achieving 100Gbit/s, with the perspective of increasing that figure reaching higher orders of magnitude. As for optical switches, current high end products already offer 320x320 optical circuit switches with 40Gbit/s transceivers, numbers that are going to scale upward as it becomes more mainstream for datacenter usage.

One of the challenges that this technology faces is the re-mapping of ports (input to output). Microelectromechanical systems (MEMS) are one of the proprietary optical switch structures that allows for simultaneous connections of multiple input to output fibers in a non-blocking, all-optical cross-connect configurations. The way MEMS work is through the use of small, tilting mirrors, configured in a double 3D matrix figure.

Some of the principal advantages of this technology are the decrease of power consumption compared to electrical switches as there is no optical to electrical or electrical to optical packet conversion and packets are not processed. This can mean a reduction of over 12W per port in the switch.

The principal downside, as it has been stated, is that reconfiguration take a few milliseconds each time. Which can slow down the process in certain situations.

Another alternative is the use of tunable lasers combined with an Arrayed Waveguide Grating Router. Tunable lasers can switch channels in tens of nanoseconds, however, during the switching step, the laser can sweep through other channels, sending false information and inducing data transmission errors, thus the need to incorporate isolation from the network during the process. This step ranges from 1 to 10ms in speed.

One of the negative aspects of optics is the fact that it is more expensive to build an optical network compared to a full electrical one. However, this issue has been narrowing over the last few years: optical transceiver's price has dropped more than 90 percent in the last ten years. It is expected that, as this technology reaches volume manufacturing as a result of high demand, due to the high bandwidth requirements, its price will go down even further, making it a viable option.

2.2. Presentation of the three infrastructures that will be compared

To begin with, a short description of the three will be presented and, after that, an on paper comparison will be made. Later on, through the results provided by the tool that has been developed, a more formal comparison will be made.

However, it should be noted that most of the new architectures that are appearing are experimental and have not been tested in real world environment, as some of the components required to build these cutting edge networks still do not exist.

2.2.1. c-Through

The architecture of c-Through is presented as an enhancement to current data center networks. It comprises both a circuit (optical) and a packet-based (electrical) networks conforming, as it has been said earlier, a hybrid solution.

Both infrastructures connect the ToRs to different networks, on one hand the electrical network is conformed in a traditional tree topology with Electrical switches. On the other hand, the optical network connects the ToRs through a MEMS switch which can only provide a matching on the graph of racks. This implies that connections among racks will be such that pairs of racks with high bandwidth demands.

Therefore a traffic monitoring system is required. An optical configuration manager collects traffic measurements and determines how the paths should be configured. In order to organize connections accordingly with the traffic demand this problem is formulated as a "*maximum weight perfect matching problem*". The solution is found through Edmons' algorithm, which takes a few hundreds of milliseconds to calculate a full DC with over 1000 racks.

For this network to provide benefits the traffic has to be "pairwise concentrated", that is, there have to be racks that require this one to one high bandwidth communication. Studies have shown that this condition is given regularly in cloud applications.

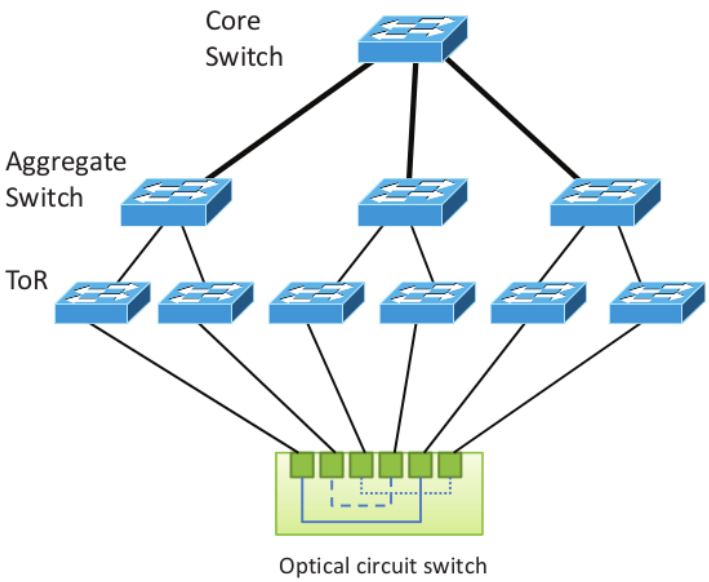


Fig. 2 c-Through diagram

2.2.2. Helios

Helios' structure is similar to the one shown in c-Through. It also uses MEMS-based Optical Circuit Switching and has a part electrical part optical network. The two main differences are the following ones:

Firstly, instead of a simple electrical tree topology, Helios uses a 2-level multi-rooted tree or leaf-spine topology for the packet switching side of the network. The core switches in this network can be either electrical packet switches or optical circuit switches, compensating the weaknesses of each kind with the strengths of the other.

Secondly, instead of using a MEMS switch that limits the connection to exclusive pairing it allows for multiple interconnections among racks. This is all because the use of optical circuit switching (OCS).

Additionally, the use of wavelength division multiplexing (WDM) many channels of information can be packed into a single optical fiber, reducing the amount of fiber links that must be deployed and thus, reducing the cost of the overall infrastructure by adding multiplexers and CWDM transceivers.

Through the use of WDM, the "superlinks", conformed by the aggregation of said optical channels can provide up to *w* x 10Gbit/s (where *w* ranges from 1 to 32), whereas electrical packet switches are limited to 10Gbit/s.

In the same way as c-Through, some kind of traffic control must be provided. In this case a Topology Manager is used to dynamically monitor shifting communication patterns, implemented through TCP fixpoint algorithm and Edmonds' algorithm.

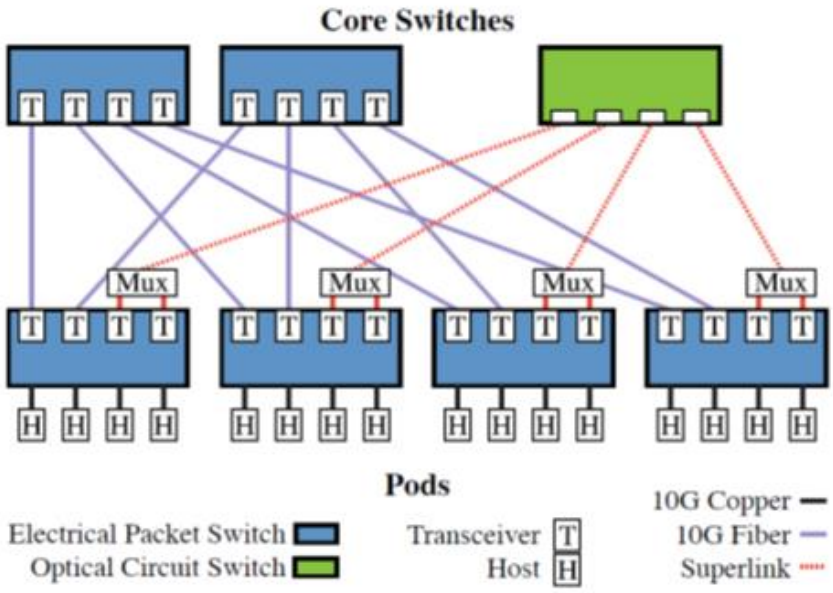


Fig. 3 Helios diagram

The figure shows a simplified example of a Helios infrastructure. Each pod has a number of hosts (labelled "H") connected to the pod switch by short copper links. The pod switch contains a number of optical transceivers (labelled "T") to connect to the core switching array. In this example, half of the uplinks from each pod are connected to packet switches, each of which also requires an optical transceiver. The other half of uplinks from each pod

switch pass through a passive optical multiplexer (labelled “Mux”) before connecting to a single optical circuit switch.

2.2.3. Lightness

Lightness is conformed as a flat architecture that combines both OPS and OCS to deliver short-lived as well as long-lived traffic. On one hand, the design of OPS switch targets high port count and low latency, which is employed for switching short-lived packet flows. On the other hand, the OCS switches aims at handling long-lived data flows. The computing servers are connected to the hybrid OCS/OPS ToRs, which perform an application-aware classification to either short- or long-lived traffic. These are also connected to the Intra-inter DC interface.

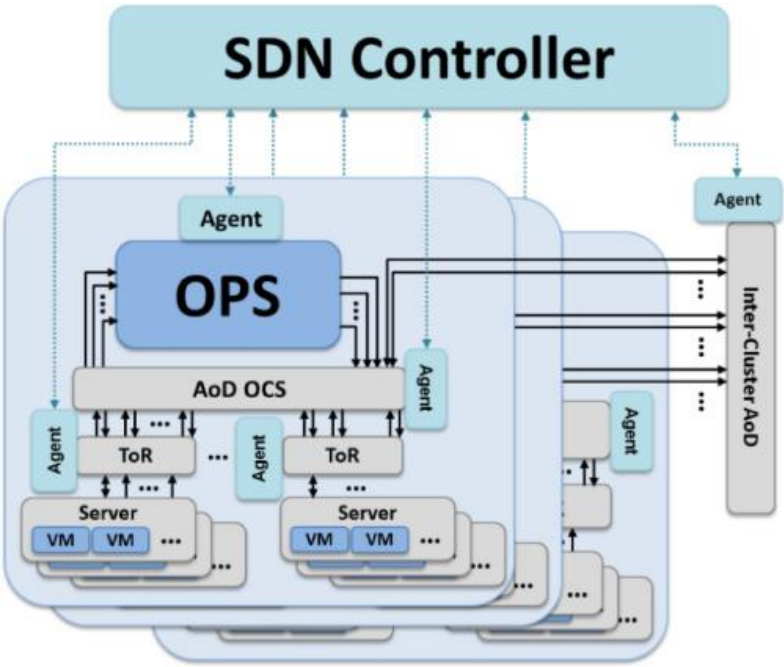


Fig. 4 Lightness diagram

This architecture uses multiple OCS/OPS switches to allow for scalability and interconnect all the racks. This allows the number of wavelengths to/from each ToR to be increased when the port count of OPS or OCS switches become the limiting factor.

Unlike Helios or c-Through, Lightness is a more versatile solution, which can deal better with different types of applications, as the combination of OPS and OCS technologies are linked through an Architecture-on-Demand (AoD) node.

The AoD node consists of an optical backplane of MEMS switches with a large port count (up to 512x512), with signal processing and/or switching modules such as AWG-based MUX/DEMUX, OPS and ToRs connected to it. Overall, this infrastructure allows for traffic switching in space, frequency and time, which translates as a way of reorganizing the network, as its own name says, on demand.

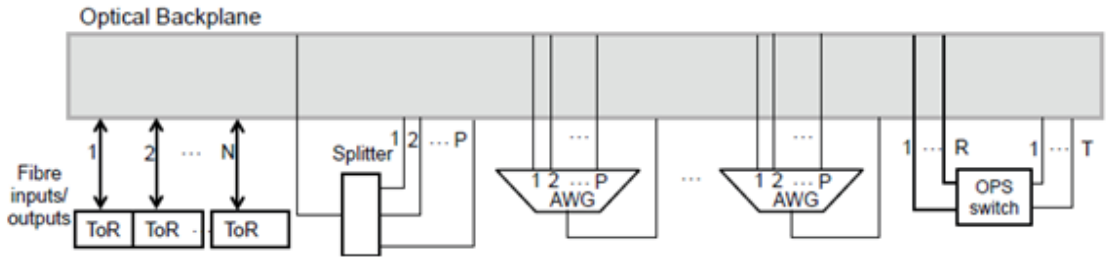


Fig. 5 AoD diagram

As for Top-of-the-Rack switches, the design expects the use of high-speed field programmable gate array (FPGA).

2.3. Simulation tool: Design and development

In order to perform this study a simulation tool has been designed to aid in the comparative aspect of the project. Before describing its functioning some insights should be shared in order to properly understand the decisions and compromises that have been made during this process.

CISCO describes a server workload as a virtual or physical set of computer resources, including storage, that are assigned to run a specific application or provide computing services for one to many users. A workload is a general measurement used to describe many different applications, from a small application to a large computational private cloud database application. It is estimated that, by 2019, more than four-fifths (86 percent) of all workloads will be processed in cloud data centers.

This tool's objective is to provide the costs of implementation of a datacenter network with one of the infrastructures described previously through its virtual implementation. Additionally, the software aims at comparing the workload capacity of the different infrastructures and stablish cost/performance ratios tied to the networking capabilities.

In order to simplify but, at the same time, provide of helpful understanding, it has been decided to focus on virtual server hosting, specifically, cloud allocation of virtual machines in a network configuration. Each set of virtual machines connected to each other is called Virtual Data Center or VDC.

Once the costs of the infrastructure are calculated, the Data Center is filled with VDCs according to the availability of the resources in the Data Center. The generation of these VDCs is random within certain fixed parameters, providing realistic virtual machines and network requirements among them. After no more VDCs can be allocated the software calculates, once again, the costs of the network devices in usage. This allows us to obtain insightful data to compare how the infrastructures behave under realistic loads.

3. Methodology / project development:

In this section there will be a description of the steps followed in the design of the software, as well as an explanation of how it works.

3.1. Software design

3.1.1. Simplifications and compromises

In order to implement a realistic approach to the project at hand some compromises have to be made:

Simplification of infrastructures

It is not realistic to represent a complete Data Center, it is too complex and far from the scope of this project. Our approach simplifies the infrastructure aspect, the model that has been implemented takes into account only the necessary components to evaluate the difference in cost in the structures, as well as enabling the study for VDC occupation in the DC.

Static Data Center design

Since the implementation focuses on allocable Virtual Data Centers to cost analysis, the design is static, that is, once a VM is placed in the DC it does not time out or die. The objective is to compute the maximum amount of allocable VMs for a given infrastructure.

Wavelength continuity

Optical traffic that goes through the MEMS or OCS cannot change its carrier. This is a very important point and has to be considered, otherwise the results could come out indubitably biased.

OPS traffic aggregation

In Lightness' case packet traffic has to be treated differently from the other two. Since aggregation is possible there must be a certain quality of service (QoS). Otherwise packet collision could be considerable and deny any realistic attempt at transmitting information. With that in mind it has been decided to allow aggregation to up to a 60% of the total link capacity for packet traffic, or up to 100% for not aggregated packet traffic.

Virtual Machine distribution inside the Data Center

It has been considered appropriate to split the VMs of every VDC into different servers (and racks) if possible. That is done in order to provide certain redundancy to the system in case a server (or even a rack) fails.

Virtual Machine allocation

The algorithm used in the allocation of virtual machines in the Data Center is a simple "first fit" design with the considerations previously stated. The execution of the simulation ends when it has not been possible to allocate a certain number of VDCs proposed by the application.

3.1.2. Infrastructure and technical aspects

Before describing how have the three architectures been modeled there will be a small explanation of common points among them, such as virtual machine types and requirements and server characteristics, as well as how have the network elements been implemented. These are all common factors in the three infrastructures.

As mentioned previously, in order to model more realistically the VDCs that are allocated inside the DC, it has been decided to add some randomness to the items that conform them. Below there is a table detailing the different possible VM configurations. These are realistic values obtained from services already available such as Amazon's EC2.

Virtual Machine resource requirements			
	CPU (Logical cores)	**RAM (GB)**	**Storage (TB)**
VM1	2	4	400
VM2	4	8	850
VM3	6	17	420
VM4	8	15	1500
VM5	12	34	850
VM6	20	7	1700
VM7	30	60	2000
VM8	30	117	4800
VM9	32	60	2000

Table 1 Virtual Machine resource requirements

In regards to the servers, each one has the following resources, also based in up-to-date obtainable configurations, using what could be the equivalent to running dual high end Intel Xeon processors, high capacity drives in redundant configuration and typical ECC server memory.

Server Resources		
CPU (Logical cores)	**RAM (GB)**	**Storage (TB)**
48	144	12000

After some preliminary tests it has been clear that CPU will be the limiting factor in most cases.

In regards to the network equipment it has been considered that switches do not have restrictions besides the amount of ports. On the other hand, the links that interconnect every device do.

Electrical links are limited at 10 Gbit/s, whereas optical links are considered to have available up to 10 Gbit/s for each wavelength in use. The amount of wavelengths per optical connector is variable but the more realistic values range around 32 per fiber.

3.1.2.1. Architecture implementation description

c-Through

The c-Through implementation has been faced with the following approach:

An optical switch is added to the simple three-layer electric tree topology. Every ToR is connected to it independent of the cluster. If possible, when a connection is over a certain threshold (about 8 to 10 Gbit/s) it is studied if it can fit through the optical network, otherwise it is routed through the packet network.

Since the design is static c-Through is penalized in one aspect, because the optical switch does not allow for multiple rack to rack communication, that is, three racks connected between them optically, most of the traffic ends up through the packet route.

On the other hand, as electric switches do not have throughput limitations in our layout, it will be benefited in this aspect.

Helios

Helios is a bit more complex. There is no cluster differentiation and all ToRs are connected to both the electric and optic networks. More than one electric switch is possible to add redundancy and diminish the theoretical limitations of throughput, reducing the oversubscription intrinsic to this design. However, only one optical switch is available but it is more flexible than the c-Through one, as it can connect multiple racks together through OCS.

The optical switch will have one port for each rack. For very big designs this might be irrational, as MEMS over 500 ports are not yet available but since the cost computation is based on price per port, splitting the racks between two MEMS would not have a big impact on performance.

In the same way as in the previous architecture, every connection with capacity requirements between 8 and 10 Gbit/s will be routed through the optical network if possible. Otherwise it will be over the electrical network instead.

Lightness

Out of the three, this last architecture is the more complex. The design is much different from the other two, as it implements OCS-AoD and MEMS for each cluster, as well as an additional OCS-AoD to interconnect all the clusters.

In this case, each ToR is connected to the OCS-AoD of the own cluster, which is at the same time connected to the MEMS through multiple fibers (one or more for each rack) and to the OCS-AoD that connects the one in each cluster together.

Connections that require connectivity over 8 Gbit/s will be OCS, whereas the ones that require less bandwidth will be OPS, routed through MEMS as described in *Simplifications and compromises*.

3.1.3. Simulation workflow

The process that the designed software follows to achieve the objective will be explained now in more detail.

When the application is run a Windows Form allows the user to select which of the three infrastructures wants to simulate. Depending on which infrastructure is selected some textboxes will appear to gather the required information to design the structure, the amount of servers per rack or how many racks each cluster can hold are some of these parameters. To provide more consistency to the study an additional parameter has been implemented: times the simulation will run. Since some of the parameters are generated randomly there must be a way to average the obtained results and make them more coherent. By simulating many times each scenario a mean can be drawn and the results will represent better the behavior of the scheme.

After inputting the specifications and clicking the “Start simulation” the simulation will begin.

The first step is to generate all the items that conform the Data Center. According to the parameters inputted the servers, racks, clusters, switches, etcetera… will be generated. The following step is to connect all the components as the selected infrastructure requires so.

After the Data Center is created the loop of VDC generation and allocation begins. The VDCs are generated in the following manner: first, three virtual machines are created using the template of one of the configurations previously mentioned. Next they are all interconnected and a network requirement is assigned to each link. This requirement is also decided randomly, between 1 to 10 Gbit/s in steps of 1Gbit/s.

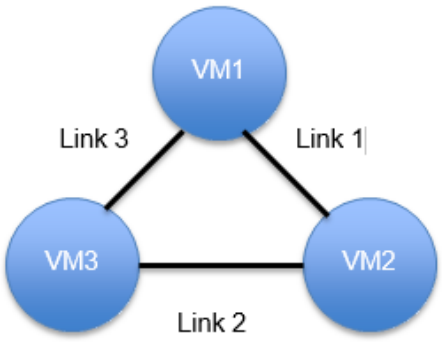


Fig. 6 VDC design

Once the VDC is generated the algorithm searches for the first three servers that have the capacity to assimilate the VM. It is imperative for the servers to be in different racks. Then it is verified that there are available network resources to connect the three servers, if not, each of those that do not accomplish the requirements are replaced others. Once all the conditions are met the VDC is allocated, the resources deducted and the process starts again.

If, after running through the whole architecture, it has been impossible to allocate the VDC, it will be discarded. Since some times that will not imply that the DC is full and no more VDCs can be allocated this operation will be repeated a few times to allow for other VDCs to be allocated.

Once this process concludes, that is, a certain number of VDCs have been rejected, it will start again without the restriction of putting the VMs in different racks. This is done as an addition to check how the network limits possible allocations instead of the resources per server. The more VDCs that fit this way the more the network acts as a limit in our case study.

When the infrastructure is completely full the program calculates both, the total cost of the infrastructure and the cost of the used infrastructure, this provides meaningful insights as it is shown how much of redundancy the network has and displays possible tuning.

As the simulation runs a text label informs the user every time a new VDC has been successfully been inserted up until the simulation finishes, then the results will be outputted in a text document. If it is configured to run multiple simulations it outputs both the information of each one and the average.

3.1.4. Software implementation

The software implementation has been carried out with the object oriented language C#, through Visual Studio Community 2015. Besides the fact that OOP was perfect for programming this application, C# has a lot of free documentation available and the integrated development environment is completely free.

A windows form application has been built to ease the input of the datacenter parameters.

3.1.4.1. Software results

The software obtained is the one shown in the next picture. Upon clicking the pull-down menu the three architecture options appear.

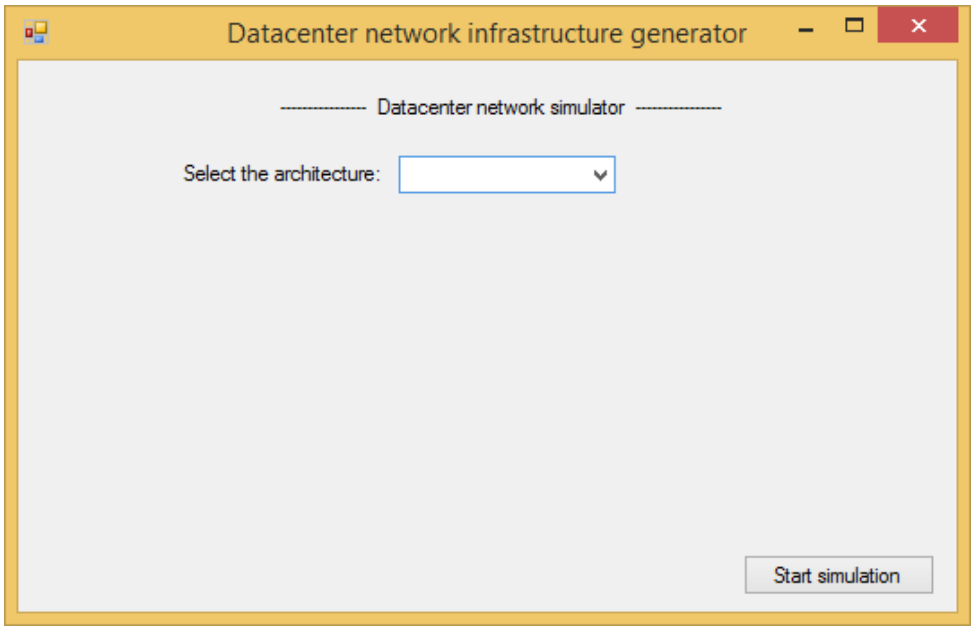


Fig. 7 Software initial window

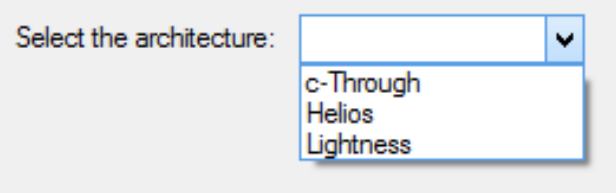


Fig. 8 Architecture selector

Once a selection has been made, the panel is filled with the options to input all the corresponding parameters.

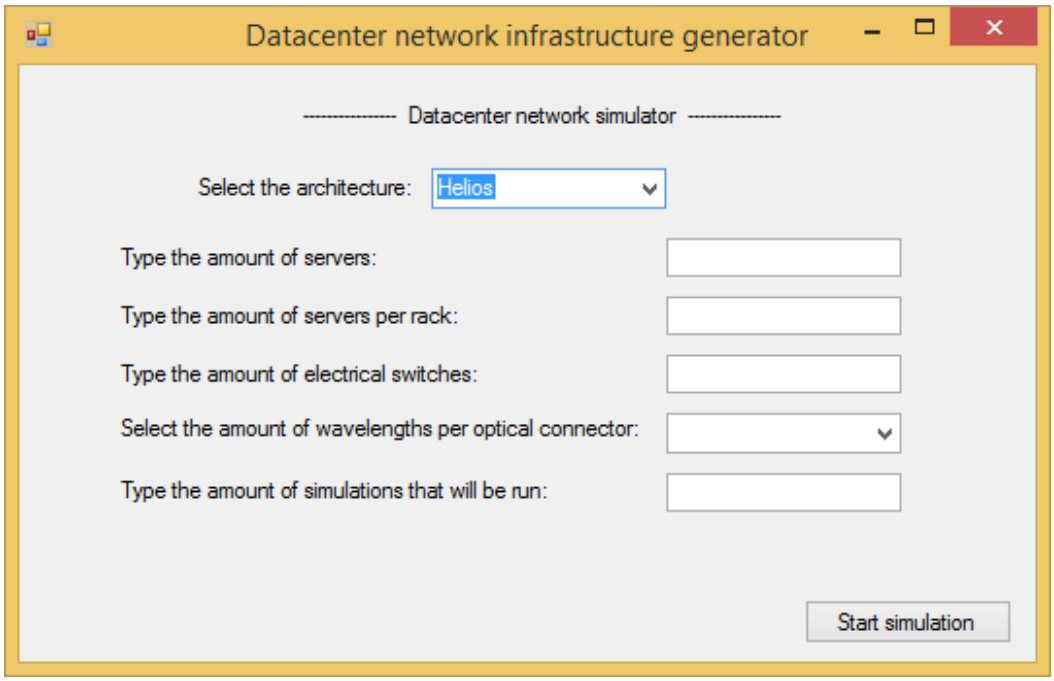


Fig. 9 Parameter input step

After inputting all the required parameters the simulation begins pressing the “Start simulation” button.

Once it is complete a window will inform the user that all calculations have been done and will close the application.

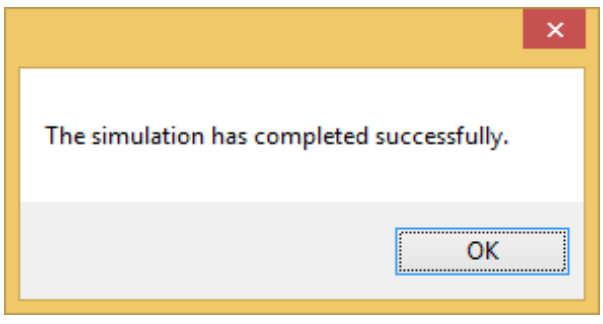


Fig. 10 Completion window

4. Budget

4.1. Components cost

The calculation cost of each infrastructure has been carried out in the following manner:

Once each DC simulation has been complete, all of the components used have been multiplied by its value.

Before describing the specific numbers of each architecture some background will be provided.

The cost of the electrical switches depends on the amount of ports the component has. It is calculated based on the cost of the application specific integrated circuits (ASIC) for one side, the mainboard, which is about 1000€ and the additional cost of components such as fans, casing, etc… which add another 400€.

The ASIC cost, as it has been said, depends on the ports of the switch. For switches up to 128 ports the costs are the ones shown in the Table 2.ASIC costs More than one mainboard of 128 ports switch has to be used to conform larger switches, as there exist space and power dissipation limitations. With this, the cost of switches over 128 ports scales rapidly. In order to implement a 256 switch port, six 128 port mainboards have to be used, that is, six times the cost of one 128 port switch.

If the implementation requires switches with 512 ports, its cost will be of six times the one of a 256 ports, which is thirty six times the cost of a 128 port switch.

Next there is the table for the ASIC pricing:

Amount of ports	Cost (€)
26	410
48	870
64	1140
128	1587

Table 2.ASIC costs

In regards to optical switches, the 32x32 OPS' estimated cost is 80.000€ based on InP based generic integration technology. As for OCS, according to the commercially available products distributed by Polaris, the cost of an OCS is around 340€ per port.

Other components' cost is described in the following table:

Components	Cost (€)
SFP	120
QSFP	339
CXP	650
Electrical cable	13
Optical Fiber	82

Table 3 Connection component costs

4.2. Infrastructure cost calculation

In this section it is explained how the calculation cost of each infrastructure has been done.

4.2.1. c-Through

- Connections

To begin with, all the servers are connected to the ToR through 10Gbit Ethernet. From there the connections to the aggregated switches are done through optical fiber, with the corresponding transponders in each case. Later, all aggregated switches are connected to the core switch, again, with optical fiber and the corresponding transponders in each case.

- Switches

Electrical switch costs are calculated as described earlier. In this way, 128 port switch cost 2.987 € each, 256 port switch cost 17.922 € and, finally, 512 port switches cost 107.532 €. Therefore, ToRs and aggregated switches cost will vary depending on the amount of servers per rack and racks per cluster.

The optical switch in c-Through (one per cluster) is considered an OCS and, as such, its cost depends only in the amount of ports it has, which is twice the number of racks per cluster.

4.2.2. Helios

- Connections

The Helios architecture is structured similar to c-Through. To begin with, all the servers are connected to the ToR with 10Gbit Ethernet. From the ToR to the electrical switches the connection is done with optical fiber and the corresponding transponders. However, only those that are actually used will be counted, therefore the redundancy is not penalized in this study and it has to be taken into consideration when comparing the architectures.

The connection from the ToRs to the MEMS is done with optical fiber as well but the transponders are CXP to allow for the required bandwidth.

- Switches

Electrical switch costs are the same as in c-Through. However, since every electrical switch (known as spine in this architecture) is connected to all the ToRs it is considered to already be 512 ports big.

The optical switch in this infrastructure, as it has been stated previously, is a MEMS and, as such its cost is 80.000 €.

4.2.3. Lightness

- Connections

Lightness' architecture differs greatly from the other two. Since it uses only optical switches the amount of transponders is reduced and that has a big decrease in costs.

Only optical fiber is used, from the server to the ToR, from the ToR to the AOD and to the MEMS as well.

- Switches

On one hand the ToR switches are represented in the same way as the ones in the other two architectures and have a cost variable according to the amount of servers per rack.

On the other hand OCS switches, that accounts for both AoDs, the inter-rack and the inter-cluster, have a cost per port already stated. MEMS switches have a fixed cost of 80.000 € each, and there is per cluster.

5. Results

This section gathers the results obtained through simulating different scenarios with the developed software. The objective is to try to arrive to coherent results that allow us to compare the three architectures and comment the cost of each one contrasted to its performance. Each scenario is repeated 100 times for each architecture to ensure more rigorous results and minimize the variance intrinsic to randomness in our simulations.

The main parameters analyzed are the amount of VDC allocated on the DC, which are classified between inter-rack and intra-rack, the optical connections achieved, and the total cost of the infrastructure used.

The scenario that is going to be studied is a simple Data Center based on 288 servers, distributed in racks of 48 each, in a cluster configuration.

After that, another more complex scenario will be used to understand if there is scalability and if the first simulation can be used to predict bigger implementations.

The second simulation is, therefore, a step further: it implements 3 of the aforementioned DCs. It is based on 864 servers distributed in racks of 48 servers each, in 3 clusters of 6 racks (for c-Through and Lightness).

In order for the simulation to be representative, the Helios architecture for the second scenario uses 3 electrical switches.

Those results that greatly differ from the average sample will be omitted.

1st Scenario – Results obtained after 100 simulations			
	c-Through	**Helios**	**Lightness**
Allocated VDC	198,3	203,6	233,2
VDC allocated intra-rack	44,8	11,6	31,3
Optical connections	6,9	85,5	82,7
Total cost of the infrastructure used (€)	276.279	508.331,9	542.556

Table 4 – 1st Scenario Results

It can be seen from the simulation results that the infrastructure that has performed better in (allocated VDC) / (cost of the infrastructure) is c-Through, followed by Lightness and Helios being the last one.

However, c-Through is not a clear winner, as it is the network that has more allocated VDCs inside the same rack, and the one that has the lowest amount of optical connections.

Lightness comes on top in the amount of VDCs allocated, which has to be contemplated as, even if the common infrastructure cost is not being taken into account, it also accounts for a big part of the investment in a new DC. In this case the full optical infrastructure allocates almost a 14% and 18% more VDCs than Helios and c-Through respectively.

2nd Scenario – Results obtained after 100 simulations			
	c-Through	**Helios**	**Lightness**
Allocated VDC	641,4	568	727,3
VDC allocated intra-rack	105,9	4,8	18,4
Optical connections	24,4	279,4	252,1
Total cost of the infrastructure used (€)	868.946,2	1.729.286	1.628.522

Table 5 – 2nd Scenario Results

From the second scenario we can see that results change quite a bit. While the performance of both c-Through and Lightness scale quite naturally, Helios' does not. It is in part because Helios only has one MEMS switch, which ends up limiting the optical connections and hurting the overall performance.

Lightness is still the architecture that allocates more VDC and c-Through still shows that cannot compete with the other two in regards to high bandwidth communications.

6. Conclusions and future development

The conclusion and future development part is divided into two different sub-sections. In the first one the software is studied, criticizing the weak aspects and possible improvements or modifications for the software to better model the virtual implementation of these architectures.

In the second sub-section a post-results analysis of the implementation has been done after obtaining the results, again, discussing the lacking aspects of it.

6.1. Software

After analyzing the results obtained with the simulator some aspects come to mind as to what could be improved in order to increase reliability and obtain more significant results with the software. To compare and dimension more accurately the different Data Center network architectures the following points should be applied:

- Components could be implemented considering certain limitations, for instance, the throughput in Aggregation Switches, not contemplated in this project.
- The possibility to propose more VDC structures, such as star or ring virtual network implementations, with variable number of VMs and connectivity.
- Rely on a dynamic model instead of a static design, where VDCs had a random serve time and new ones were continuously added to the DC.

6.2. Hybrid/Optical networks

Once the results have been obtained it has been clear that a static implementation has severe disadvantages when comparing this kind of infrastructures. Some aspects, like the latency or the oversubscription existing in these architectures cannot be properly weighed in the cost to performance analysis.

Another important aspect that lacks visibility after the results obtained is

Below now there will be a short study at each network particularly:

c-Through

Out of the three architectures studied this is the one that benefits the most from the model that has been chosen because of the following aspects:

- Even if it barely implements optical connections the only way that performance is examined is in the form of allocated VDC and, if we relate it to the cost, the simpler architecture for the same server infrastructure is the one bounded to produce better results.
- As it has already been said, this study does not consider some of the aspects in which the c-Through infrastructure falls behind with the other architectures. Redundancy, latency and oversubscription (or bisection bandwidth) are the weak points of this design towards the others.

Helios

Helios is a more complicated network structure than c-Through, it implements higher redundancy and the optical part is more complex, as it adds WDM and higher end MEMS.

This model used in the simulations has some flaws regarding the Helios infrastructure as well:

- Current MEMS do not yet reach high port count. This model simulates a 32 port MEMS and thus, configurations over 32 racks would not be realistically modeled.
- Since electrical switches do not have throughput limitations, the scenario of a single switch is possible but probably not realistic in configurations with high rack count. Therefore, it has to be taken into account when simulating, as it has been done in the second scenario described at the Results section.

Lightness

The overall design of Lightness is very complex and it is hard to judge with a static analysis the performance of said infrastructure. However, during the preliminary simulations that have been carried out while tweaking the software the behavior of the network has been the expected.

Probably, in the same line of thinking, the potential of this network could be further studied in a dynamic software simulation, as one of the strong points of this network is the flexibility to configure itself to satisfy the required demands, which is clearly not shown here.

Bibliography:

[1] Cisco Global Cloud Index 2012. [Online] Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/globalcloud-index-gci/Cloud_Index_White_Paper.html.

[2] Limei Peng, Chan-Hyun Youn, *Member, IEEE*, Wan Tang, and Chunming Qiao, *Fellow, IEEE "A Novel Approach to Optical Switching for Intradatacenter Networking"*

[3] Guohui Wang, David G. Andersen, Michael Kaminsky, Konstantina Papagiannaki, T. S. Eugene Ng, Michael Kozuch, Michael Ryan. "*c-Through: Part-time Optics in Data Centers*" Rice University, Carnegie Mellon University, Intel Labs Pittsburgh.

[4] Nathan Farrington, George Porter, Sivasankar Radhakrishnan, Hamid Hajabdolali Bazzaz, Vikram Subramanya, Yeshaiahu Fainman, George Papen, and Amin Vahdat "*Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers*" University of California, San Diego.

[5] Fulong Yan, Wang Miao, Harm Dorren, Nicola Calabretta. "*On the cost, latency, and bandwidth of LIGHTNESS data center network architecture*" Beihang University, Beijing, China. COBRA Research Institute, Eindhoven University of Technology, Eindhoven, the Netherlands.

[6] "*A Novel SDN enabled Hybrid Optical Packet/Circuit Switched Data Centre Network: the LIGHTNESS approach*"

Shuping Peng, Dimitra Simeonidou, George Zervas, Reza Nejabati, Yan Yan, Yi Shu -High Performance Networks Group University of Bristol, United Kingdom.

Salvatore Spadaro, Jordi Perelló, Fernando Agraz, Davide Careglio - Universitat Politècnica de Catalunya Barcelona, Spain.

Harm Dorren, Wang Miao, Nicola Calabretta - COBRA Research Institute, Eindhoven University of Technology, Eindhoven, the Netherlands.

Giacomo Bernini, Nicola Ciulli – Nextworks via Livornese 1027 Pisa, 56122, Italy

Jose Carlos Sancho[1], Steluta Iordache[1], Yolanda Becerra[2], Montse Farreras[2] - [1]Barcelona Supercomputing Center -[2]Universitat Politècnica de Catalunya, Barcelona, Spain

Matteo Biancani, Alessandro Predieri - Interoute S.p.A., via Cornelia 498 Roma, 00166, Italy

Roberto Proietti, Zheng Cao, Lei Liu, S. J. B. Yoo - Department of Electrical and Computer Engineering University of California, Davis, California 95616, U.S.A

[7] Albert Pagès, Sergio Jiménez, Jordi Perelló, and Salvatore Spadaro, Member, IEEE. "*Performance Evaluation of an All-Optical OCS/OPS-Based Network for Intra-Data Center Connectivity Services*" Advanced Broadband Communications Centre (CCABA), Universitat Politècnica de Catalunya (UPC).

[8] Mohammad Naimur Rahman, Dr. Amir Esmailpour. "*A Hybrid Electrical and Optical Networking Topology of Data Center for Big Data Network*" ASEE 2014 Zone I Conference, April 3-5, 2014, University of Bridgeport, Bridgeport, CT, USA.

[9] Christoforos Kachris, Ioannis Tomkos. "*A Survey on Optical Interconnects for Data Centers*", IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 14, NO. 4, FOURTH QUARTER 2012.

[10] Norberto Amaya, Georgios S. Zervas, Dimitra Simeonidou. *"Architecture on Demand for Transparent Optical Networks".* School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK.

[11] MSDN: Learn to Develop with Microsoft Developer Network, [Online] https://msdn.microsoft.com

Glossary

- GCO: Optical Communications Group
- TSC: Signal Theory and Communication Department
- DC: Data Center
- VM: Virtual Machine
- VDC: Virtual Data Center
- OPS: Optical Packet Switching
- OCS: Optical Circuit Switching
- MEMS: Microelectromechanical Systems
- ToR switch: Top of the Rack switch
- SDN: software-defined networks
- AoD: Architecture on Demand
- NFV: Network Function Virtualization
- WDM: Wavelength Division Multiplexing
- AWG: Arrayed Waveguide Grating
- FPGA: Field Programmable Gate Array
- ECC: Error-Correcting Code
- CPU: Central Processing Unit
- OOP: Object-Oriented Programming
- ASIC: Application Specific Integrated Circuit