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 D5.0 – Definition, Design and Test Plan for Use Cases

Due date of deliverable: 31/12-2015
Actual submission date: 21/12-2015

Start date of project: January 1, 2014

Duration: 36 months

Lead contractor for this deliverable: DTU

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

The demonstrator work in COSIGN is organised in three different timeframes reflecting the expected time to implementation in datacentres. These timeframes are 'short-', 'medium-' and 'long-term'.

The short-term demonstrator will confirm the performance of the TU/e Ethernet switch and act as reference for the other demonstrators by implementing a legacy-type Ethernet network based on the compact high radix TU/e Ethernet switch developed in COSIGN.

The medium-term demonstrator will introduce optical switching in the DCN as an optical overlay to a more conventional Ethernet-based network provided by a large port-count switch from Polatis. This demonstrator will be integrated with the COSIGN control plane and orchestration mechanisms to allow demonstrations of several use-cases: the Virtual Application use-case, the Virtual Datacentre provisioning use-case and the Virtual Datacentre VM migration use-case. The final industrial use-case demonstration of the medium-term scenario will be hosted by Interoute in their datacentre facility.

The long-term demonstrator will focus on the implementation of an all-optical DCN using multiple advanced optical technologies. It will be carried out pursuing two different approaches to demonstrate different achievable benefits by the large-scale introduction of optical technologies in DCNs.

One approach will pursue a fundamentally different network topology allowed by advanced optical fibers and switches: a ring-based topology combined with multilayer optical connectivity will be demonstrated. This approach allows all-optical connectivity while significantly reducing network complexity.

Another approach will pursue a DCN structure based on all-optically interconnected clusters. Here optical connections of various capacity can be established between servers by going though different domains of inter-/intra- cluster switches. This modular approach is highly compatible with the trend of disaggregating resources in data centres.

A number of test-cases have been defined related to the use cases addressed by COSIGN. These test-cases will be implemented to verify the compatibility and performance of the demonstrator scenarios with respect to the different use-cases.

For the Virtual Application use-case visual Ping or video streaming will be implemented as test cases where performance can be evaluated.

For the Virtual Data Centre use-case a mechanism for provisioning of VDCs on a reconfigurable optical DCN will be implemented allowing for performance evaluation.

A Virtual Data Centre VM migration test-case will be implemented to allow utilization of reconfigurable optical connections in the DCN.

All the mentioned test-cases will be evaluated according to KPIs specified in this document.

Legal Notice

The information in this document is subject to change without notice.

The Members of the COSIGN Consortium make no warranty of any kind with regard to this document, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The Members of the COSIGN Consortium shall not be held liable for errors contained herein or direct, indirect, special, incidental or consequential damages in connection with the furnishing, performance, or use of this material.

Possible inaccuracies of information are under the responsibility of the project. This report reflects solely the views of its authors. The European Commission is not liable for any use that may be made of the information contained therein.

Document Information

Status and Version:	D5.0_v4_final	
Date of Issue:	21/12/2015	
Dissemination level:	Public	
Author(s):	**Name**	**Partner**
	Michael Galili	DTU
	Valerija Kamchevska	DTU
	Cosmin Caba	DTU
	Anna Manolova Fagertun	DTU
	Yaniv Ben-Itzhak	IBM
	Salvatore Spadaro	UPC
	Albert Pagès	UPC
	Fernando Agraz	UPC
	Giada Landi	NXW
	Giacomo Bernini	NXW
	Chris Jackson	UNIVBRIS
	Yanni Ou	UNIVBRIS
	Dimitra Simenidou	UNIVBRIS
	Nicola Calabreta	TUe
	Oded Raz	TUe
	Domenico Gallico	IRT
	Alessandro Predieri	IRT
	Matteo Biancani	IRT
	José Aznar	I2CAT
	Amaia Legarrea	I2CAT
Edited by:	Michael Galili	DTU
Reviewed by :	Alessandro Predieri	IRT
	Chris Jackson	UNIVBRIS
Checked by:	Sarah Ruepp	DTU

Table of Contents

1 Introduction

This document describes the use-cases and demonstrator plans to be pursued in the COSIGN project. The document lays out the key elements of the planned demonstrators in COSIGN and describes demonstrator requirements and available components and facilities for the demonstrators.

The document is organized as follows:
Section 2 summarizes the overall COSIGN architecture and general approach to advancing the state-of-the-art in data centre networks.
Section 3 describes the use-cases which will be addressed in the demonstrator work and describes the concrete test-cases and performance metrics which will be applied.
Section 4 describes the architecture and proposed implementation of the short- and medium-term demonstrator scenarios in COSIGN.
Section 5 describes the long-term demonstrator architectures which will be implemented. Plans for the concrete implementations are also presented.
Finally, Section 6 summarizes and concludes the document.

1.1 Reference Material

1.1.1 Reference Documents

[1]	http://oss.oetiker.ch/smokeping/index.en.html
[2]	based on: https://blog.equinix.com/2011/11/latency-bandwidth-keys-to-user-experience/
[3]	Takara, H. et al. “1.01-Pb/s (12 SDM/222 WDM/456 Gb/s) crosstalk-managed transmission with 91.4-b/s/Hz aggregate spectral efficiency”. *Proc. European Conference on Optical Communications 2012 (ECOC 2012),* postdeadline paper Th3.C.1 (2012).
[4]	COSIGN Deliverable D2.3
[5]	V. Kamchevska et al., “Experimental Demonstration of Multidimensional Switching Nodes for All-Optical Data Centre Networs,” in *Proc. ECOC 2015*, Tu.1.2.2, Valencia, Spain, 2015.
[6]	Anna M. Fagertun, Michael Berger, Sarah Ruepp, Valerija Kamchevska, Michael Galili, Leif K. Oxenløwe and Lars Dittmann, “Ring-based All-Optical Datacenter Networks”, P.6.9, ECOC 2015.
[7]	Plexxi, “Plexxi Pod Switch Interconnect”, Datasheet available at: http://www.plexxi.com/wp-content/uploads/2013/11/Plexxi-Pod-Switch-Interconnect.pdf
[8]	Dongxu Zhang, Tingting Yang, Hongxiang Guo, Jian Wu, “Enabling Traffic Optimized Topology Reconstruction with the Optical Switching based Small World Data Center Network”, P.6.6, ECOC 2015.

1.1.2 Acronyms and Abbreviations

The most frequently used acronyms in the Deliverable are listed below. Additional acronyms may be defined and used throughout the text.

CPU	Central Processing Unit
DC	Data Centre
DCN	Data Centre Network
DoW	Description of Work
DSP	Digital Signal Processing
HTTP	HyperText Transfer Protocol
IaaS	Internet as a Service

LN	Lithium Niobate
MIMO	Multiple Input Multiple Output
NIC	Network Interface Controller
ODL	OpenDaylight
OF	Open Flow
OVN	Open Virtual Network
OVS	Open vSwitch
PLZT	(Pb,La)(Zr,Ti)O3
QSFP	Quad Small Form-factor Pluggable
RTT	Round Trip Time
SDM	Space Domain Multiplexing
SDN	Software Defined Networks
TDM	Time Domain Multiplexing
ToR	Top Of Rack
VDC	Virtual Data Centre
VLAN	Virtual Local Area Network
VM	Virtual Machine
WAN	Wide Area Network

1.2 Document History

Version	Date	Authors	Comment
01	08/12/2015	See the list of authors	Integrated version with some contributions still not received
02	13/12/2015		Integrated version for internal review containing all contributions
03	17/12/2015		Version integrating reviewer comments
04	21/12/2015		Final version

2 Description of the COSIGN architecture

The COSIGN architecture is the major outcome of WP1. There, partners identified application, business and functional requirements to study and architect the possible future intra-DC control and management mechanisms based on a converged IT and network orchestrator that operates with the future DCN network technologies following the SDN paradigm. Novel data plane approaches are developed and proposed.

The proposed architecture (see the high level model on Figure 1) aims to introduce disruptive transformations in the data plane, significant advances to the control plane and major innovations in the DC management and service orchestration layers, which will drive the evolution of the Data Centres Networks (DCNs).

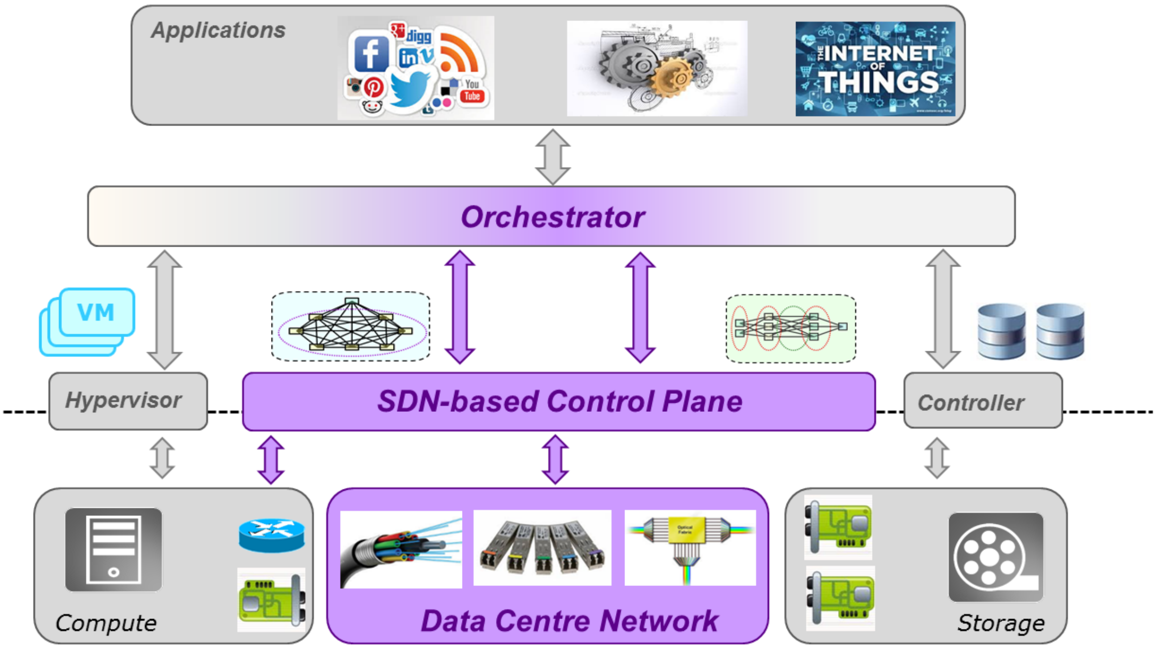


Figure 1: High level COSIGN architecture model view.

Additionally, the COSIGN architecture model has been established to serve as reference for the rest of the work packages while accommodating future intra-DC workloads and improving ToR, Cluster and Core switches (WP2), identifying the control plane features to expose a logical, abstracted view of the network resources towards the orchestrator (WP3) and also defines the orchestrator entity to automate the provisioning of IT and DCN resources in a unified way. We have demonstrated the benefits that the adoption of novel technologies brings to the different layers of the architecture. Furthermore, we have shown how improved resource virtualisation and provisioning, resulting from intelligent integration of the layers, benefits applications and services

Finally, WP1 has also identified the roadmap evolution and adoption of future SDN based optical DCNs and the general vision consisting of a sequential migration (based on the technologies available in COSIGN consortium – see Figure 2) from electronic towards all-optical DCNs through several progressive scenarios. The short, medium and long-term proposed scenarios refer to a prospective horizon for deployment/implementation to be demonstrated in WP5 through the different test cases that are being prepared.

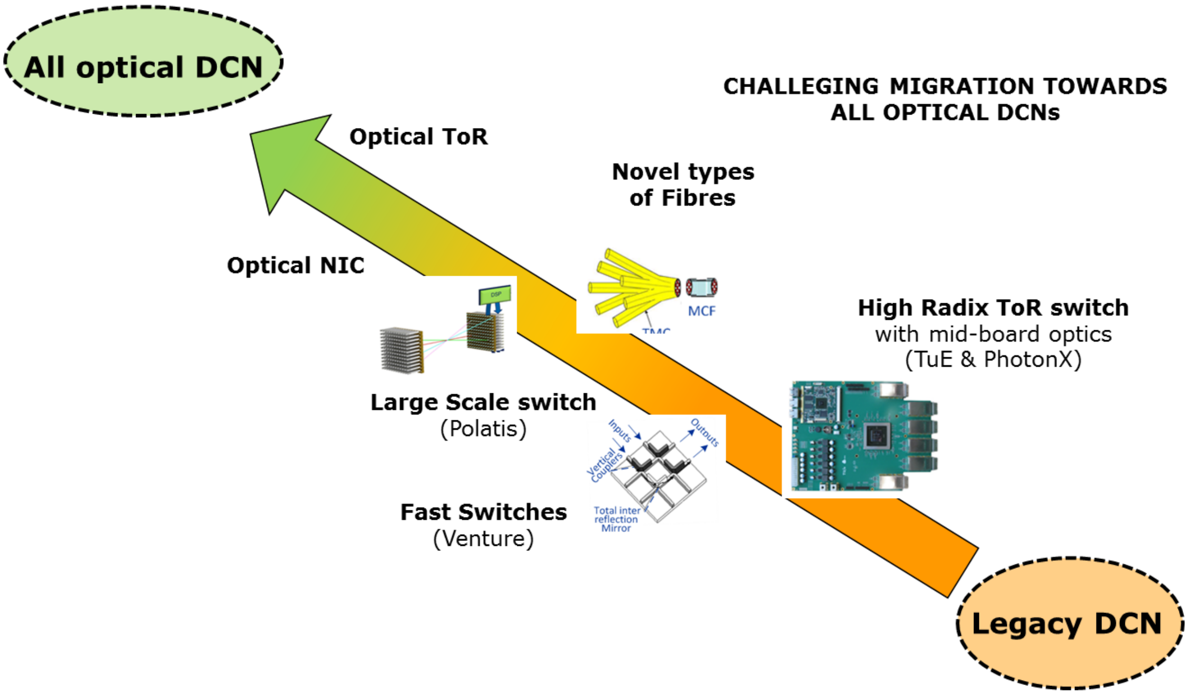

All optical DCN
CHALLEGING MIGRATION TOWARDS ALL OPTICAL DCNs
Optical ToR
Novel types of Fibres
Optical NIC
MCF
TMC
DSP
High Radix ToR switch with mid-board optics (TuE & PhotonX)
Large Scale switch (Polatis)
Inputs
Outputs
Vertical Couplers
Fast Switches (Venture)
Total inter reflection Mirror
Legacy DCN

Figure 2: COSIGN architecture design vision: progressively introduction of optical data plane resources towards an all-optical DCN approach. The arrow indicates the evolution towards more optical data plane technologies and increased performance.

2.1.1 Short term scenario: Legacy networks in combination with the TU/e ToR

In the short-term scenario, the TU/e ToR electronic switch (described in D2.2) with on-board optics is employed to replace the traditional ToR switch but still follow the layered architecture. The switch is still essentially an electronic switch but several major improvements have been made in both hardware and software aspects in order to:

- Enable great savings in power consumption
- Save costs by reducing the number of optical modules
- Switch size reduction
- Run OpenFlow, which means they can be controller using an SDN controller. This will lower the power consumption of the switching ASIC (SW change).

2.1.2 Mid-term scenario: Introducing the optical overlay

The mid-term scenario includes all of the advantages of the short-term scenario but offers an additional optical overlay network based on Polatis switches, which can be configured in a dynamic manner. Since the optical overlay switch is also SDN controlled, one can orchestrate the ToRs and overlay switches together for specific use cases.

Thus, combined with TU/e ToR, a more flattened architecture is designed, introducing large scale optical cluster switch.

Comparing with the layered structure, an optical-enabled flattened DCN architecture could avoid the bandwidth bottleneck caused in the aggregation/core switch by increasing the connectivity of ToR switch and further reduce the traffic latency as well. The description details of the short and medium-term demonstration scenarios are provided in section 4.

2.1.3 Long-term scenario (Towards all optical scenarios)

Finally, long term scenarios have also been identified which aim to augment the data plane by adding novel optical technologies including fast optical switching, SDM-type fibres and WDM technology. In this scenario, the overall architecture approach is still valid but it is foreseen the need to extend the control plane to support new features and the new devices coming from the data plane, hiding complexity and making them available at the orchestration level. The specifics of the long term demonstrator descriptions and implementation details are provided in section 5.

3 Description of use cases addressed by the demonstrators

3.1 Test case 1 – Virtual Application

In this section we describe the scenarios, which will be used to demonstrate in a visual way the benefit of optical circuits (in terms of both latency and bandwidth) for the selected elephant flows in the Virtual Application (vApp) use case. To that end, we are going to use either visual ping or video streaming scenarios, which are described below.

3.1.1 Visual Ping

Ping is usually used to measure Round-Trip-Time, which indicates the latency between two hosts. Therefore, running repeatable ping with visualization of the output RTT (e.g., smokeping [1]) can be used in order to demonstrate the latency improvement of the optical circuit compared to a traditional Ethernet-based network.

3.1.2 Video Streaming

Video streaming offers a better solution as compared to pre-download a video file: prerecorded or live video starts playing almost immediately, does not take up any permanent storage, and can be made available/unavailable as needed. However, streaming video is much harder on the network.

The typical way that video is delivered is through a continuous sequence of small HTTP downloads. Under perfect conditions, the video is transferred over the network at exactly the same rate as it is played back. If the network cannot keep up with video playback, the player usually switches to a lower bandwidth / lower quality version of the video stream. But lack of raw bandwidth is rarely a problem: a reasonable video quality can be achieved at rates of one or two megabits per second. The real challenges with streaming video are packet loss and excessive buffering.

In some cases network performance can be poor (due to other applications concurrently imposing load on the network), such that new data does not arrive before the buffer is empty, and then the video has to pause while the player waits for more data. To avoid having to pause again very quickly, the application first fills up its buffer again before it resumes playing. Some applications indicate that they are “re buffering” at this point.

3.1.2.1 How latency and bandwidth affect video Streaming user experience [2]

To assess the effect of bandwidth and latency on user experience, a simulated WAN environment is used, which allows to adjust both (see Figure 3). It is worth noting that this only produces a one-way stream of content. For distributed and/or chatty applications, the impact of latency is multiplied.

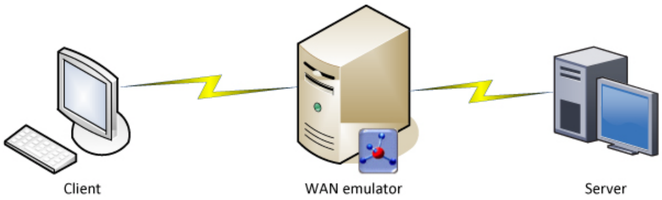


Figure 3: Network Simulation Environment

WANem utility is used (http://wanem.sourceforge.net) as emulator. For testing purposes, 1Gbps connection is set – essentially unlimited bandwidth. Then, the latency is stepped up from 0 to 160ms. These tests are repeated with bandwidths of 100Mbps, 6Mbps, and 2Mbps. Not surprisingly, when latency increased both video and audio began to get choppy. The results were similar as the available bandwidth is dropped. The test results are presented in Table 1 along with links to the video captures. Since the original clip is 30 seconds long, the playback time is an additional indication of how much the video was affected. (Due to some initial buffering, problems are not always apparent until a few seconds in.) Latency effects became increasingly pronounced as the available bandwidth approached the bit rate of the video. Finally, when the available bandwidth is dropped below the video bit rate the results were terrible.

Table 1: Video Streaming User Experience vs. Network Latency and Bandwidth

Available Bandwidth	Round-Trip Latency (ms)	Playback Time (m:ss)	Link to Captured Video	Notes
1 Gbps	0	0:30	http://www.youtube.com/watch?v=NmHi7-QLlh4	Smooth video & audio
1 Gbps	25	0:30	http://www.youtube.com/watch?v=0ejJNu4rx4A	Still good playback
1 Gbps	50	0:35	http://www.youtube.com/watch?v=L_XnDYBE8tQ	Audio gets choppy
1 Gbps	80	0:55	http://www.youtube.com/watch?v=NFQ3MnHasMc	Both video and audio are choppy
1 Gbps	160	1:52	http://www.youtube.com/watch?v=Z2Om2XW3hOw	Extremely choppy; awful
100 Mbps	50	0:33	http://www.youtube.com/watch?v=5UrNhfLIIZc	Similar to 1Gbps results with same settings
100 Mbps	80	0:57	http://www.youtube.com/watch?v=f4QrAOGMftA	Similar to 1Gbps results with same settings
6 Mbps	0	0:30	http://www.youtube.com/watch?v=SIO2v4nuIi0	Good video & audio
6 Mbps	12	0:34	http://www.youtube.com/watch?v=j1mqjvtPLfk	Some audio stutter
6 Mbps	25	0:44	http://www.youtube.com/watch?v=PyfHRNHBnR0	Bad audio; video okay
2 Mbps	0	1:35	http://www.youtube.com/watch?v=_isXG648vNM	Unwatchable

3.1.2.2 How Video Streaming will be used to demonstrate vApp

In order to demonstrate vApp benefits, a video stream is transmitted between two guests, each running within different host/compute node in the COSIGN test-bed. Since video streaming usually last at least several seconds, its corresponding network flow will be detected as elephant flow by the virtual observer. Then, in order to improve the video streaming latency and bandwidth flow, the physical observer can assign an optical circuit to this flow. Hence, improving the user experience and demonstrate visually the optical circuit network benefit.

3.1.3 Test bed requirements

In this use-case, at least two physical servers are required in order to transmit at least one flow which benefits from the optical circuit shortcut. Furthermore, additional hardware resources are required for the control and management planes.

Therefore, the test-bed hardware requirements are:

- Two devstack compute physical nodes: each with 1CPU/2GB

And additional hardware requirements for the control and management planes:

- Virtual observer: 1CPU/1GB
- Physical observer: 1CPU/2GB
- devstack controller: 2CPU/4GB
- ODL: 2CPU/4GB

The required software include: devstack controller and compute nodes with OVN support, physical and virtual observer (which includes sFlow collectors), and ODL.

The benefit of optical bypasses in the data plane is expected to be significant when the Ethernet switches, which directly connect to the optical switch, are highly loaded. To that end, other servers are required in order to generate background traffic to overload the corresponding switches. Alternatively, one can logically slice the physical switch in order to both overload the switches and overcome the lack of TU/e switches resources, as described in the following sub-section.

3.1.3.1 Logical Switch Slicing

In order to overcome the limited TU/e switch resources for the demonstrators, we need to logically "slice" each physical switch into several logical switches. To that end, ports of the same switch can be physically connected by an optical link. Based on that physical connection, the switch is configured with corresponding open-flow rules. For instance, Figure 4 demonstrates the "slicing" of the switch into two logical switches. Whenever a packet is required to be transmitted from *Logical Switch 1* to *Logical Switch 2*, the packet is transmitted through port 32, which in turn transmits the packet through the physical optical link back to port 33, which is considered to be *Logical Switch 2*. Therefore, such ingress packet from port 33 is considered to be transmitted from *Logical Switch 1*. Each such egress and re-ingress transmission contributes latency which consists of the optical-to-electrical conversion, queuing, packet processing, and switching. In extreme case where much higher switch hops are required, one can define logical switches which consist of only two ports to be used for receive packets on one port and forward to the other port. Hence, such minimal logical switch can be used for building up latency. However, such "slicing" technique complicates the controller decisions; it should be aware of the physical links between ports of the same switch, and how the physical topology maps into the "logical sliced" topology. Furthermore, the open-flow rules scope of each logical switch should be limited to its ports range only; e.g. *Logical Switch 1* cannot forward packets to ports higher than 32.

An additional option to the logical switch slicing is to create a closed loop in the switch, such that a flow will be transmitted through the same physical switch infinitely. To that end, the Open Flow matching rules should modify the TTL field in order to keep this flow alive in the switch. Such an approach can further increase the switch overload. Moreover, one might be able to generate such flow by a server for a limited time period, which after that the flow will be self-existed in the switch without being generated by a server.

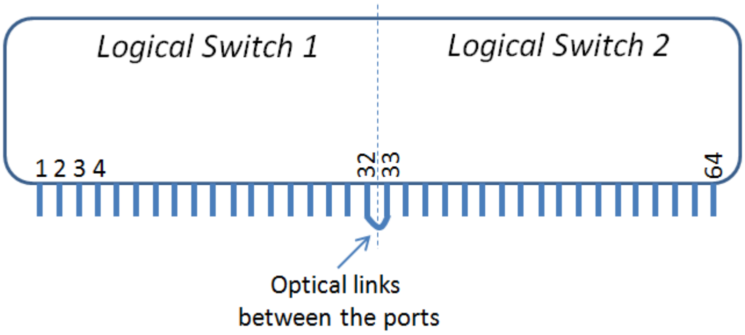


Figure 4: Logical Switch Slicing Example

3.1.3.2 OVN Compute Node Setup

Each devstack compute node with an OVN deployment must be configured with an OpenvSwitch bridge dedicated for OVN’s use, called the integration bridge. System startup scripts may create this bridge prior to starting ovn-controller if desired. If this bridge does not exist when ovn-controller starts, it will be created automatically with the default configuration suggested below. The ports on the integration bridge include:

- On any chassis, tunnel ports that OVN uses to maintain logical network connectivity ovn-controller adds, updates, and removes these tunnel ports.
- On a hypervisor, any VIFs that are to be attached to logical networks. The hypervisor itself, or the integration between Open vSwitch and the hypervisor takes care of this.

Other ports should not be attached to the integration bridge. In particular, physical ports attached to the underlay network must not be attached to the integration bridge. Underlay physical ports should instead be attached to a separate OpenvSwitch bridge.

3.1.4 KPIs and Benchmarks

The benchmark for all the KPIs is the same and can be either traffic traces from data centres (if they become available), or on-line client-server video streaming application.

1. Network improvement in terms of bandwidth/latency for different flow types (Mice vs. Elephants).
   Usually will be obtained by the used application/benchmark, or by employing synthetic network probe along with the application/benchmark.
2. Number of used optical circuits.
   Depends on the path calculation by the physical observer
3. Utilization of the used optical circuits
   Can be obtained by sFlow statistics from the TU/e switch which are directly connected to the optical switch.
4. Time to establish optical circuit path (benchmark: manual config.)
   Time breakdown to:
   i. Physical observer time to calculate/reuse an optical circuit - reported by the physical observer.
   ii. Time to compute the optical path by the ODL - reported by ODL.
   iii. Time to provision the path - reported by ODL.

3.2 Test case 2 – Virtual Data Centre Provisioning

This section describes the test scenario that aims to demonstrate the capacity of the COSIGN architecture to request and configure complex virtual optical slices, comprising both compute resources and network assets. This use case, identified in the COSIGN project as Virtual Data Centre (VDC) use case, serves the purpose to allow the data centre (DC) operator to lease part of his infrastructure to external entities (tenants) so they can exploit innovative Infrastructure as a Service (IaaS) solutions to develop their own business models. For this, a tenant requests for a virtual infrastructure represented by some virtual nodes with computational capabilities, i.e. a number of virtual machines (VMs), and virtual links interconnecting the virtual nodes, stating the desired capacity in terms of bit-rate. Then, it is the responsibility of the DC operator to find the most suitable mapping of the requested virtual resources onto physical ones: servers for the VMs and network resources for the virtual links.

In this regard, the proposed test scenario involves the on-demand creation of VDC instances on top of a shared DC infrastructure. There, co-existing VDC instances are mapped to the physical resources with the purpose to satisfy the requested resources and the isolation between the different tenants (see Figure 5). In this regard, the proposed test will be used to assess that independent configuration of the different VDC instances is provided, maintaining the isolation between the different virtual infrastructures. Moreover, in order to fully exploit the capabilities of the underlying physical infrastructure, the COSIGN orchestrator implements a set of provisioning algorithms that provide the most optimal mapping of the virtual slices onto the physical resources. The key feature of such algorithms is that they provide a joint optimization of both compute and network resources.

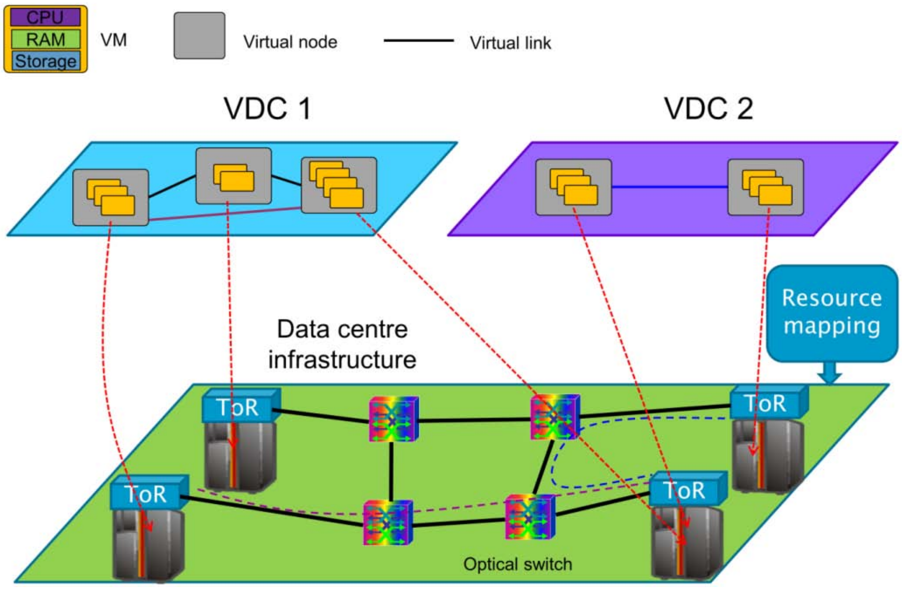


Figure 5: Virtual Data Centre provisioning scenario

The following section details the full workflow of the demonstration and validation test as well as all involved entities.

3.2.1 Work flows

The starting point for the demonstration will be the dashboard service of the orchestrator. There, a new VDC instance will be specified utilizing the graphical editor of the dashboard (e.g. OpenStack Horizon). For the new VDC instance, it will be possible to specify the quantity of the virtual nodes as well as the number of virtual machines per virtual node. For each virtual machine, it will be possible to state the desired resources in terms of CPU cores, storage capacity and memory (e.g. utilizing the OpenStack flavour templates) as well as the desired image for the operating system. Once the virtual nodes have been specified, the connectivity between the virtual nodes will be specified as well in terms of virtual links. In this regard, it will be possible to state both the source/destination and bit-rate for each virtual link.

When the VDC request has been fully specified, the dashboard service will contact the orchestrator core, asking for the provisioning of a virtual slice that satisfies the requirements of the VDC request. In this regard, the orchestrator core (Heat) will contact the provisioning algorithms, which are the responsible for deciding on the optimal mapping of virtual resources to physical ones. The algorithms module will utilize the topological and resource utilization that can fetch from the physical infrastructure controllers (Nova for the compute and Neutron for the network) and will run the algorithm using this information as input. The output will be passed to the services responsible for provisioning the resources. On one hand, Nova will take care of the VM allocation onto the physical servers, while, on the other hand, Neutron will collaborate with the OpenDaylight controller to configure and establish the optical slice that corresponds to the realization of the virtual links.

Once the whole VDC instance has been successfully deployed, the dashboard service will be informed, providing the details about the provisioned virtual infrastructure as well as the corresponding mapping. After this process, new VDC instance can be configured onto the same physical infrastructure. For this, the dashboard service allows for independent request of VDC instances in order to keep separated the requests that belong to different tenants. The process for provisioning a new VDC request is the same, with the difference that the physical infrastructure utilization has been changed. In this regard, the algorithms module will have to be updated with the newest utilization information in order to adjust the mapping decisions to the current status of the physical infrastructure.

3.2.2 Test bed requirements

Figure 6 shows the topology of the testbed which is used for the demonstration and experimental evaluation of the VDC use-case.

The data plane is composed of 2 POLATIS switches and 4 TU/e ToR switches interconnected as shown in the picture with the blue lines. Each ToR switch is interconnected to 2 OpenStack Compute Nodes (i.e. servers), for a total of 8 available compute nodes. 6 of them are interconnected through 10Gbps link, while 2 of them are interconnected through 1Gbps link. These interconnections are represented in the picture through the red lines.

At the control and orchestration plane, a server is dedicated to the OpenStack Controller Node while another server runs the OpenDaylight SDN Controller. Each entity is also interconnected through a management and control network which is required for the interaction between OpenStack Controller Node and OpenStack Compute Nodes and for the interaction between OpenDaylight and the network data plane elements, as shown by the dotted back lines. The OpenDayligth controller interacts also with all the Compute Nodes for the configuration of their OVS instances, but these interactions are not represented in the picture for readability reasons.

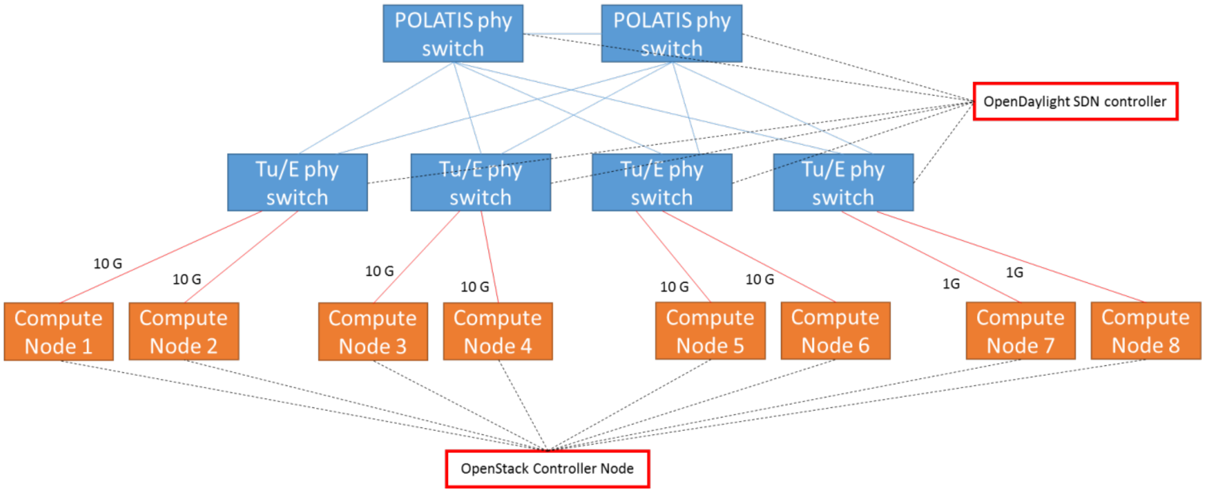


Figure 6: Topology for VDC testbed

It should be noted that all the computing nodes represented in the picture are actually deployed as Virtual Machines (VMs) on the Interoute physical infrastructure (discussed in section 4.1.1). In order to exploit the two 10G NICs available in each of the 3 physical servers, the compute nodes from Compute Node 1 to Compute Node 6 need to be instantiated in pairs of two VMs in each physical server. Moreover, the internal configuration of virtual switch in each server must be configured so that the traffic from each VM representing a compute node is directed to one of the 10G network interfaces. Compute Node 7 and Compute Node 8, as well as the OpenStack Controller Node and the OpenDaylight SDN controller can be instantiated without any location constraint and the associated traffic can use the 1G network interfaces in the servers. In terms of capabilities, the VMs for the SDN controller and the OpenStack controller node should have 2 CPUs and 4GB, while 2 GB are enough for the compute nodes.

In terms of physical infrastructure, the deployment of the interconnections between Compute Nodes deployed on VMs and TU/e switches, as described above, requires the usage of an additional TU/e switch in the physical infrastructure, statically configured and fully transparent for the COSIGN system. In particular, the traffic from the VMs which will host the Compute Nodes will be tagged with different VLANs (still through a static configuration: VLAN X for the traffic from Compute Node X) and will exit from a given, statically configured, network interface as shown in Figure 7. Based on this VLAN, the Infrastructure TU/e switch will forward the traffic on the suitable outgoing port which is connected to the desired TU/e switch of the COSIGN data plane.

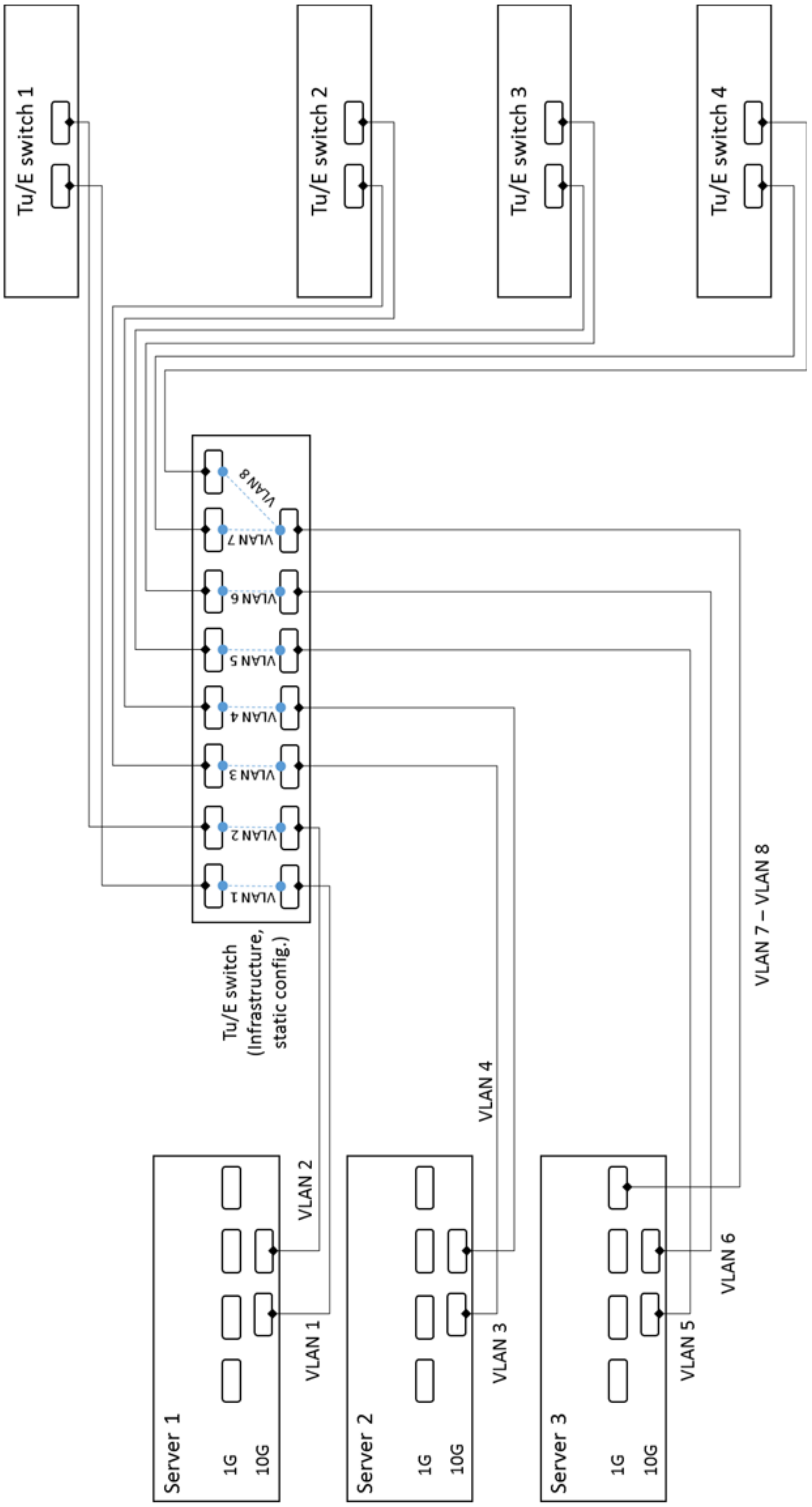


Figure 7: Testbed for VDC use case – simplified physical interconnections between servers and ToRs

3.2.3 KPIs and Benchmarks

The main KPIs that will be assessed in the proposed test are the following:

1. Time to provision a new VDC instance. This KPI will be experimentally measured in the demonstrator as an output of the efforts done in WP5, breaking up the provisioning time in its

several components to better appreciate the individual contributions of the different software modules in the whole provisioning operation. Basically, the provisioning time can be divided in:

- Time to calculate the mapping of the VDC instance by the orchestrator algorithms.
- Time to configure and deploy the VMs.
- Time to configure the optical slice.

2. Execution time for commands to operate the VDC instance (optional). Once a VDC instance has been deployed, it will be tested how long it takes to perform configuration commands over the provisioned virtual slices as a means to achieve a customized configuration of the virtual network
3. Isolation among VDC instances (optional). To assess that independent virtual slices are being provisioned, a test of the isolation between VDC instances will be performed. For this, two different VDC instances will run on top bandwidth hungry applications (e.g. high-quality video streaming), showing that no degradation on the service is appreciated once both application run in parallel.

The benchmarks for the KPIs are the following:

1. Provisioning times found in commercial DC services provided by real DC operators (possibly to be provided by IRT).
2. Compare the utilization of the physical resources considering both a joint IT and network mapping process as well as one that maps them separately.

3.3 Test case 3 – Virtual Data Centre VM migration

This test case aims to further demonstrate that the requirements for Virtual Data Centres have been achieved. As well as orchestration and allocation, there must also be support for runtime service recovery of VDC instances. Furthermore, we aim to demonstrate that the addition of optical circuit cut-through increases the performance of a VDC host in terms of how quickly Virtual Network instances can be redeployed in the event of physical hardware failure.

3.3.1 Work flows

Initially the network must allocate server nodes for virtual machines as well switches to support the virtual network. In this test case we do not wish to measure the performance of allocation as this is tested in case 2, see Section 3.2. Hence, in this case we will assume that a VDC instance and all relevant resources have already been successfully allocated.

The workflows will involve the introduction or simulation of an error condition in the VDC instance. The result of the condition is that the VDC instance must coordinate the network hypervisors and SDN controller to redeploy the virtual components. In summary the actions of the COSIGN orchestrator and SDN controller in this scenario are as follows:

1. Physical fault detected or predicted by the hypervisors (server failure) or SDN controller (switch failure).
2. The orchestrator is notified of the fault and determines to take action.
3. The VDC tenant is notified of impending live migration.
4. VMs instances cloned and moved to alternative servers.
5. Alternative switches that connect the new servers are programmed with flows to implement the virtual network.
6. The new VM instances are started.
7. The initial VM instances are halted.
8. The initial switches have their flows removed.

9. The VDC tenant is notified of successful migration.

3.3.2 Test bed requirements

The testbed must be able to support communicating VM instances that run on separate server nodes, supported by a virtual network. There must be sufficient resources to replicate this virtual infrastructure on separate hardware to the initial instances. The following diagram illustrates the test case:

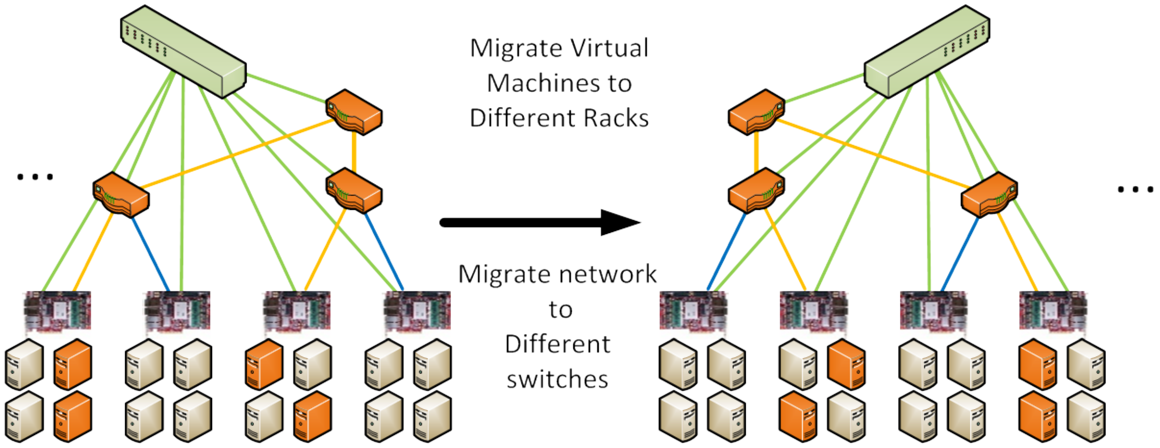


Figure 8: VDC VM migration use case

The migration of both network and virtual machine instances can be seen, though this precise arrangement is not necessary. Here, a VDC spread over 2 racks, using 4 physical servers and three packet switches is demonstrated.

Because we are interested in the performance of the migration process, the VM cluster has low performance requirements in terms of servers. The executing application(s) need only demonstrate the connectivity between the VMs.

3.3.3 KPIs and Benchmarks

We will measure the total latency to migrate a Virtual Machine. The components of this latency:

- Time to halt the VM
- Time to copy from source to destination node
- Time to restart the VM

By measuring all the components separately we will see what proportion of the time of a VM migration is dependent on the network and therefore the maximum amount of performance improvement that can be gained by increased network performance.

These measurements will be taken with and without circuit cut through enabled. We should vary:

- the size of the VM(s)
- the network congestion level
- the complexity of the virtual network (number for flows required to implement it)

This will allows us to better understand under what circumstance VM migration benefits from optical circuit switching.

4 Short/medium term demonstrator

This section describes the DCN architectures which will be implemented as the short- and medium-term demonstrators. These two demonstrators will host most of the use case demonstrations in COSIGN.

4.1 Description of the architecture, topology and planned implementation

The architectures considered in the short/medium term demonstrators are shown in Figure 9. The left-hand side architecture will be used for the short term demonstrator. This is a spine-leaf architecture that will make use of the Ethernet switches built at TU/e. The medium term demonstrator aims towards introducing an optical circuit switch (i.e. a Polatis switch) as shown on the right-hand side of Figure 9. For both demonstrators the different use cases will be implemented and the relevant KPIs will be quantitatively assessed. In order for this to be achieved, the demonstrator relies on previously developed and working hardware/software prototypes. Successful integration tests among the different hardware or among the different software components are essential for the demonstrator to perform vertical integration. The following subsections describe in details the physical equipment needed as well as the software requirements in order to implement the different use cases in the short/medium term demonstrator.

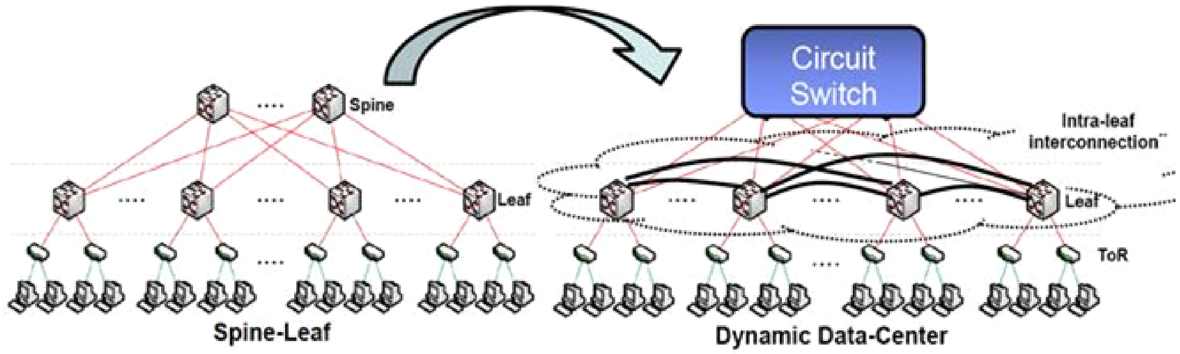


Figure 9: Short term and medium term data plane architectures

4.1.1 Test bed specifications

Both the short term and the medium term demonstrators require hardware and software components for the vertical integration demonstration, as described in details below.

The data plane for the short/medium term demonstrator is composed of servers used to generate traffic, TU/e Ethernet switches used for interconnection and for the medium term demonstrator, one or several Polatis switches acting as spine switch. An example of a simplified data plane demonstrator test-bed with the hardware components used is given in Figure 10. It is important to note that due to the limited number of hardware components available, such as servers or TU/e Ethernet switches, the demonstrator may have slightly different component deployment as compared to the test-beds shown in Figure 10.

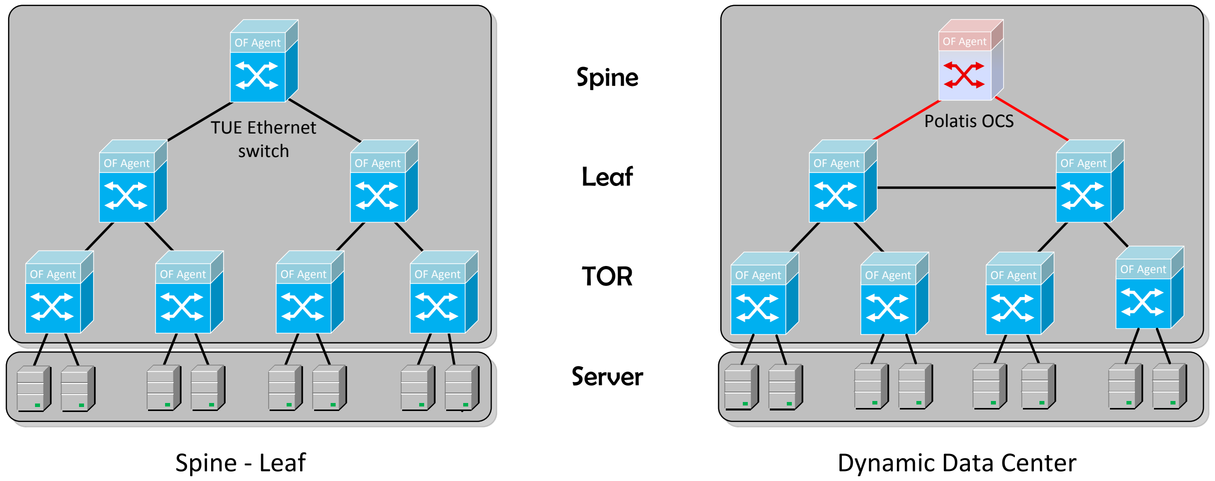


Figure 10: Short term and medium term data plane demonstrator test-beds

Interoute Testbed and Servers

Interoute testbed is located in the production infrastructure of the Interoute Milan PoP where a co-location area is dedicated to host servers and network equipment for the COSIGN demostrator. All Interoute co-location sites are provided with AC & DC power as standard, both of them are designed with N+1 resilience configuration in order to ensure redundancy and guarantee 99.99% power availability. Temperature and humidity control is provided by a heat rejection chiller system with N+1 redundancy configuration that ensure a temperature of 23°C (+/- 3°C) and humidity rate of 50% (+/- 10%).

The demonstration testbed provides a dedicated hardware infrastructure that will simulate a virtualized DC environment, where all the software components and hardware equipment can be tested and evaluated. To address this purpose, 3 hardware servers will be made available, each of them with the following specifications:

- (2x) 10 Core CPU
- 256 GB of Memory
- 1TB storage HDD
- (4x) 1GbE Ports + (2x) 10Gb-T Ports
- iLO Chassis Lights Out Management Card

In addition, an Ethernet Switch with 48x1GbE ports + 2x10Gb-T ports can be used to interconnect the servers.

The rack space reserved for the testbed equipment is 15 U, 5 of which will be used to host the three servers, the switch and a router, so that a free space of 10 U can be used to accommodate any other device provided by other partners (e.g. Polatis and TU/e switches, further servers, etc.).

On top of the hardware devices, a virtualized environment will be deployed and managed using OpenStack. This approach will ensure a higher level of flexibility and optimization of resources in order to significantly increase the number of Virtual Servers available for the demonstrator's purposes. The COSIGN project partners will have access to the virtualized environment managed by Interoute that will provide a transparent, isolated, fully reserved and statically configured virtual infrastructure able to emulate a DC environment, with the possibility to deploy multiple virtual servers as VM instances in the Interoute's OpenStack platform.

In Figure 11 is reported an example of topology where four TU/e Switches are linked to two different Virtual Servers using the two 10Gbit-T ports of each physical server. The number of compute nodes for each virtual server can be modified depending on the number of Virtual Machines needed to perform the demonstrator.

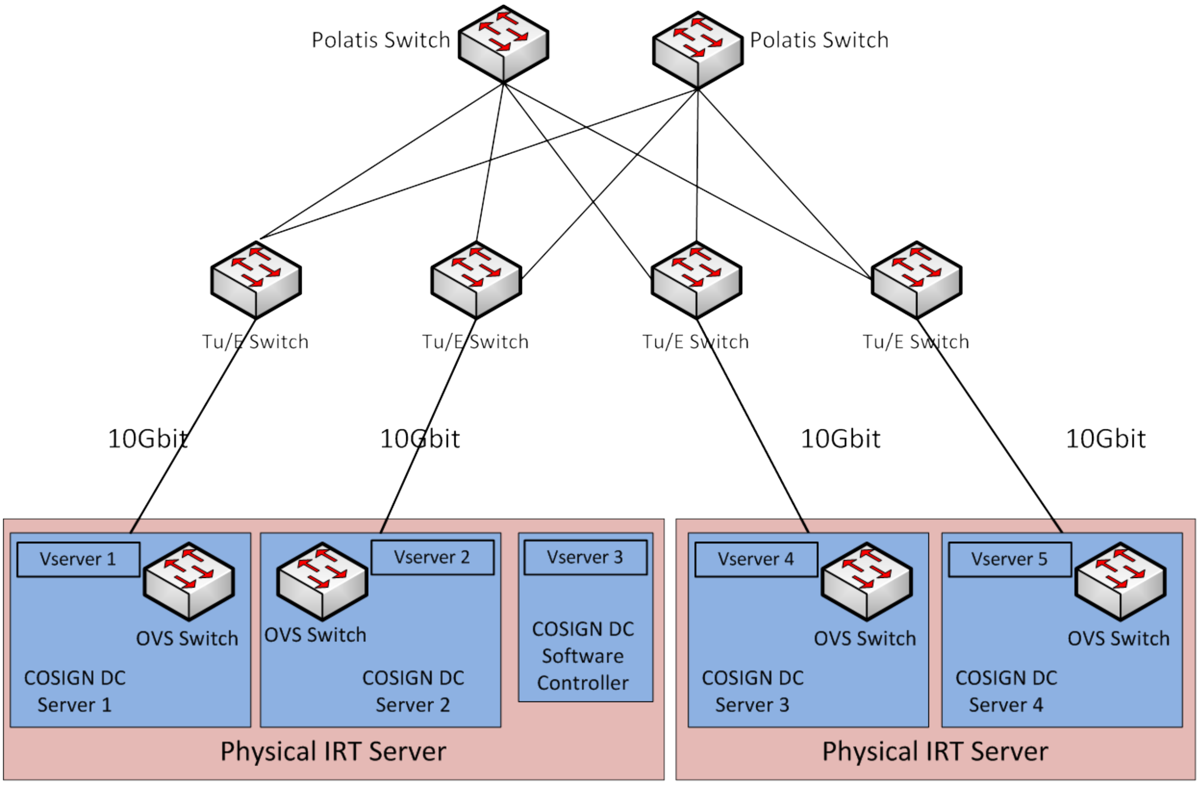


Figure 11: Topology example

The three physical servers which will be available for the demonstrator activities allow the creation of several virtual servers (up to 60 vCPUs, 768 GB of memory and 3 TB of storage is made available to the final demonstrator). Figure 7 and the associated discussion in section 3.2.2 above explains how we intend to represent a larger number of separate compute nodes by using the smaller number of more powerful servers. Using this scheme we expect to be able to emulate a sufficient number of servers to allow relevant demonstrations.

Ethernet switches

The relevant details and specifications of the TU/e Top-of-Rack (ToR) Ethernet switches, the hardware interfaces to interconnect the switches with the OCS, and the OpenFlow compatible control software are discussed below. Table 2 reports the list of available resources provided by TU/e and the quantity to be employed to build the demonstrator. The ToR switch has been designed and built based on the Trident I switch ASIC. It supports the connection of up to 64x10Gbps Ethernet links for 1st generation, and 128x10 Gbps for 2nd generation. Moreover, to solve electrical ToR switch to optical switches wavelength/fibre mismatch, new hardware has been design and fabricated. The interface allows adapting CXP transceiver at 850 nm to 2x QSFP transceivers at 1300 nm as shown in Figure 12. The main features of the ToR prototype developed by TU/e are summarized in Table 3.

Table 2: List of available resources provided by TU/e

Resources	**Available quantity**
ToR Ethernet switch - 1st generation	5
Port count	64
Data rate	10 Gb/s
Multi-mode optics - for 1st gen ToR	15 modules
Adaptable wavelength/fiber interfaces between ToR switch and optical switches	6
ToR Ethernet switch - 2nd generation	3
Port count	128
Data rate	10 Gb/s
Multi-mode optics - for 2nd gen ToR	12 modules

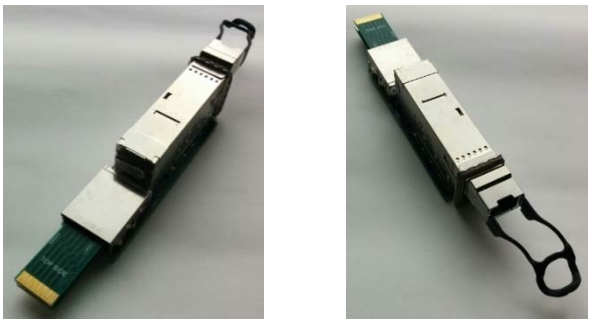


Figure 12: Adaptable wavelength/fibre interface between ToR switch and optical switches

Table 3: Main features of TU/e ToRs

ToR switch – 1st generation	
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)
SW features	Supports Open Flow 1.3.4
optical transceiver form factor	CXP
ToR switch – 2nd generation	
Size	20x20cm
Power consumption	At maximum load 150 Watts (120 w/o optical interfaces)
Number of optical interfaces	11 BOA (each one including 12 Tx and 12 Rx lanes)
SW features	Supports Open Flow 1.3.4

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. 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 13. 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. 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 13 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.

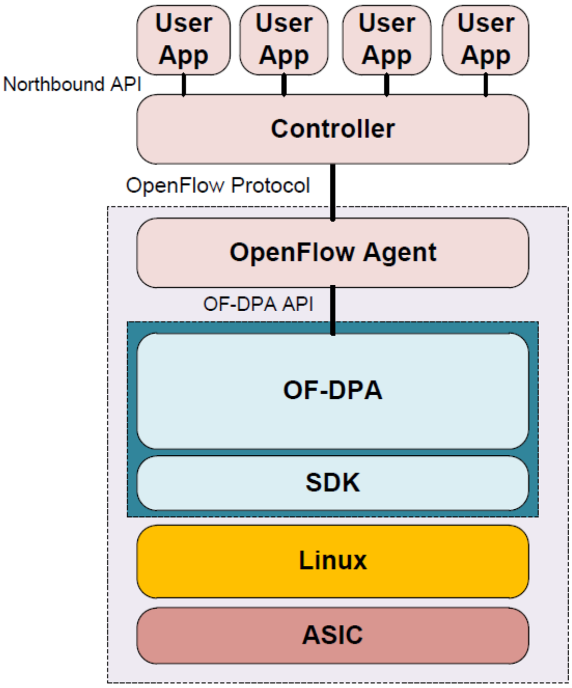


Figure 13: Broadcom OFDPA component layering

Optical circuit switches

The available quantity and port count of the Polatis switches that will be provided for the demonstrator are listed in Table 4. On the software side, the switches are compatible with an OpenDaylight Lithium SDN controller currently being customized and extended in Task 3.2 and provide support for Open Flow 1.0/1.3. An agent supporting Open Flow 1.4 is under development.

Table 4: Main features of the Polatis switch

Polatis switch	
Available quantity	Min. 2 (more switches available if relevant)
Port count	1 unit up to 384 ports, others with 48 ports
SW features	Supports Open Flow 1.0/1.3 (1.4 being developed)
SDN controller compatibility	OpenDaylight Lithium

The southbound plugins for the OpenDaylight controller have been created to interface with an OpenFlow agent and to demonstrate OpenDaylight control of the optical circuit switch. A faster and more efficient version of the OpenFlow 1.0+ agent which integrates directly with the Polatis user services API as shown in Figure 14, rather than connecting indirectly via the TL1 interface is under development. The agent size has also been reduced by a factor of ~3 by eliminating redundant code. The agent is currently being ported to the network interface on a 192x192 optical switch. It is intended to finish the development of the embedded OpenFlow agent to be compliant with the next stable release of OF 1.4/1.5, which will provide direct support for optical circuit switching, as opposed to the vendor/experimenter-specified extensions used with OF1.0+. Additional user interfaces will also be created to add REST and NETCONF support for configuration management.

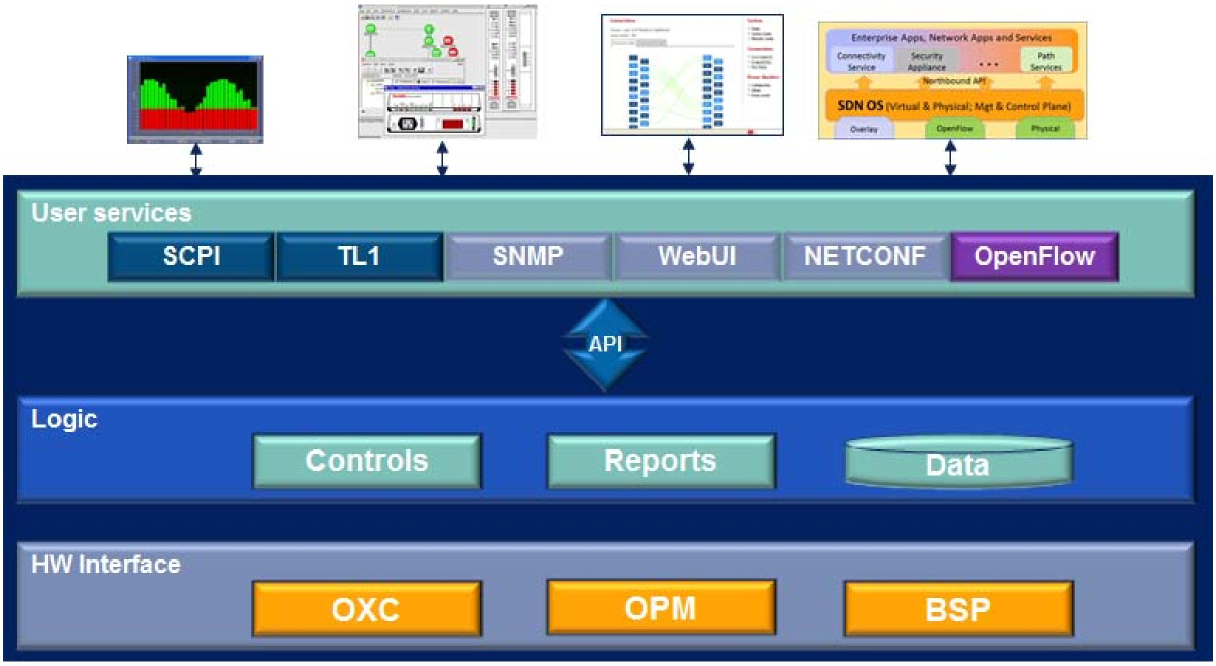


Figure 14: Polatis Optical Switch User Interface Structure

Software Requirements

We first define a set of generic software requirements for the short/medium term demonstrator, and detail only the specific requirements for each use case where needed. Figure 15 shows the hardware and software components required for the integrated demonstration in the medium term scenario. Starting from the data plane, the first requirement is to have the corresponding agents for the devices used in this demonstrator:

1) Agent for the Polatis switch.
2) Agent for the TU/e switch.

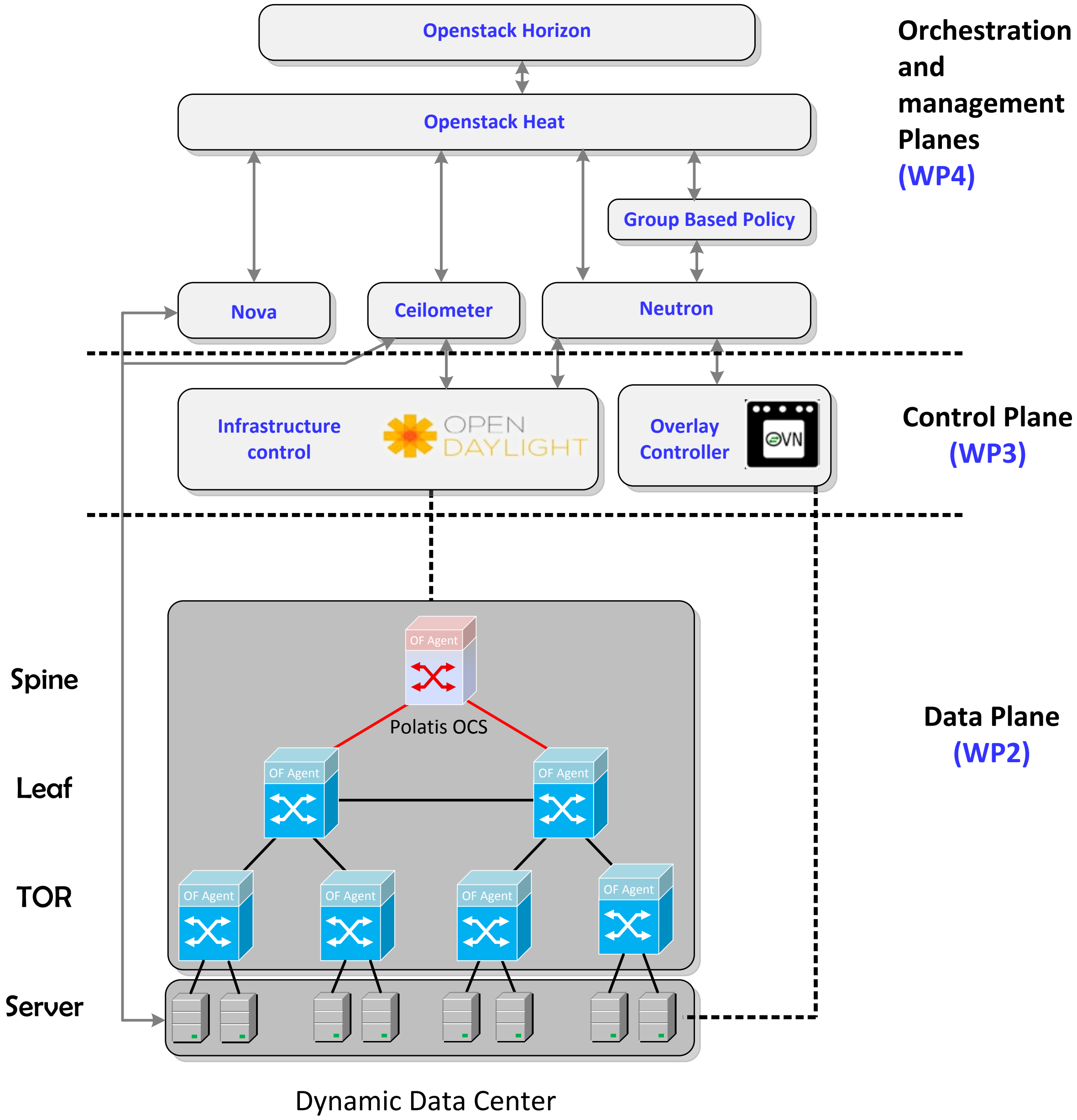


Figure 15: Vertical integration of hardware and software components in the medium term scenario

With respect to the ODL controller platform, which is using stable version Lithium, the following requirements must be met:

VDC use case requirements for the infrastructure control plane:

1) ODL has the correct southbound plugins to interoperate with the two switch agents mentioned above. Some plugins that are needed are: OF with extensions for Polatis, OF for the ToR, OVSDB for the OVS software switches in the VDC use case.
2) ODL must maintain a consistent topology view by using the features provided through the southbound plugins.
3) ODL must be able to compute paths in the network, over the consistent topology view.
4) ODL must be able to provision connectivity in the network. Further, the connections should span from OVS to OVS (Open vSwitch). The ToR to ToR connection should use either traditional Ethernet connection or optical connection.
5) ODL should also run the VTN plugin.

vApp use case requirements for the infrastructure control plane:

1) ODL has the correct southbound plugins to interoperate with the two switch agents mentioned above: OF with extensions for Polatis and OF for the ToR.
2) ODL must maintain a consistent topology view by using the features provided through the southbound plugins.
3) ODL must be able to compute paths in the network, over the consistent topology view.
4) ODL must be able to provision connectivity in the network. Further, the connections should span from ToR to ToR, either using traditional Ethernet connection or optical connection.

For the vApp use case, the OVN controller is required to control the OVS software switches thus the OVSDB plugin is not needed. Consequently, for vApp, ODL provides connectivity from ToR to ToR.

VDC requirements for the orchestrator:

1) Openstack with at least Heat, Neutron (with ML2 plugin and ODL driver), Nova and Ceilometer components, apart from the infrastructure specific components such as Keystone and the database.
2) VDC algorithms module.

vApp requirements for the orchestrator:

1) Openstack with the same basic components, as mentioned above for the VDC use case.
2) Additionally, as specified in D4.2, Openstack may need to integrate also the Group-Based Policy component.
3) Neutron must also have the OVN driver.
4) Other vApp-specific components such as the physical and logical observers.

4.2 Timeline and practical plans

The short and medium term demonstrators will be carried out jointly. The two demonstrators are closely linked and being able to do direct comparisons will have significant value.
Two testbed infrastructures will be available for integration tests and experimentation related to the demonstrators. The infrastructure provided by IRT as mentioned above will host the final demonstration. DTU will host an integration testbed for the ongoing effort of integration between software and hardware components in preparation for the final demonstrator. This testbed will provide a number of servers necessary to confirm the performance of the demonstrator and will be available for the different aspects of system integration.

Regarding the timing of the demonstrator effort, the setup of the hardware platform will commence early 2016. This will then be available to test individual components (both software and hardware) and their integration. As early versions of the integrated software platform (described above) become available the different use cases will be tested preparing for the final demonstrators.

The goal of the short-/mid-term demonstrators is the successful implementation of the test-cases outlined above. These will not necessarily be implemented simultaneously but functionality and performance of all test-cases should be tested. A key goal is to host one or more of the demonstrated test-cases at IRT’s data centre facility.

5 Long term demonstrator

The long term demonstrators in COSIGN will demonstrate implementations of all-optical DCNs as disruptive approaches to harvesting benefits of introducing advanced optical technologies in data centres. This section will present two approaches for realizing an all-optical DCN which will both be demonstrated in COSIGN.

5.1 Ring-based DCN Demonstrator

One of the long-term demonstrators that will be pursued in COSIGN implements a ring based network topology to realize a network with high flexibility and low cabling complexity. The ring based DCN will implement Space Domain Multiplexing (SDM) and Time Domain Multiplexing (TDM) to combine high capacity and high granularity of data connections in the network. Connections of different capacity tailored to fit the needs of an application can be provided based on doing optical switching in the space and time domains. The connections which can be offered in this way range between having a full optical path (fiber core or mode) continuously connecting two network nodes and establishing a connection consisting of a time slot in a frame of repeated time slots. This way the duration of the time slot and the size of the frame will determine the net bandwidth of the connection. Many connections requiring low bandwidth can thus share the same optical path in the network and a single optical transceiver e.g. in a ToR can connect to many different points of the network depending on the timing of transmission and detection. Figure 16 shows a schematic of the ring based topology combined with hierarchically structured optical switching nodes comprising optical switching functionality at both SDM and TDM level.

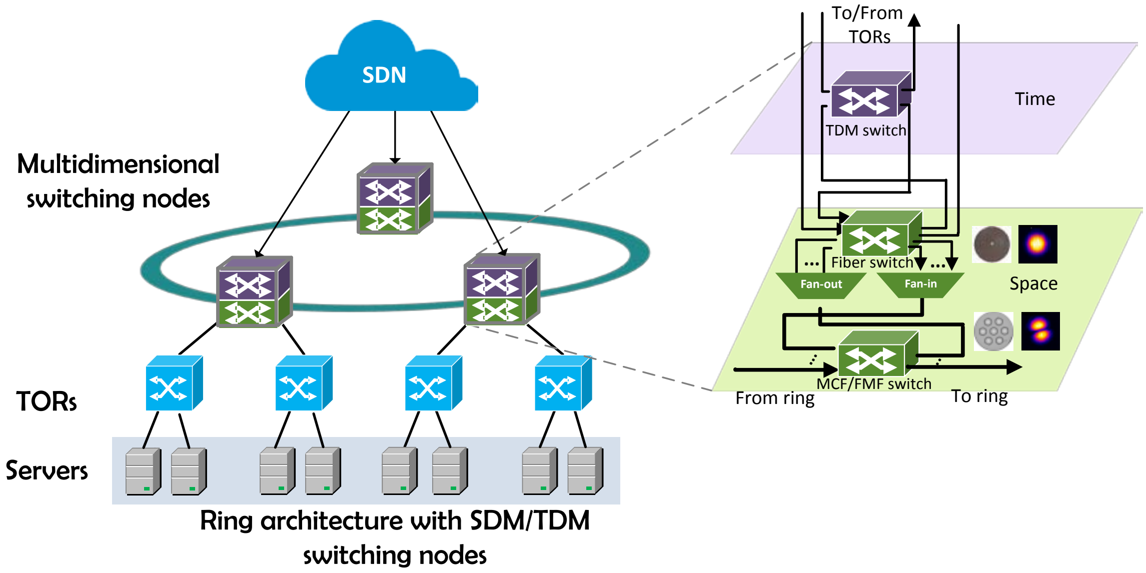


Figure 16 Proposed architecture for the long term demonstrator

Two approaches to realizing SDM will be demonstrated. The first approach relies on the use of multicore optical fibers for combining several spatially separated channels into a single strand of fiber. This technology is characterized by having very good performance in terms of optical crosstalk between channels allowing for long distance transmission of large amounts of data e.g.[3]. The other SDM technology to be demonstrated is based on optical mode multiplexing in a single fiber core. Concretely this will be investigated in FMF supporting a small number of modes (e.g. 4 LP modes) which couple only weakly, allowing for the signals contained in the modes to be detected without the need for digital signal processing to compensate the cross talk. Multicore fiber offers superior performance in terms of crosstalk compared to FMF allowing for much longer transmission distances. On the other hand multicore fiber is more complex and consequently more expensive to manufacture and handle (rotational alignment is critical for multicore fiber). In the context of data centres with many very short data links it is thus relevant to investigate both options to clarify the relevant applications of the two technologies. In Figure 16 the SDM layer is illustrated as based on MCF, which is the more mature of the two technologies. Work will also be done on employing SDM based on FMF in the bottom layer of the hierarchical switch structure. Table 5 summarizes the typical values for loss and crosstalk in MCF and FMF. Here we focus on the most promising FMF being the 4 mode step index fiber. A more comprehensive discussion of this issue is found in [4].

Table 5

	Crosstalk [dB/km]	Loss [dB/km]
4-mode step index FMF	~ -14	~ 0.2
Typ. 7-core MCF	~ -54	<0.3

5.1.1 Test bed specifications

The testbed to be used for this demonstrator scenario is being established at DTU and will comprise a number of key components described below:

- **Fast optical switch**
  The fast optical switch used in this scenario is intended to be the 4x4 switch made by Venture Photonics within the COSIGN project. The switch will act as a TDM circuit switch manipulating the time slots used to establish connections with sub-wavelength granularity. The electrical driver and high-speed control circuit for the switch is developed by DTU while the agent software and SDN controller plugin is developed by UNIVBRIS. The whole switch system comprising the optical switch, drivers and controller is compatible with OF and work is ongoing to make it compatible with ODL Lithium. Given that the Venture switch remains under development all initial experiments [5] done in preparation for this demonstrator have been carried out using alternate optical switching technologies. Thus far, LN and PLZT based switches have been used as substitute for the Venture switch. These switch types have different properties compared to the Venture switch, but for laboratory demonstration purposes they can be made functionally equivalent. In case of the LN switch polarization control is required and in both cases optical amplification must be applied to achieve lossless operation which is one of the attractive prospects of the Venture switch. As a third alternative an SOA based switch developed in the LIGHTNESS project for optical packet switching could be modified to serve the purpose of optical TDM circuit switch in COSIGN. Work on this demonstrator scenario can thus continue and if necessary be concluded while the Venture switch is being realized.

- **Fiber switch for MCF**
  A proof-of-concept switch based on Polatis' switching technology but supporting switching all cores in a MCF is being developed in a collaboration between UNISOUTH and Polatis. This switch or its functional equivalent will be incorporated as the bottom switch layer in the hierarchical switching nodes.

- **Core switch for MCF**
  An integrated silicon switch capable of coupling directly to individual cores of multicore fibers at the input and output will be used to connect a given fiber core at the input to a node to a chosen core of a MCF leaving that node. In this way connections can be established across cores in the network of MCFs. This device is being developed in the project *SimcROADM* at DTU and is made available to COSIGN where it is being integrated with control electronics and SDN interface. The agent and ODL plugins are adapted versions of the ones used for the fast switch.

- **FMF multiplexer/demultiplexer**
  Similar to the integrated core switch mentioned above a mode multiplexer/demultiplexer has been developed with the purpose of enabling transmission on several modes in a FMF with low

crosstalk over distances relevant to datacentres. These devices will enable tests of point-to-point transmission of multiple spatial channels on different fiber modes without the need for digital signal processing (DSP) to correct for mode coupling. The initial devices are expected to support stable transmission on two separate fiber modes without requiring MIMO DSP.

5.1.2 Results from simulation studies on the ring DCN

Preliminary performance evaluation was conducted for the proposed Ring-based architecture via event-based simulation. The Ring interconnect was compared to a traditional Fat Tree (FT) interconnect for an all-optical Data Centre with equal-size switch elements, performing WDM switching. The simulations served 2 main purposes. First, to illustrate the benefits of the ring-based interconnect, compared to a tree-based interconnect in an all-optical data centre. And second, to evaluate the scaling properties of the ring-based interconnect.

The results of the simulations were presented at the ECOC 2015 conference [6] and summarized in deliverable D 1.2 "*Comparative analysis of optical technologies for Intra-Data Centre networks*". The main conclusions from the comparison of the two interconnects are:

- Given the same radix-size switch as building element, the Ring of Rings (RoR) has considerably smaller footprint compared to FT – it serves the same /or more servers with almost 50% less hardware (switches and links) deployed in the network;
- the RoR outperforms the FT in terms of blocked connections;
- the RoR utilizes its resources better compared to the FT (a typical drawback of standard DC architectures are the lower utilization of the deployed resources);
- the RoR outperforms the FT across network loads and traffic distribution characteristics;

Additional justification for the rationale behind adopting a ring-based interconnect can be also found in other works from academia and the industry. Plexxi [7] offers commercial products for DC interconnect based on ring structures. The works of Dongxu Zhang et al., [8] also look into ring-based data centres. A detailed summary of the simulations is included below.

The comparative simulation study consist of simulating two different topologies, a standard fat-tree and the proposed ring topology as shown in Figure 17. Two main performance measures are considered: connection request blocking and network resource utilization. Furthermore, two scenarios are investigated. Scenario 1 looks at the performance of both topologies under different capacity availability conditions in the DCN when the amount of available wavelengths per link is varied. The ratio of within-cluster/between-cluster traffic is set to 0.125 (i.e., 1/8 of the traffic is distributed within the cluster/pod and the rest is uniformly distributed between the other clusters/pods). Scenario 2 investigates the performance of the topologies under varying traffic conditions, given fixed capacity availability in the network. In this scenario the ratio of within cluster/between-cluster traffic is varied. The goal is to evaluate under what traffic patterns the corresponding topologies perform better. This indicates the suitability of the topologies in handling different types of applications.

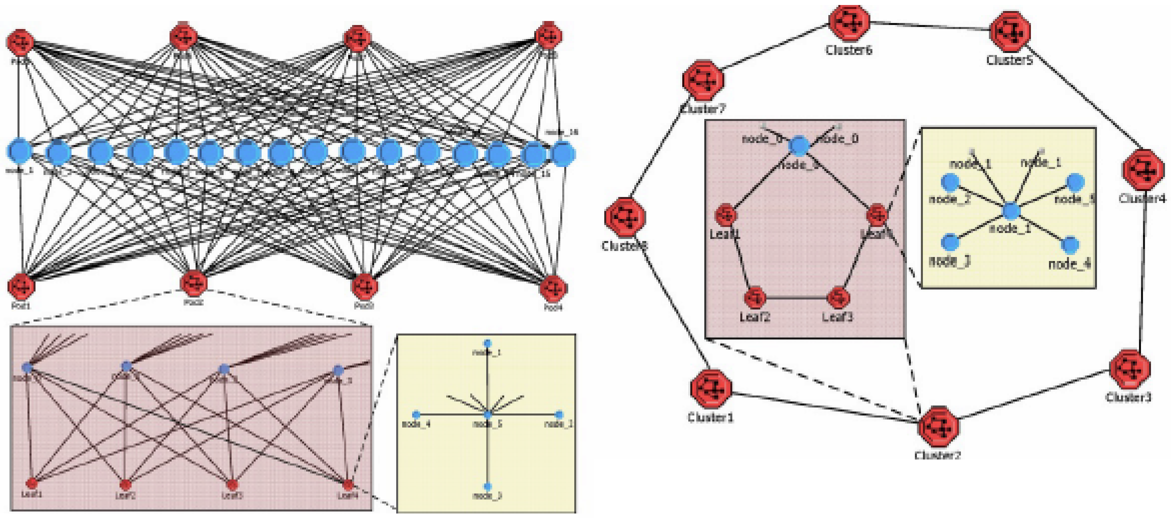


Figure 17: Simulated fat tree topology (left) and ring topology (right)

First, a numerical comparison between the proposed ring topology and the fat-tree network is performed, focusing on inventory and available capacity. From Table 6, it can be seen that for a network comprised of 8-port WDM switches, given the same number of supported ToRs, the Ring-of-Rings (RoR) topology has 45.8% less links and 50% less nodes, compared to the fat-tree. Switching from fat-tree to ring-based topology reduces the amount of needed equipment significantly, which is a clear benefit with respect to inventory management, failure localization, cabling, energy efficiency, etc. Furthermore, the bisection bandwidth (BW) of the RoR is more than 14 times lower.

Table 6: Inventory and capacity comparison

	# servers	# WDM nodes	# links	Bisection BW
FT	128	80	384	X
RoR	128	40	176	X/14.5

Next, we look at the performance of both networks under Scenario 1. Figure 18 shows the blocking ratio of the connection requests and the network resource utilization with varying resource availability in the DCN. The RoR topology outperforms the fat-tree network by 49% to 96% in terms of blocking and has 15% to 17% better utilization of its resources. Even though the fat-tree network has more links, and thus more paths between nodes, in a channel-switched scenario with millisecond long connections the available resources are not adequately utilized due to the concurrent nature of the resource reservation process. The RoR network on the other hand has less links, but the same amount of available capacity, which is better utilized to serve the requested connections. It is worth noting that the RoR network achieves lower blocking at more than 14 times lower bisection capacity.

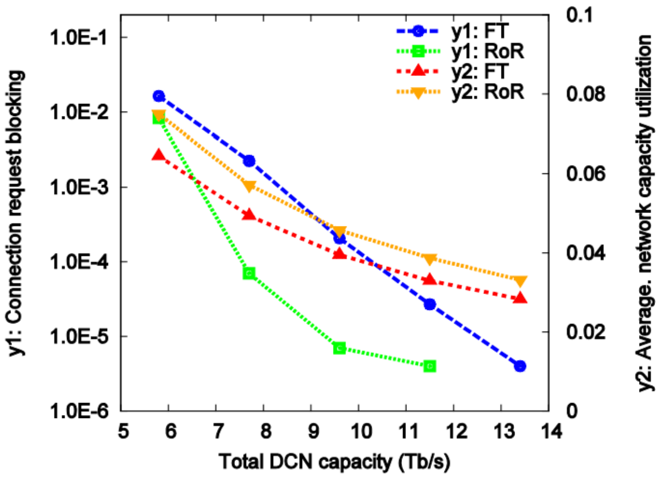


Figure 18: Performance evaluation under Scenario 1

Finally, we look at the performance of the networks under Scenario 2, where different traffic conditions are simulated. Figure 19 shows the connection blocking ratio and the average network capacity utilization as a function of local traffic ratio, i.e. when the ratio of within cluster/between-cluster traffic is changed, both for lightly loaded and for highly loaded networks. As before, the RoR DCN configuration outperforms the fat-tree architecture with respect to both performance measures (40% to 99% improvement in connection blocking and 2.6% to 17.6% improvement in resource utilization). It can be seen that at around 50% within-cluster/between cluster traffic distribution, the blocking is the lowest. At lower values, the core links of both DCN configurations get overloaded, whereas at higher values the links within a pod/cluster get overloaded, which results in higher blocking.

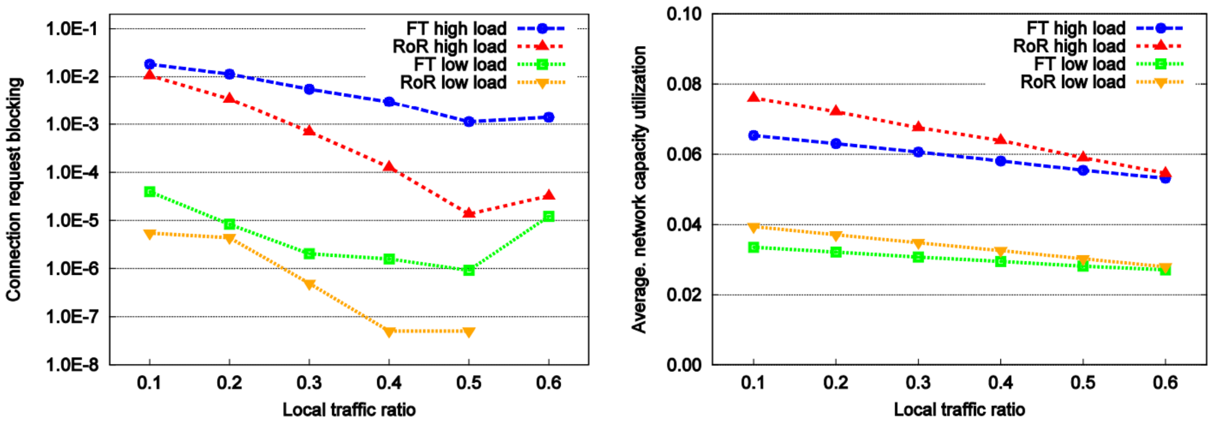


Figure 19: Connection request blocking (left) and network capacity (right) vs. within-cluster/between-cluster traffic ratio

5.1.3 Timeline, practical plans and goals for the Ring DCN

The demonstration of a ring based DCN will primarily target demonstration of physical connectivity and data plane performance. The goal for the ring-DCN demonstrator is to implement an optical DCN with SDN control and interfaces to standard Ethernet components (e.g. servers and ToRs). A best effort attempt will be done at adapting elements of the Virtual Application test case to run on the ring DCN. The performance metrics for the demonstrator are:

- Latency of Ethernet communication over the DCN
    - Min-max latency from bypassing an optical node (scalability in terms of latency)
- Throughput
- Realizable granularity of connection bandwidths

The ring DCN demonstrator will be a laboratory demonstrator mainly targeting dissemination to the research community. High visibility publications will be targeted to communicate the results to large DCN operators with research divisions following the scientific community.
Work on the ring DCN is already progressing well and a more comprehensive demonstration is planned for the first half of 2016.

5.2 Description of the cluster-based DCN and planned implementation

Figure 20 shows the first proposed architecture for intra data centre network (DCN) communication, and each DCN consists of clusters with tens/hundreds of racks networked together. All the clusters are connected together through a large-port-count fibre switch (LPFS)-based inter-cluster switch. Single-mode fibres (SMF) or multi-element fibres (MEF) can be used to connect all clusters to the inter-cluster switch. The inter-cluster switch configures the connection matrix among all clusters and provides adaptable link capacities. In addition, to realize inter-DCN application, the inter-cluster switch can send several signals from any top-of-racks (ToR) to the space-division multiplexing (SDM)-to-wavelength-division multiplexing (WDM) converter for metro/core network transmission, or can receive signals from the metro/core network and send these back to destination ToRs through the intra-cluster network.

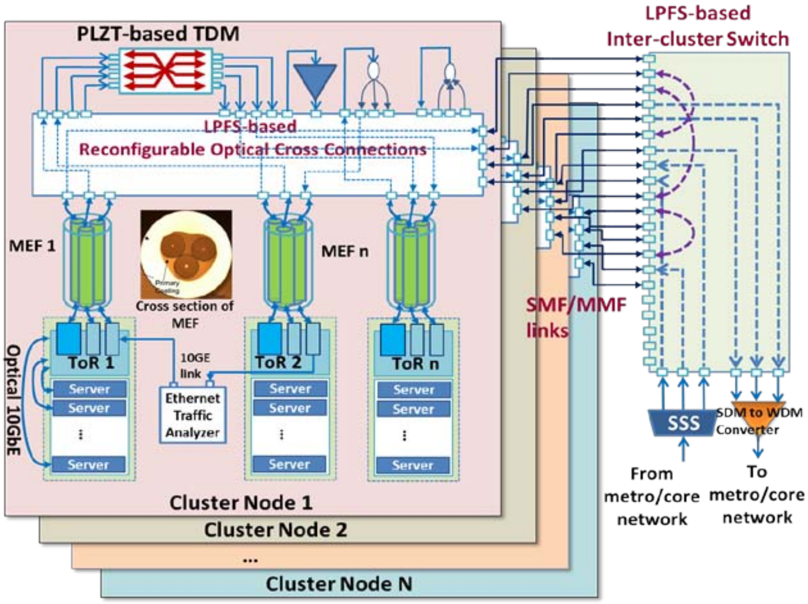


Figure 20: Intra data centre architecture based on SDM/TDM/OCS experimental setup of intra-DCN.

In each cluster, a centralized large-port-count fibre switch LPFS, as cluster switch, interconnects ToRs via MEF-based SDM links and other optical functional elements, i.e., PLZT-based time-division multiplexing (TDM) switch, and erbium-doped fibre amplifiers (EDFA). By re-configuring the connection matrix of LPFS in each cluster and inter-cluster switch, a single hop optical circuit switching (OCS) is achieved to serve the long-lived large-capacity flows both for intra- and inter-cluster communication. The OCS capacity can be reconfigured by configuring bandwidth variable transceivers (BVT) in ToRs. In addition to OCS, the re-configurability of LPFS introduces network function programmability in DCNs. According to traffic requests, interconnections between the connected components, such as optical splitters, optical couplers, and EDFAs, can be connected to OCS network, to obtain variable network functions, e.g., time aggregation, broadcasting and time de-multiplexing.

A fast PLZT-based TDM switch provides sub-wavelength switching with short reconfiguration time in order to serve short-lived low-capacity data flows for intra-cluster communication. The TDM link capacity can be varied from 100 Mb/s to 5 Gb/s by configuring the used time slot lengths. As TDM switch is synchronized with all ToRs in the cluster, more functions, such as time aggregation and time de-multiplexing, can be achieved without an optical buffer. The TDM switch is connected to the LPFS, enabling more network flexibility.

Traditional data centres have a relatively static computing infrastructure, normally a number of servers, and each with a set number of CPUs and fixed amount of memory. However, workloads and traffic patterns, in commercial as well as in High Performance Computing (HPC) data centres, show huge variation and a great degree of unpredictability. This erratic traffic inside data centres stems from the large differences that are observed between the requirements of the various applications or tasks that run in such environments. Thus, DC may safely supply adequate service during peak-demand, yet cannot fully utilize those same resources during non-peak conditions. Furthermore innovative data-intensive applications require sharing of compute and memory/storage resources among large amounts of servers, such as in modern highly parallelized computer architectures. Moving away from the traditional server-rack-cluster architecture, a need for novel arrangement and interconnection approaches has appeared.

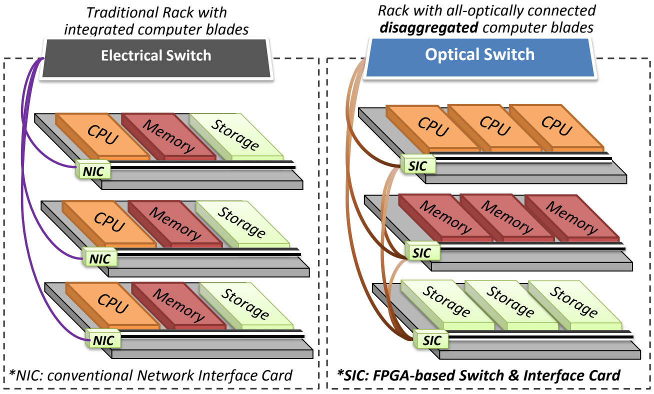

Traditional Rack with integrated computer blades
Electrical Switch
CPU
Memory
Storage
NIC
*NIC: conventional Network Interface Card
Rack with all-optically connected **disaggregated** computer blades
Optical Switch
CPU
Memory
Storage
SIC
*SIC: FPGA-based Switch & Interface Card

Figure 21: Conventional integrated servers rack (left) & all-optically connected disaggregated computer blades rack approach with FPGA-based SIC (right)

A promising approach to satisfy the above conditions, while adding modularity and upgradability, is the physical disaggregation of the DC resources (CPU, DRAM, storage) which will then be joined by an all-optical low-power consumption high-capacity interconnect with minimal chip-to-chip delays (as shown in Figure 21). It is quite obvious that ultra-low-latency connectivity and protocols need to be established, especially between processing and remote memory chips/blades, so that maximum performance is maintained. Modular and flat architectures consisting of optical and electrical technologies are able to support this notion, by incorporating dimension reconfigurable capacity-agnostic optical switches, and with fully-programmable, integrated switch and interface cards (SIC) within and among the disaggregated blades. This disaggregation of server resources can enable fine-grained resource provisioning consequently boosting the overall system performance.

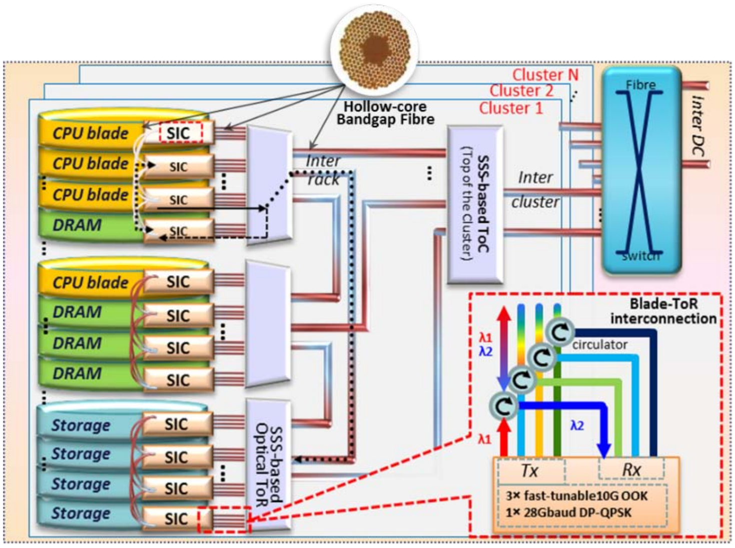


Figure 22: All-Optical Programmable Disaggregated Data Centre Interconnect utilizing Hollow-Core Bandgap Fibre

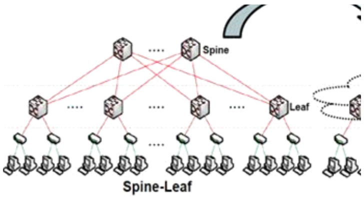
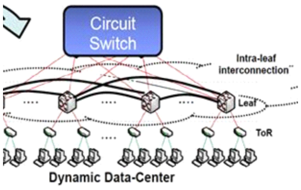
Figure 22 shows the proposed architecture with the bidirectional intra-cluster interconnection scheme of various blades/racks containing disaggregated resources for increased system modularity and performance. Hollowcore photonic bandgap fibre (HC-PBGF) links offer 30% propagation delay reduction and dimension-programmable single device 20-port NxM spectrum selective switches (SSS) were used for ToR and top of cluster (ToC) switches. The optical SSS-based ToR switches, which are bandwidth-to-port selection-flexible and able to dynamically reconfigure their I/O dimensions (e.g. 4×16, 8×12, 10×10), groom, switch and balance traffic within and between racks. This unique feature enables the architecture to adapt to different networking scenarios with diverse requirements on capacity, latency and connectivity. The identical SSS-based ToR handles and re-balances intra/inter-cluster communication and aggregates traffic towards a high-port count fibre switch, adding scalability to the network. The blade-to-ToR interface is implemented with 4 channels and circulators to exploit the bi-directionality of the SSS, thus saving resources.

By utilizing 3×fast tuneable 10 Gb/s OOK channels for either lowlatency WDM or WDM/TDM and a 28 Gbaud DP-QPSK channel on each SIC, wide (C-band), flexible-bandwidth connectivity with various granularities is achieved. Traffic is switched either via the cut through optically-connected electrically-switched back-end or via the programmable fast-tuneable dual-mode (cut-through WDM/WDM-TDM) to reach remote destinations. Apart from interfacing, SIC can also hitless switch traffic from the back-end towards the front-end and vice versa, allowing multi-hop communication to assist congested links. Furthermore, SIC is able to flexibly aggregate TDM traffic based on network requirements (QoS, latency, connectivity) by dynamically programming the size/number of slots per frame.

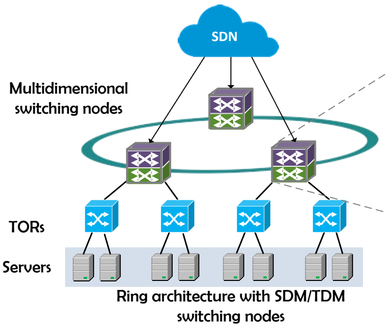
6 Conclusion

The planned demonstrator activities in COSIGN are summarized in Table 7.

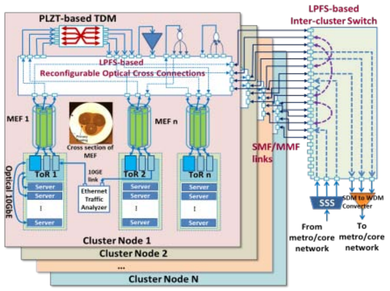
Table 7: Summary of COSIGN demonstrator efforts

<table>
<tr><th></th><th>V. App.</th><th>VDC provisioning</th><th>VDC VM migration</th></tr>
<tr><td><b>Short term:</b>
 </td><td>X
Physical demonstrator in WP5</td><td>-
Only orchestration algorithms in WP4</td><td>-</td></tr>
<tr><td><b>Medium term:</b>


Circuit Switch
Intra-leaf interconnection
Leaf
ToR
Dynamic Data-Center

</td><td>X
Physical demonstrator in WP5</td><td>X
Physical demonstrator in WP5</td><td>X
Physical demonstrator in WP5</td></tr>
<tr><td rowspan="2"><b>Long term:</b>
<b>All-optical Ring DCN</b>


SDN
Multidimensional switching nodes
TORs
Servers
Ring architecture with SDM/TDM switching nodes

</td><td>(X)
Best effort to adapt V. App test case to circuit-based DCN</td><td>-</td><td>-</td></tr>
<tr><td colspan="3">Main focus on demonstrating optical connectivity, flexibility and large capacity.</td></tr>
<tr><td><b>Long term:</b>
<b>All-optical clustered DCN</b>


PLZT-based TDM
LPFS-based Inter-cluster Switch
LPFS-based Reconfigurable Optical Cross Connections
MEF 1
MEF n
Cross section of MEF
SMF/MMF links
Optical 10GbE
ToR 1
ToR 2
ToR n
10GE link
Ethernet Traffic Analyzer
Server
SSS
SDM to WDM Converter
From metro/core network
To metro/core network
Cluster Node 1
Cluster Node 2
...
Cluster Node N

</td><td colspan="3">Aim to demonstrate all three</td></tr>
</table>

The three defined demonstrator scenarios will be implemented and tested according to the scope of the demonstrator. The short term demonstrator will be tested partly as a reference scenario for the medium term demonstrator. The medium term demonstrator will be a key demonstrator in COSIGN featuring support for all the defined use cases implementing selected concrete test cases presented in this document. The long term demonstrators will focus on the realization of an all-optical DCN and quantify key benefits of the extensive introduction of optics in the DCN.

The medium term demonstrator will be the key industrial demonstrator which will also be hosted by Interoute. The long-term demonstrators will target a broad academic audience and commercial audience.