Intra- Datacenter Challenges; System Perspective

Elad Mentovich

Mellanox TECHNOLOGIES
Connect. Accelerate. Outperform.

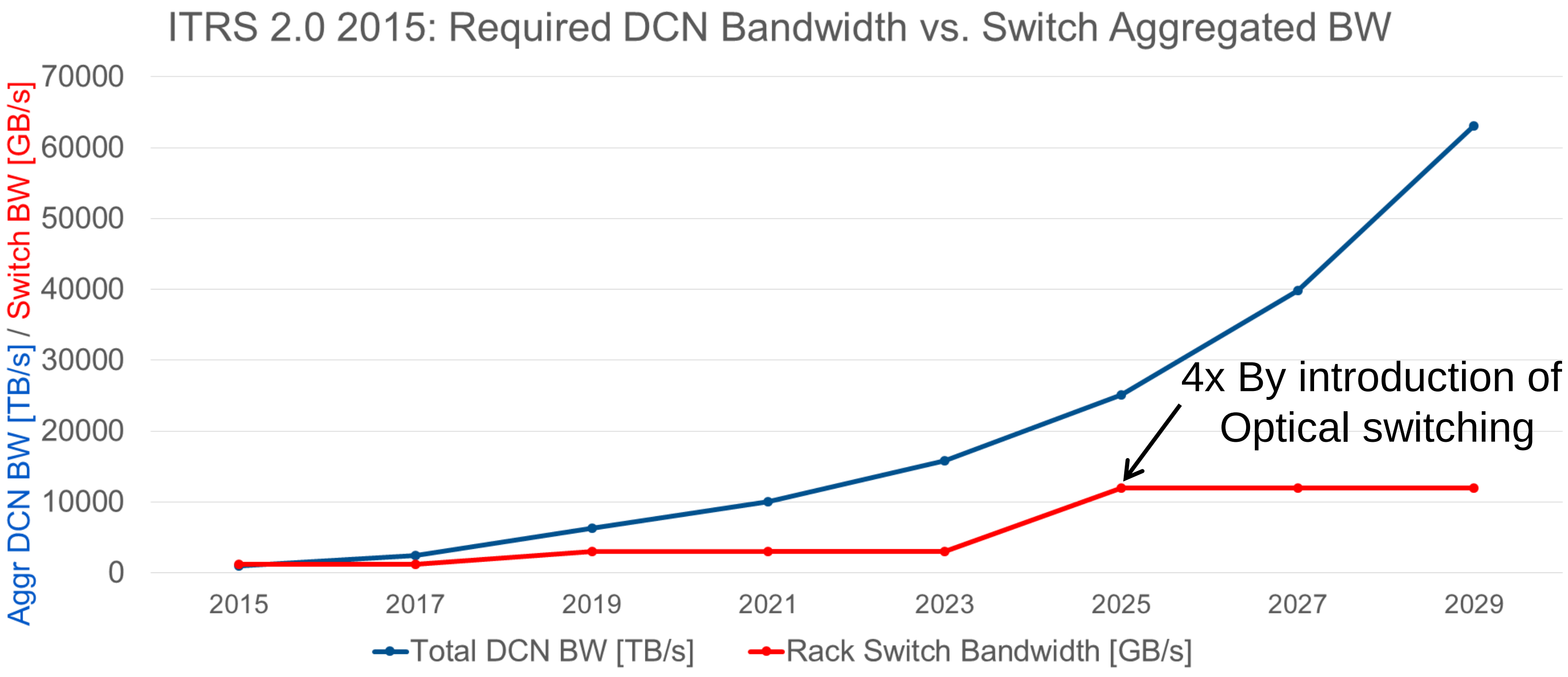
ECOC2017 workshop

Datacenter Network challenges

Mellanox TECHNOLOGIES

Challenge 1: Bandwidth

- datacenter traffic demand grows exponentially
- electrical switches do not catch up on aggregate bandwidth



feature	2012	2016	2020
Peak performance	10 PF	100 PF	1000 PF
(bidi) bandwidth	1 PB/s	20 PB/s	400 PB/s
overall power consumption	5 MW	10 mW	20 MW
network power consumption	0.5 MW	2 MW	8 MW

Challenge 2: Energy consumption

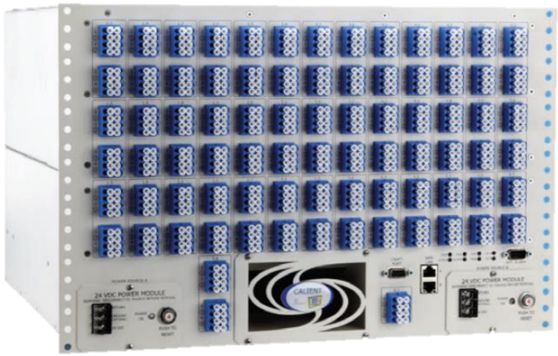
- DCN total energy efficiency has to drop from a few mW/Gb/s to less than 1 mW/Gb/s*

*P. Pepeljugoski et al., “Low Power and High Density Optical Interconnects for Future Supercomputers,” in proc. OFC 2010.

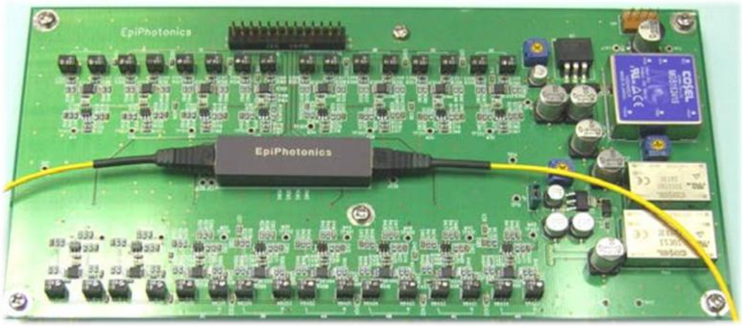
Optical switching to the rescue

The good

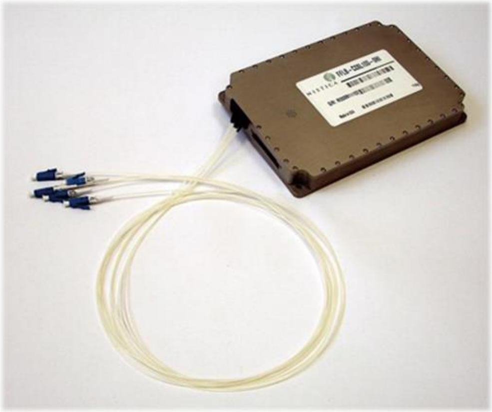
- transparent
  - *"unlimited" bandwidth*
  - *future proof*
- low power*
  - *16 ×16 MEMS module: 150 mW*
  - *36 port state-of-the-art switch: 136 W*
- low latency*
- potential for large scale switches
- compatible with photonic integration
- compatible with emerging trends
  - *single-mode optics*
  - *software-defined-networking*



320 port optical switch



2x2 switch, ns-speed



1x4 WSS, 50 GHz grid

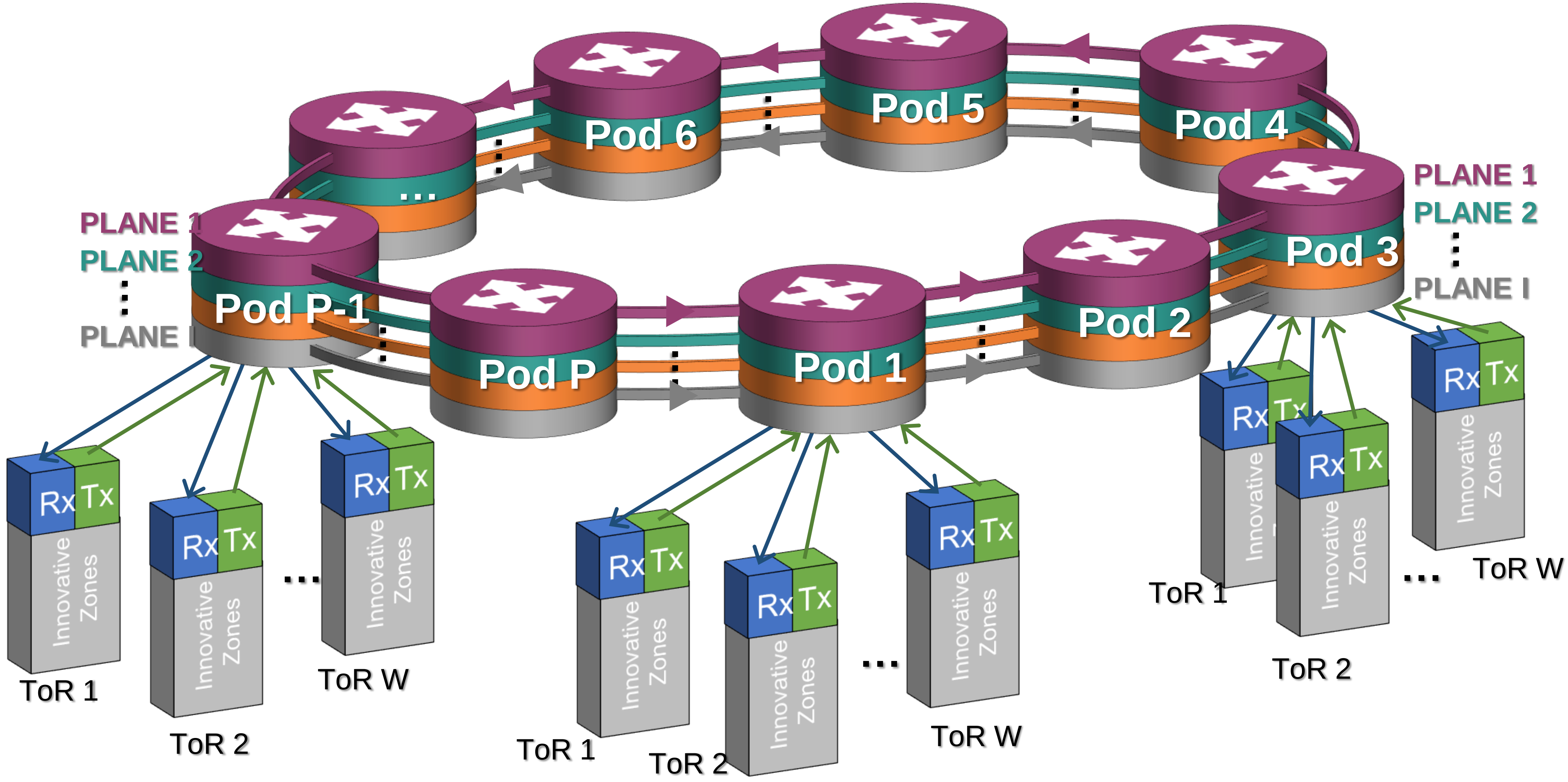
Optical switching challenges

- no one-to-one association of optical & electronic switches
  - ***no buffering***!
  - *no processing, no functionality whatsoever*
- **reconfiguration time**: speed vs. port count tradeoff
  - *fast optical switches (~ns) typically small port count. scalability…*
  - *large optical switch technologies typically slow (~ms)*
- **cost**, supply chain
  - *currently tailored to telecom applications*

How to introduce optical switching? Need to revisit the entire DCN approach

- *network* ***architecture***
- *network* ***management***
- *photonic* ***components***

The NEPHELE Data Center Network

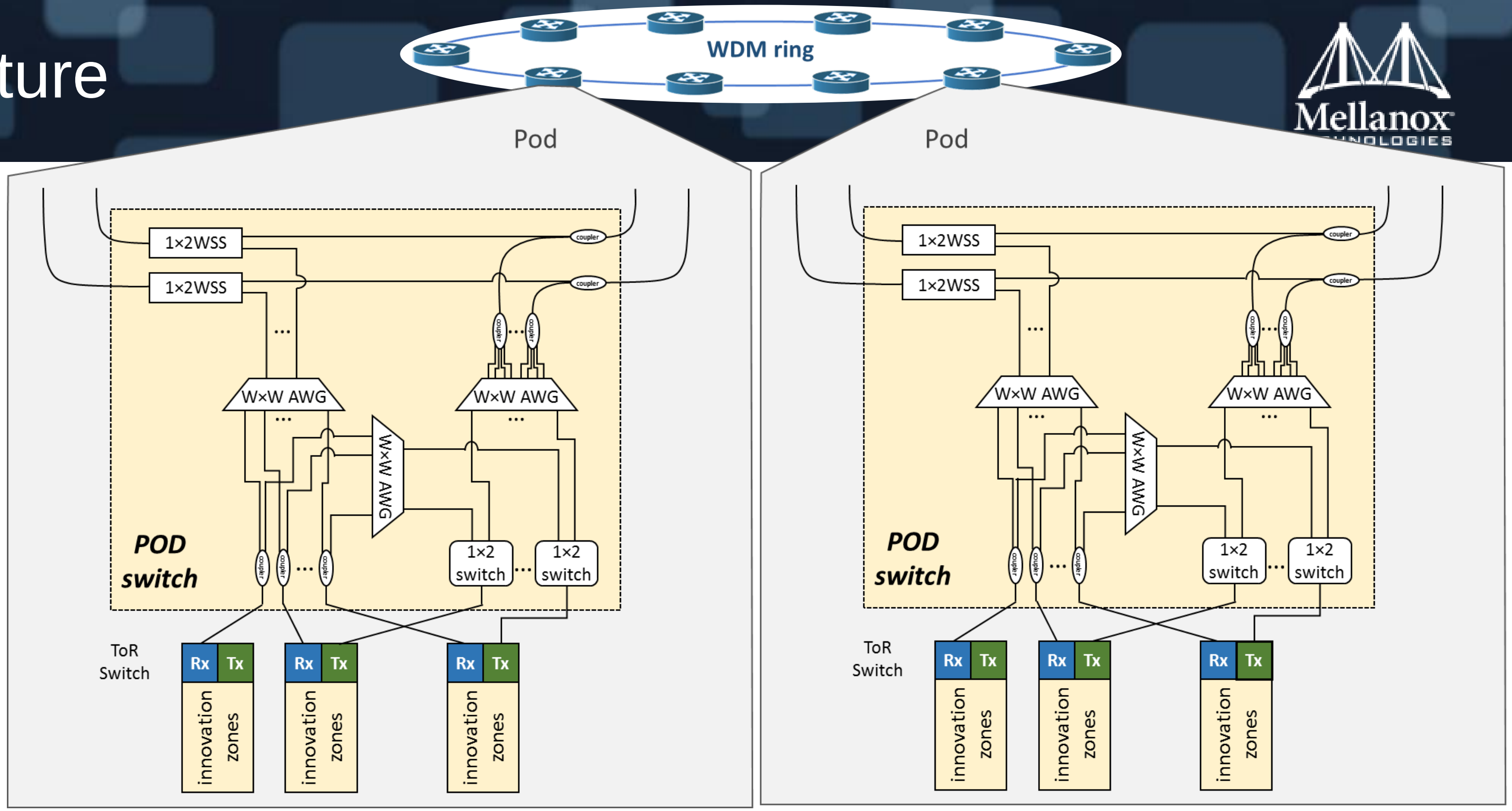


- relies on **COTS** components
- **Scalable** to >32,000 hosts
- #switches **linear** with #hosts
- TDMA and WDM

*P. Bakopoulos et. al.,"NEPHELE: an end-to-end scalable and dynamically reconfigurable optical architecture for application-aware SDN cloud datacenters", *IEEE communications Magazine,* accepted for publication.

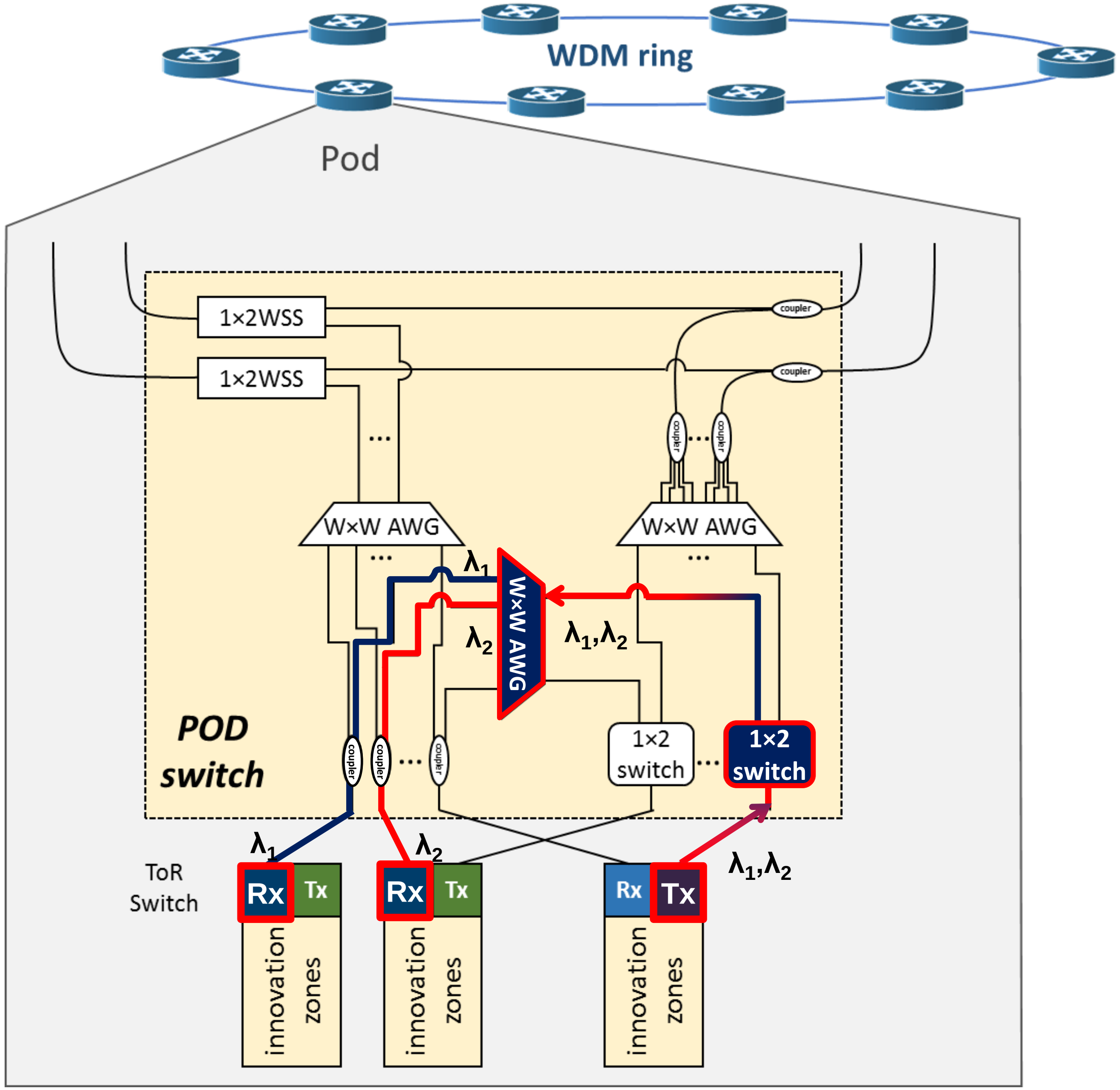
The NEPHELE Architecture

- **ToR (Rack):** up to 20 sub system ports (innovation zones)
- **POD:** self-contained small DC up to 80 racks (1,600 ports)
- **Data Center:** 20 PODs (32,000 ports)

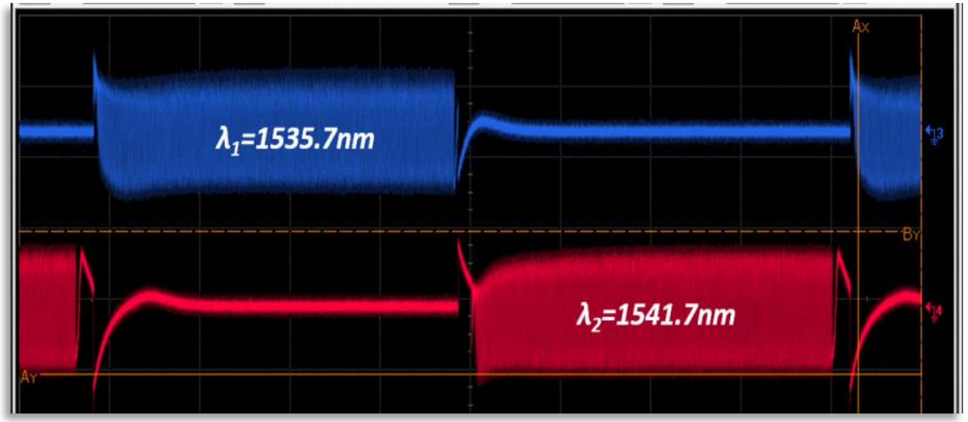


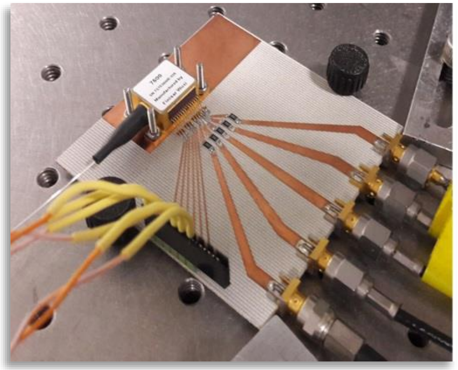
Parameter	Meaning	Typical value
Z	Number of innovation zones per ToR switch	4
S	Number of innovation zones' ports per ToR switch	20
W	Number of racks and ToRs per pod; also number of wavelengths in the system	80
R	Number of fiber rings per optical plane	20
P	Number of pods	20
I	Number of NEPHELE optical planes	20

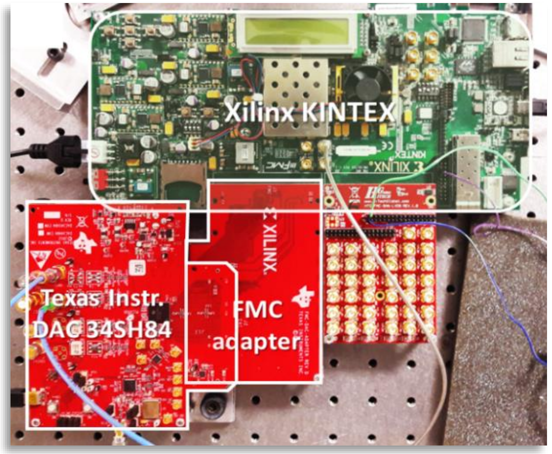
The NEPHELE Architecture



- **wavelength switching** in the pod: tunable Tx
  - *< 22 ns switching time (200 μs packets)*
  - *FPGA controlled, 80 λ LUT*

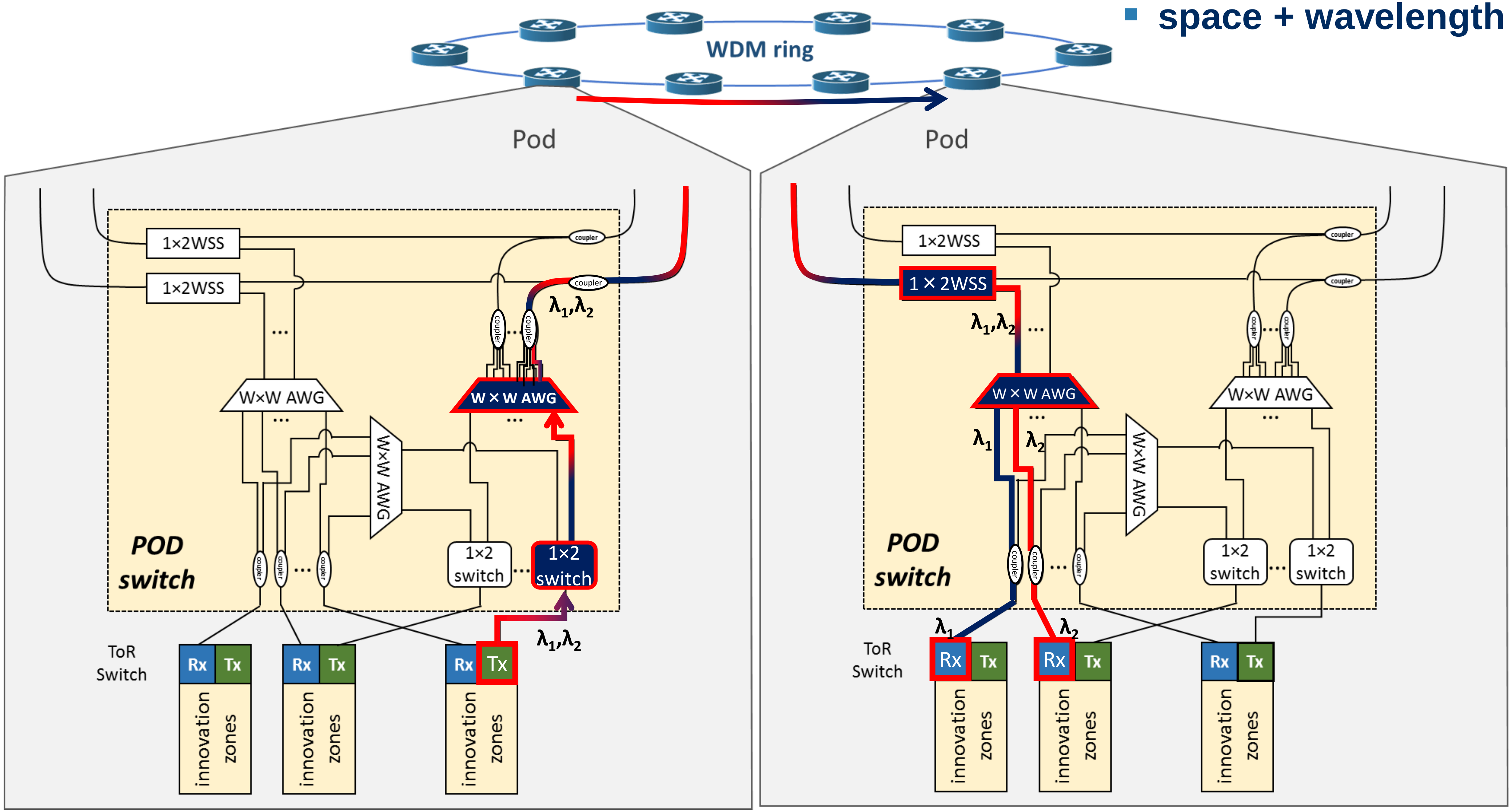






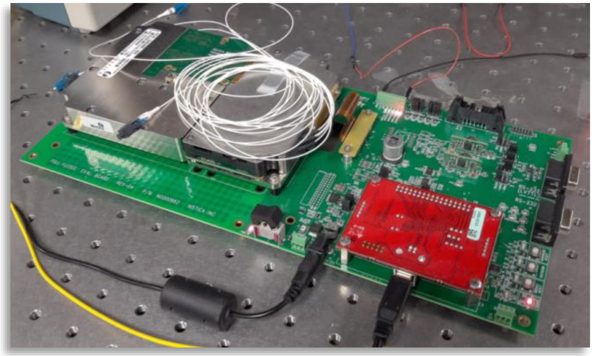
The NEPHELE Architecture

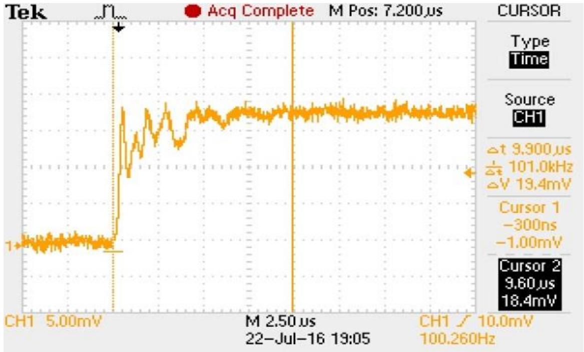
- **space + wavelength** switching in the ring



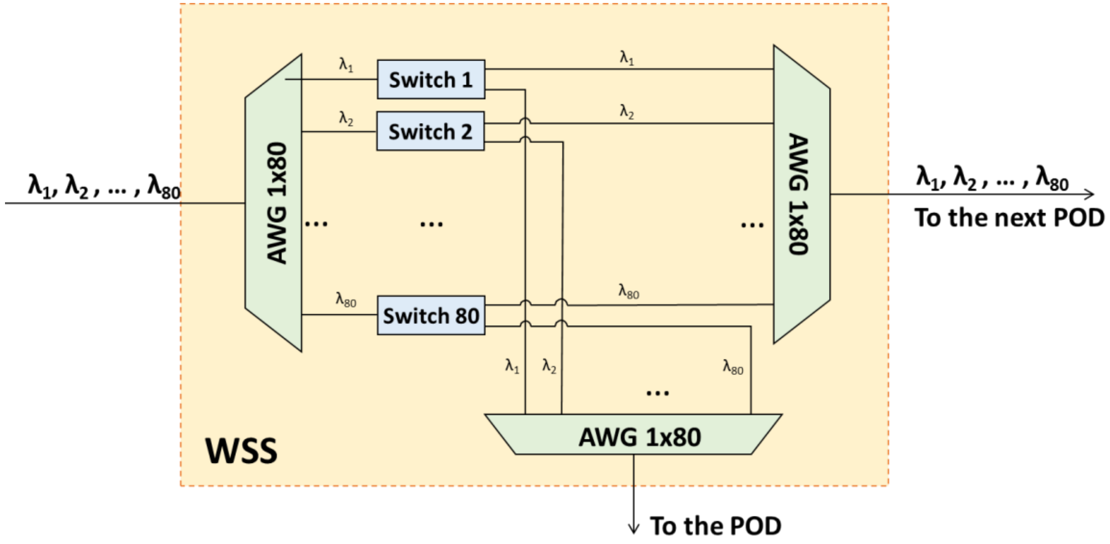
The NEPHELE Architecture

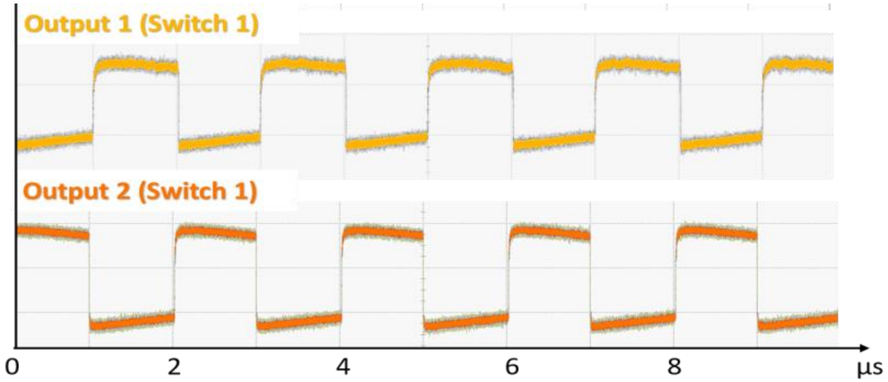
- **space + wavelength** switching in the ring: WSS





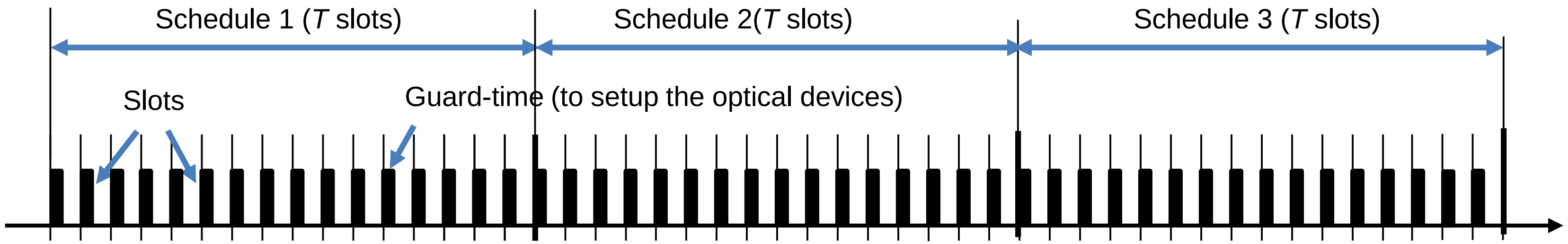
- *DLP® WSS, 10 µs switching time*





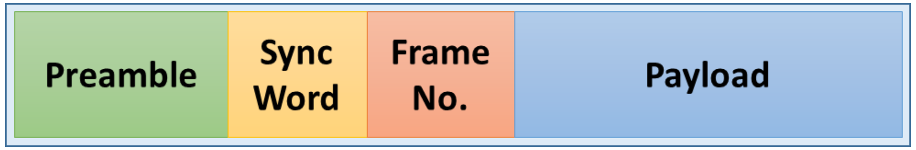
- *1×2 switch-based WSS, 10 ns switching time*

The NEPHELE Architecture – **TDMA** operation



- Each slot contains exact setting for **all** the optical devices in the entire network
- Same schedule may be used many times
- Length of schedule may change based on required traffic pattern
- Length of slot may be reduced if we don’t have data to send, too

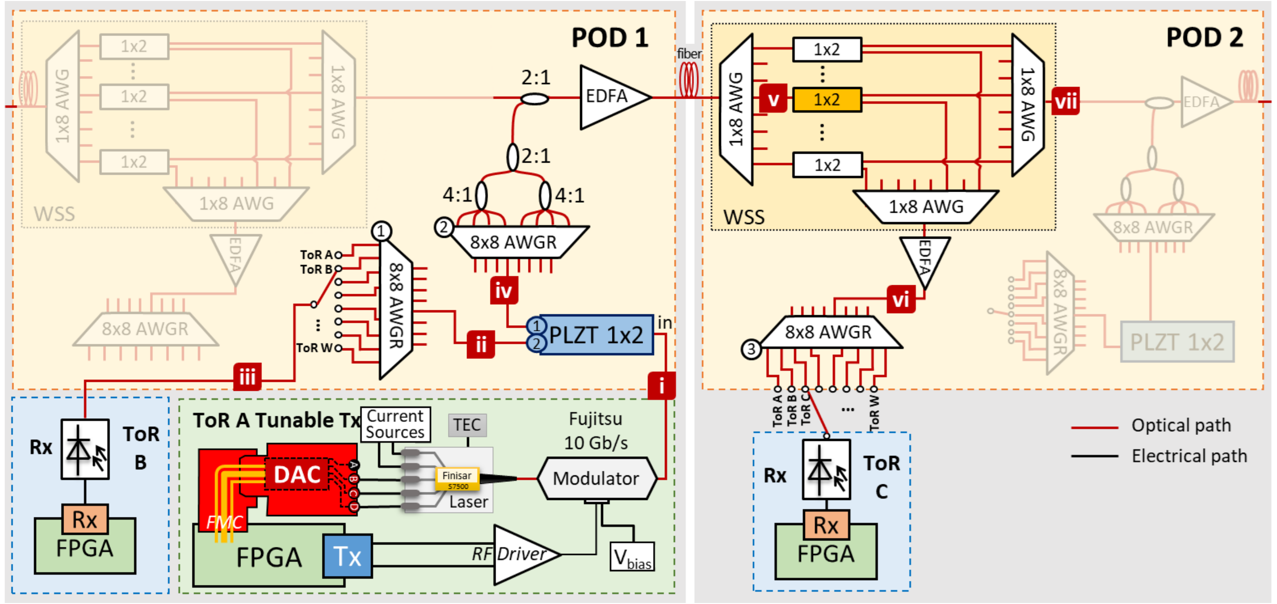
Parameter	Typical value	comment
packet length	200 µs	
guard time	10 µs	dictated by optical component with slowest reconfiguration time
schedule length	16.8 ms	worst case, all to all: 80 × (*packet* + *guard time*)

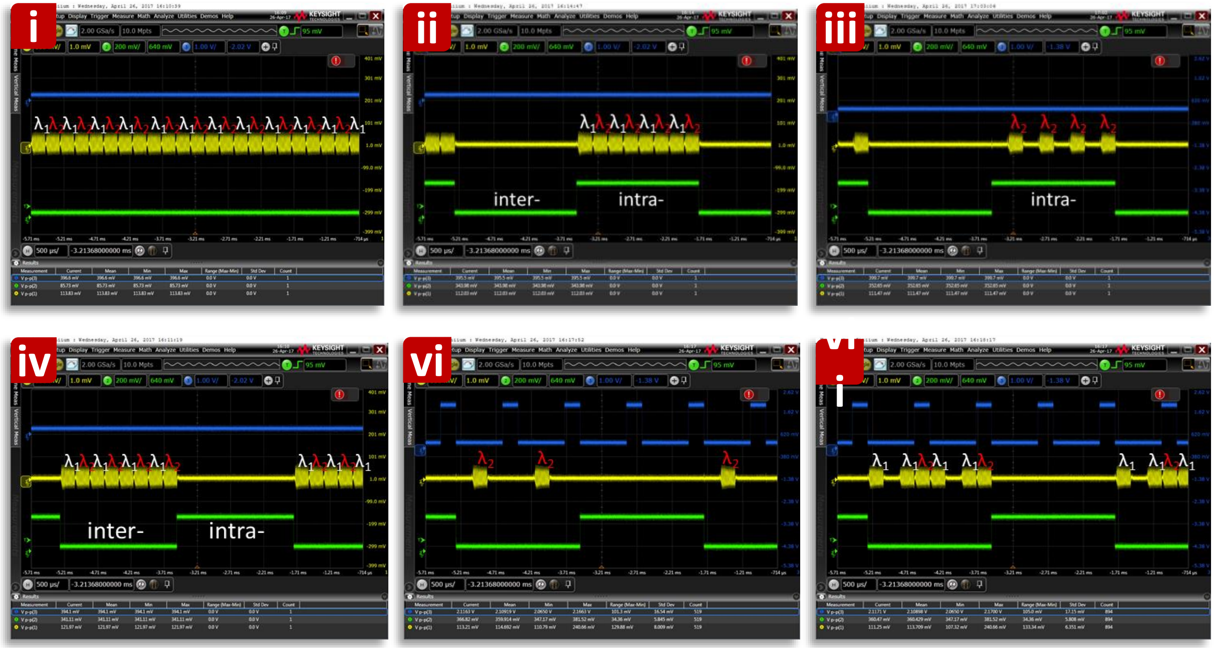


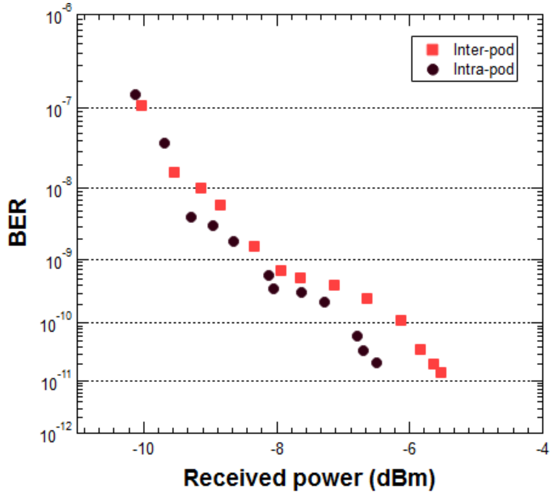
- **Transmitter**
  - ***XILINX Kintex KC707***
  - ***Generation of data packets***
- **Receiver**
  - ***XILINX Virtex VC707***
  - ***Packet loss and BER calculation***

Experimental results (I)

▪ Intra-pod & inter-pod communication







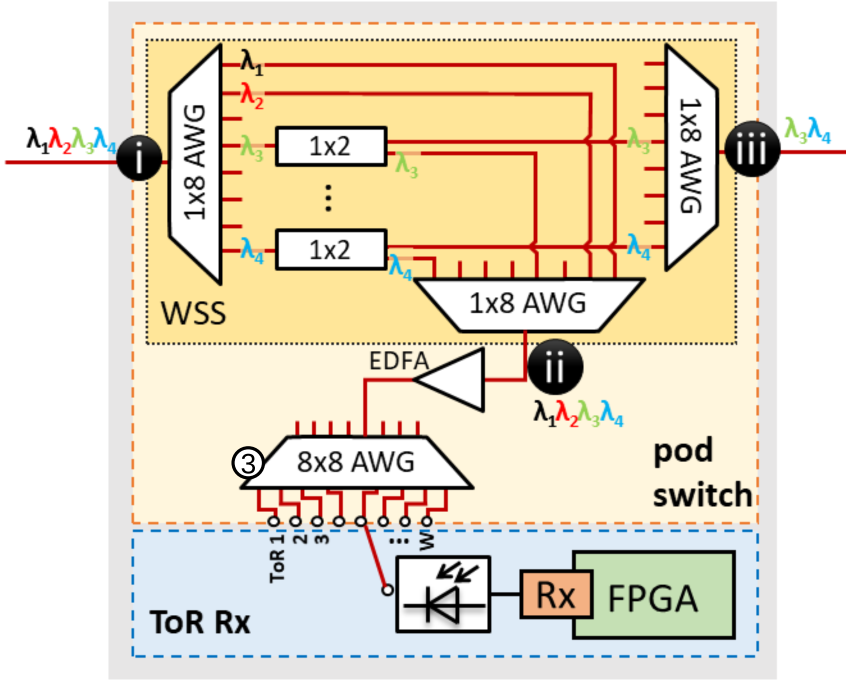
BER better than 3.10^{-13}

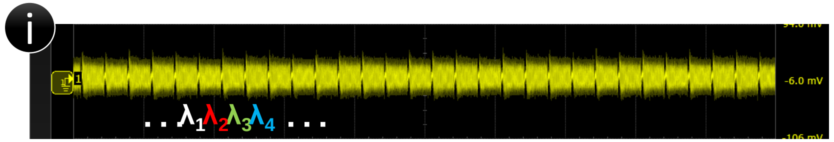
8 packets are switched towards a different POD and 8 packets remain within POD 1 alternatingly, via outputs 1 and 2 of the PLZT switch

λ_1=1546.91 nm
λ_2=1551.72 nm

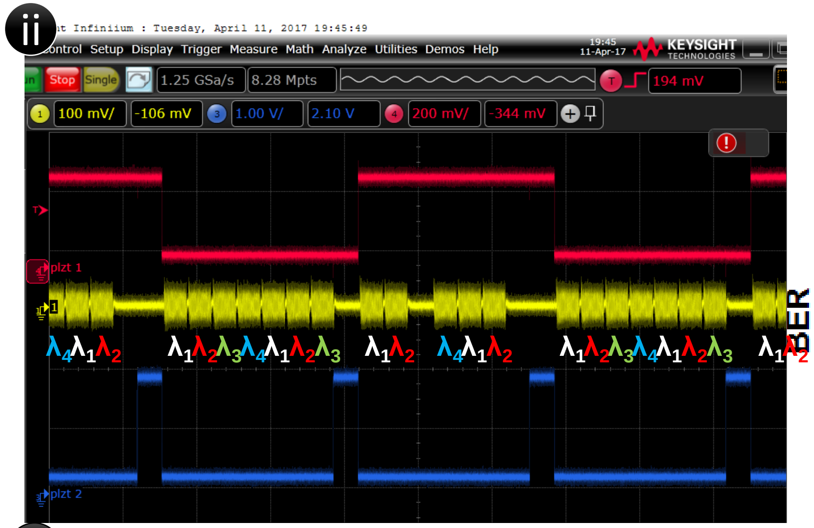
Experimental results (II)

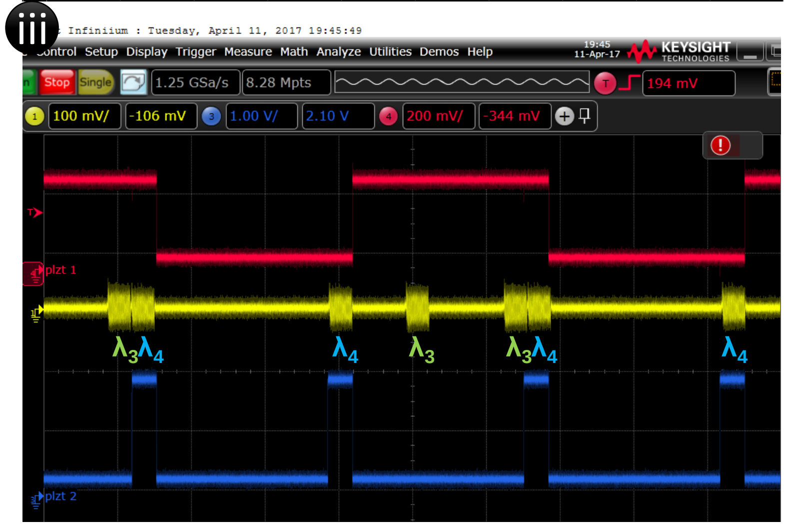
- **scaling of WSS functionality**

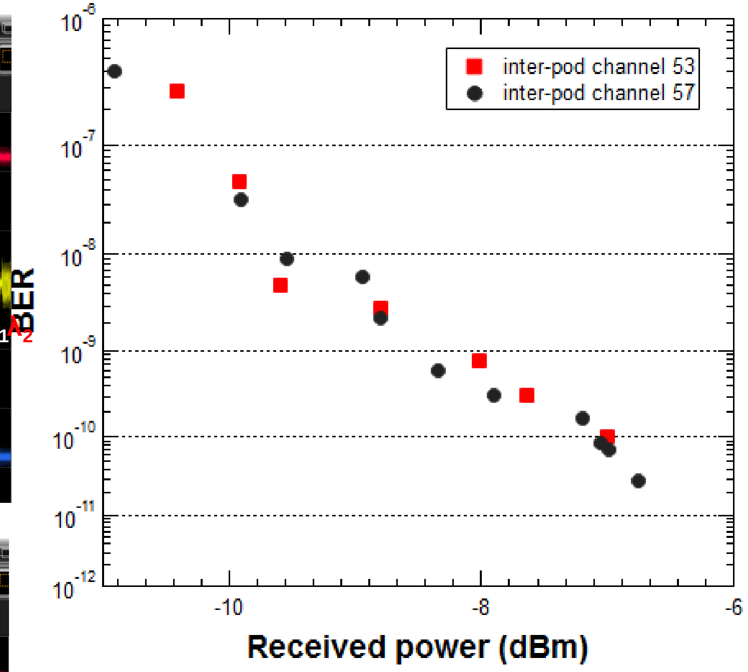




λ_1=1550.116 nm
λ_2=1548.515 nm
λ_3=1550.918 nm (ch.53)
λ_4=1552.11524 nm (ch.57)



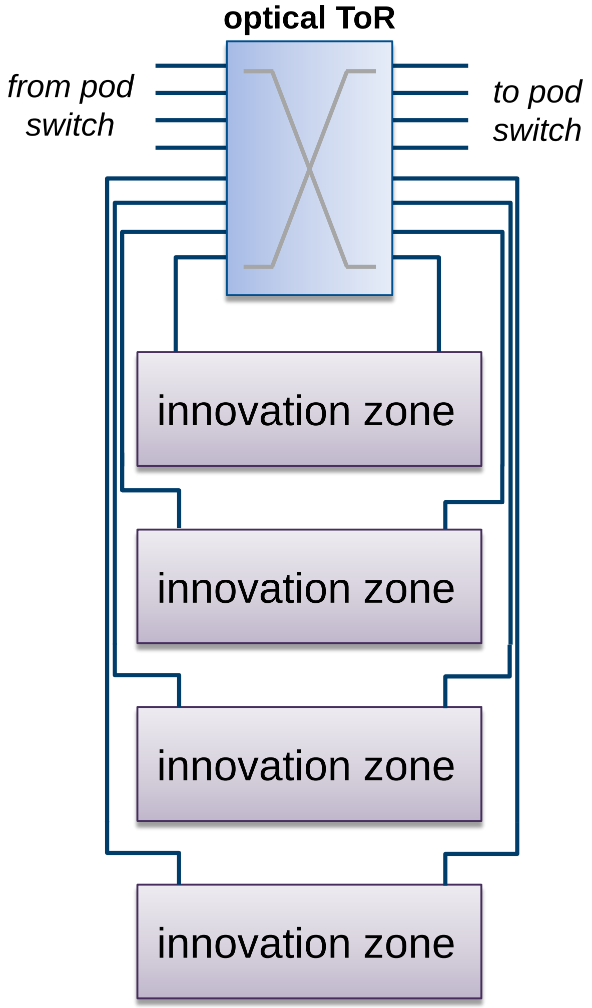


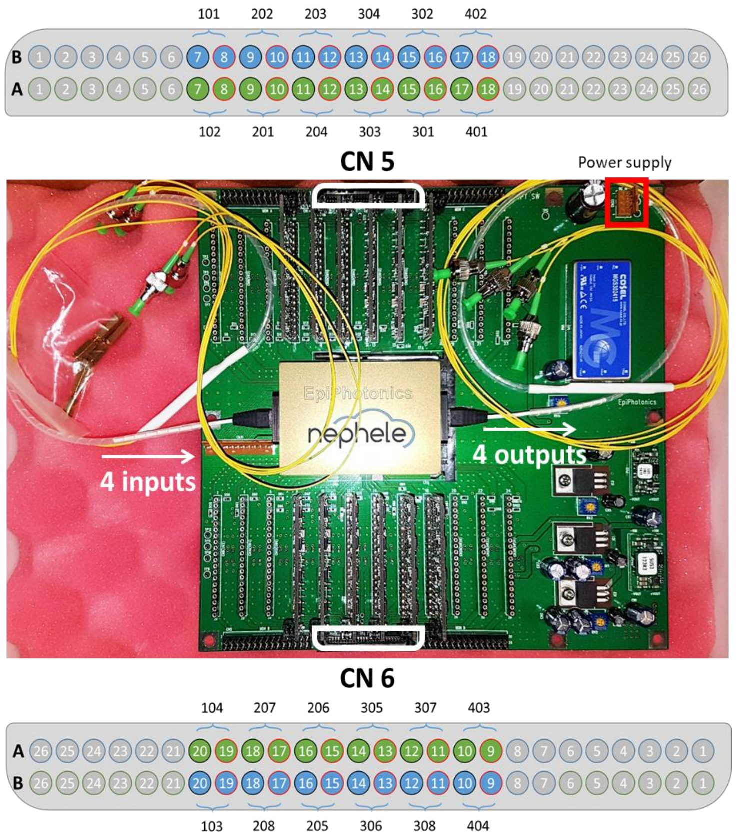


- *for optical power higher than -6.8 dBm no errors were observed (BER better than 3.10^{-13})*

Bringing optical switching to the server

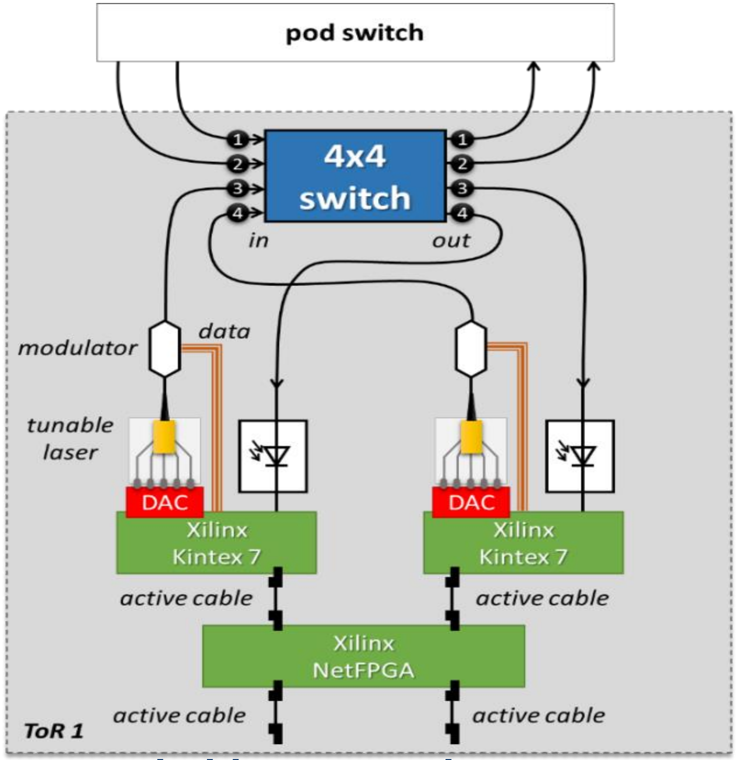
- **Modified version of NEPHELE ToR to accommodate all-optical traffic**





4×4 PLZT switch

- *~33 ns rise and fall time*
- *14.6 dB insertion loss*
- *26 dB crosstalk*

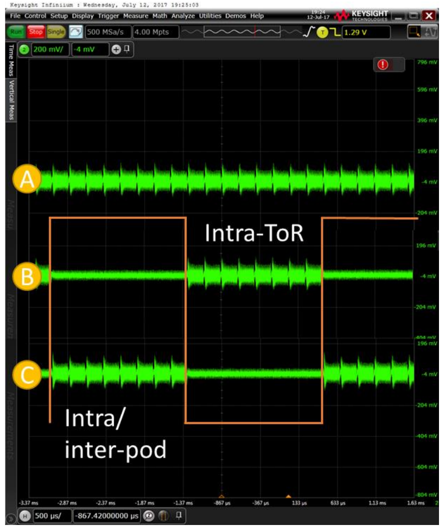


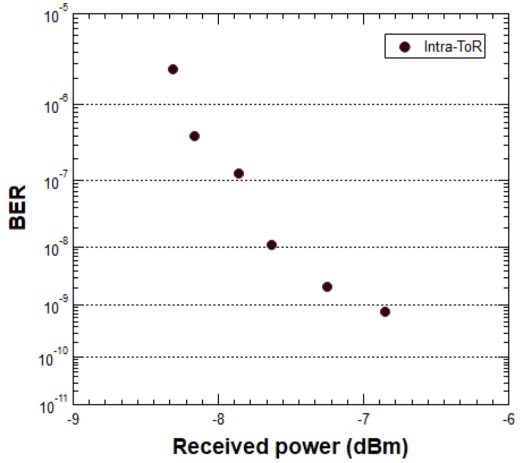
Switching scenarios

- *Two NEPHELE servers communicating via optical ToR*
- *inter-rack*
- *intra-pod*
- *inter-pod*

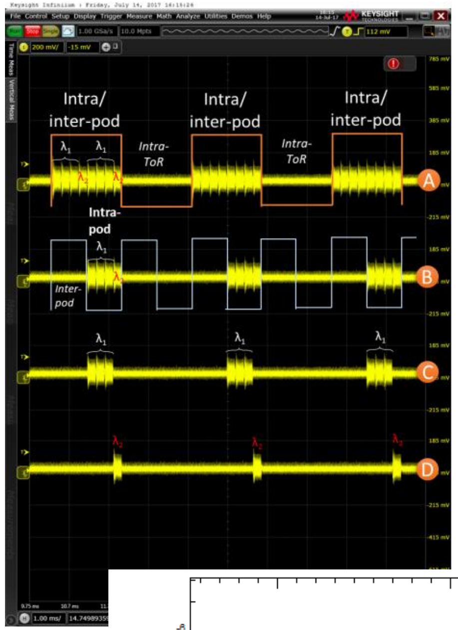
Experimental results (III)

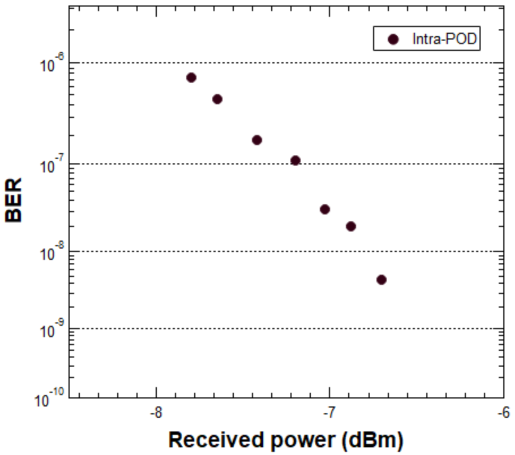
- **Inter-ToR traffic**



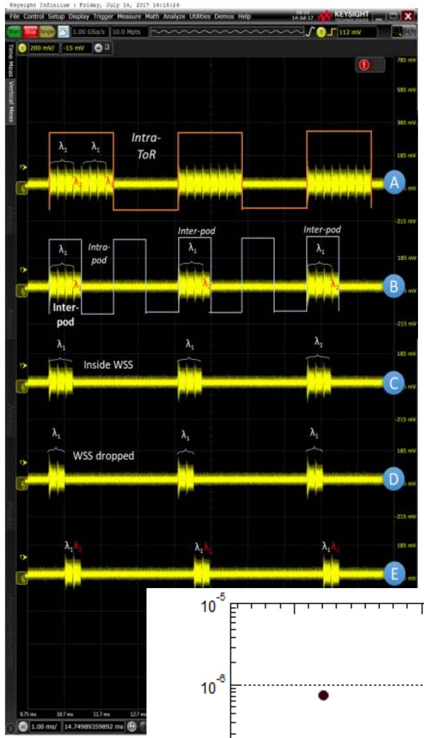


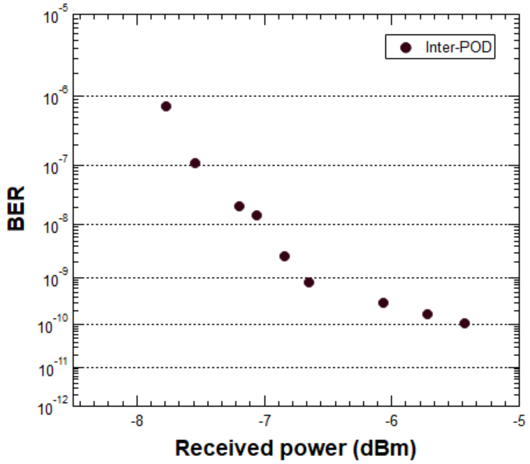
- **Intra-pod traffic**





- **Inter-pod traffic**





Controlling the optical DCN

Central vs. distributed control*

- **Distributed:** Each packet switch avoids collisions on its own by utilizing over provisioned resources
- **Central:** Allocate each of the flows with an orthogonal light circuit (end-to-end connection)

Distributed Control relies on the switches to resolve contention

- Mostly by using Tunable Wavelength Converters (TWC)
  - *TWCs are expensive, power consuming and a large number is required*
  - *Effect of cascading TWC-based switches on performance TBD*

e.g. DOS, IRIS, NTT

Central Control relies on a central controller (aka SDN)

- Simpler and possibly more feasible
- Suffers from the load of the central controller

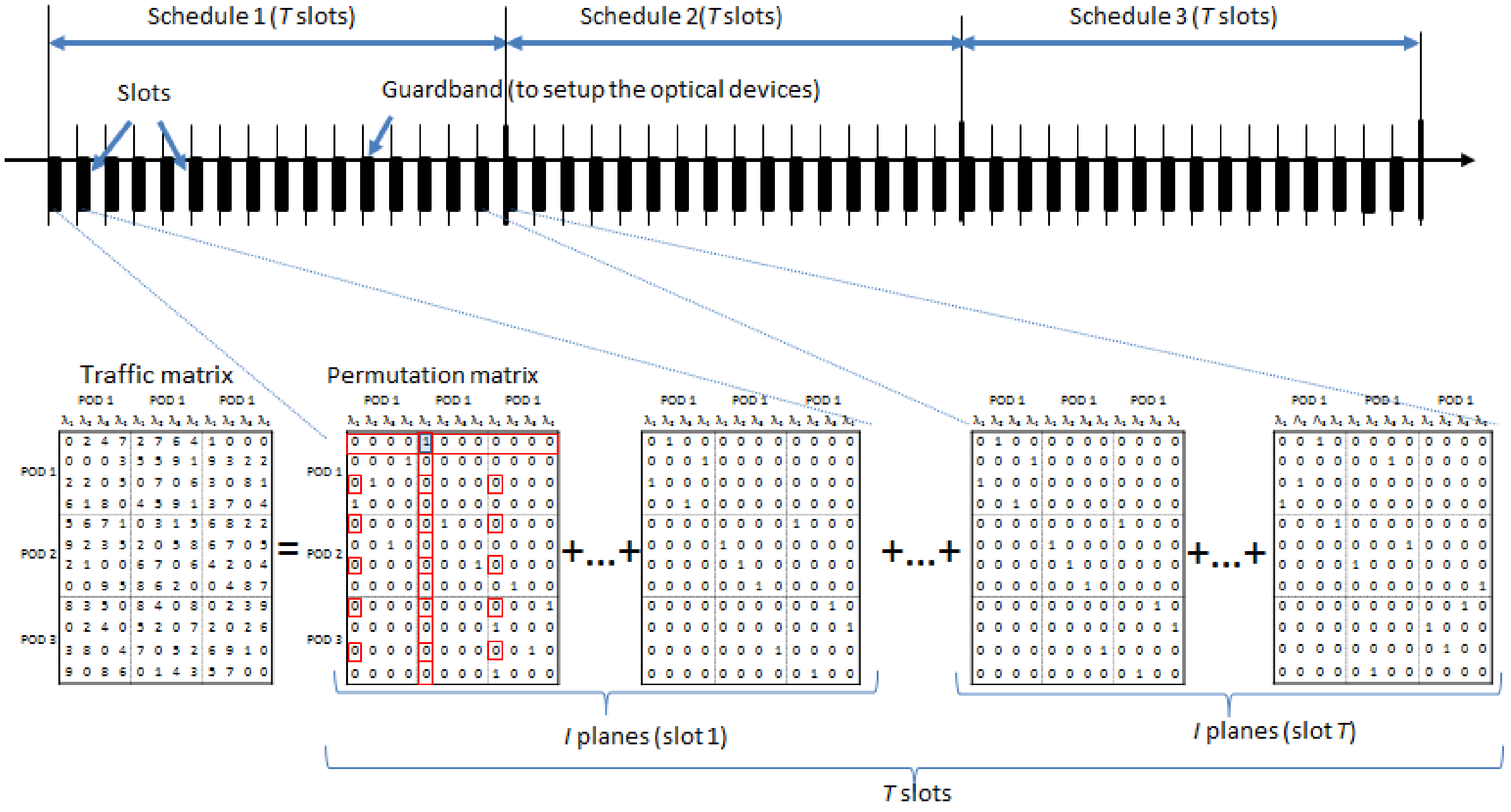
*e.g. Mordia, Lightness, Dublin City University, RotorNet***

*E. Zahavi, "Optical Data Centers," in proc. *HIPINEB Summer School* 2017.
**Central controller is avoided by applying a fixed set of permutations

Controlling the NEPHELE DCN

TDMA scheduler*

- permutation matrices represent communications over a specific plane and timeslot
- traffic matrix is the sum of permutation matrices
- each permutation matrix represents communications over a specific plane and timeslot
- The traffic matrix is periodically generated at the controller



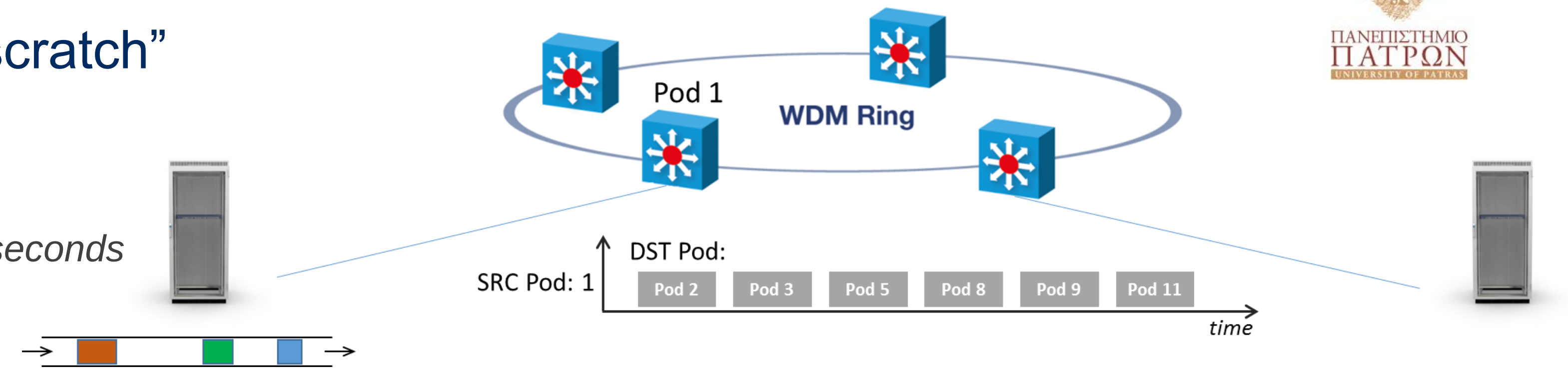
*K. Christodoulopoulos et. al., “Bandwidth Allocation in the NEPHELE Hybrid Optical Interconnect” *ICTON* 2016

*M. Varvarigos, 5th International Symposium for Optical Interconnect in Data Centres, 4th session.

Offline and Incremental algorithms developed

Offline: calculate the schedule "from scratch"

- Developed offline algorithms
  - *Optimal, Maximum Remaining Sum, greedy*
  - *Good performance but high run time – order of seconds*



Incremental: take into account the previous schedule

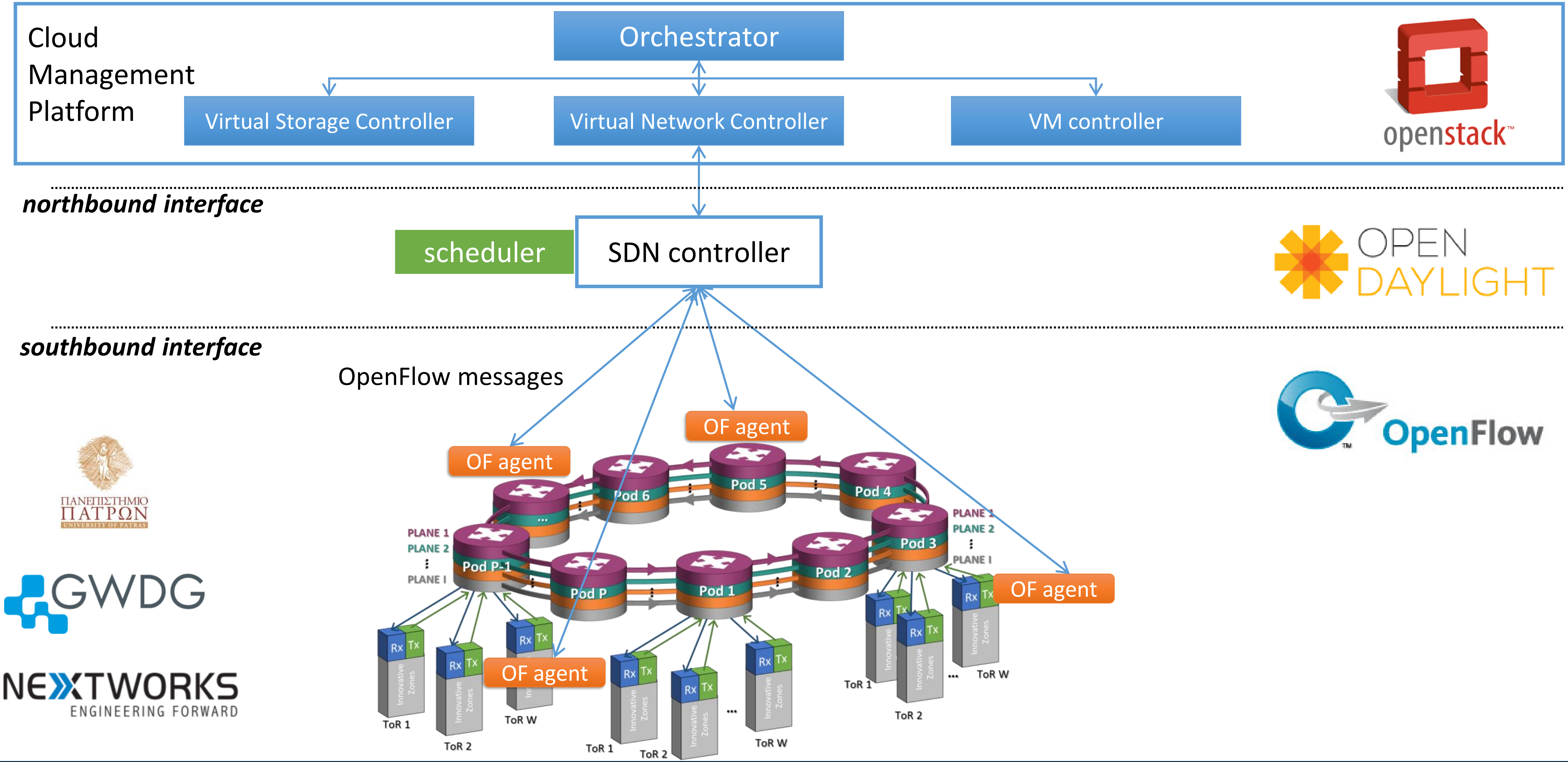
- Motivation: traffic from period to period does not change substantially – observed in real DCs*
  - *Results even for 10% change in two consecutive periods are quite promising*
- Developed incremental algorithms
  - *Optimal, randomized, greedy*

➢ The greedy has very low execution time (~0.2 sec) and very good performance

➢ A parallel implementation of the greedy algorithm in an FPGA is under development, early results are quite promising

*A. Roy, H. Zeng, J. Bagga, G. Porter, A. C. Snoeren,"Inside the Social Network's (Datacenter) Network", *Sigcomm* 2015

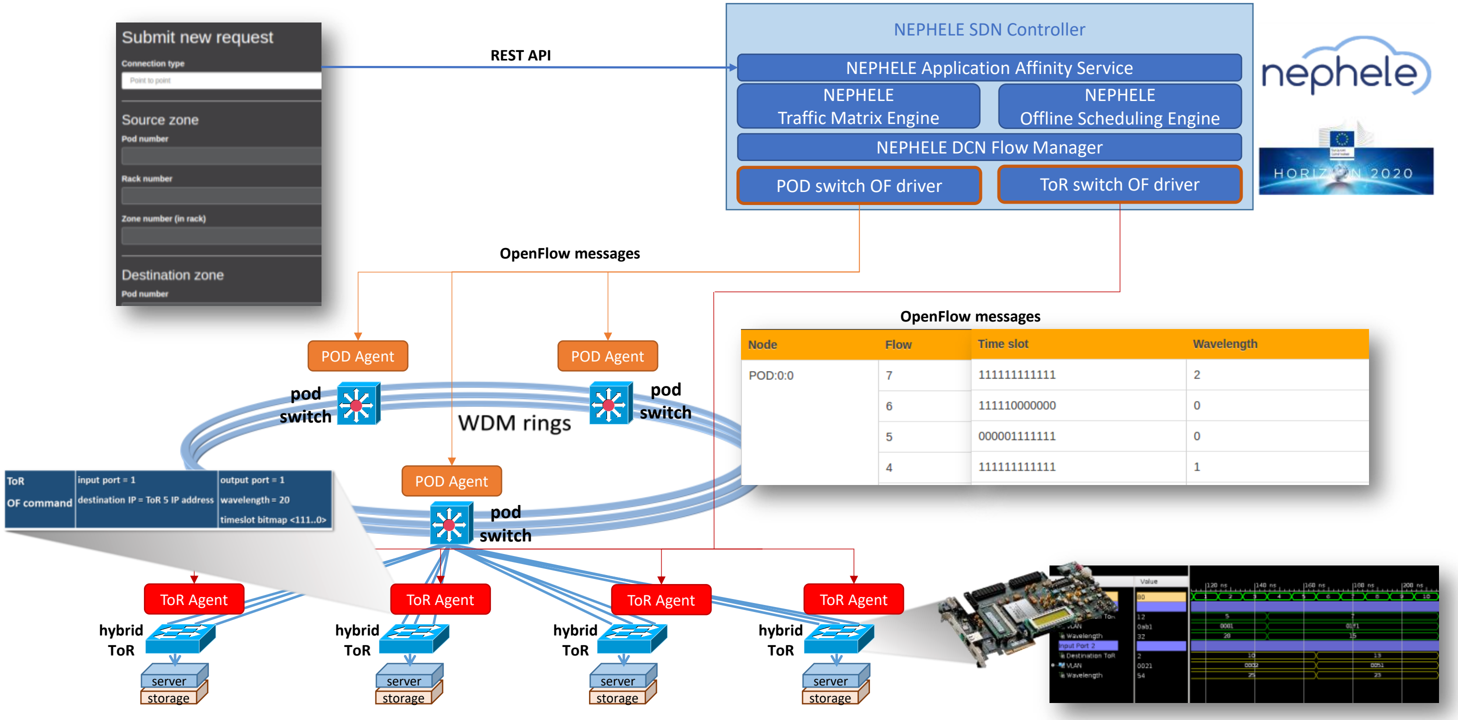
NEPHELE **SDN framework**

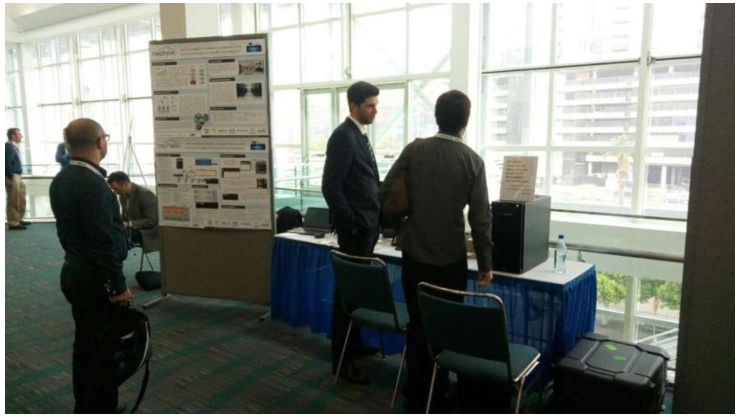


Functionalities of NEPHELE SDN framework

- abstraction of optical switching components (information models)
- translation of TDMA schedule into OpenFlow commands
- integration with open source frameworks

- **Preliminary demo at OFC2017***





*G. Landi et. al., “SDN Control Framework with Dynamic Resource Assignment for Slotted Optical Datacenter Networks”, in Proc. *OFC* 2017.

New **photonic component** technologies in the spotlight

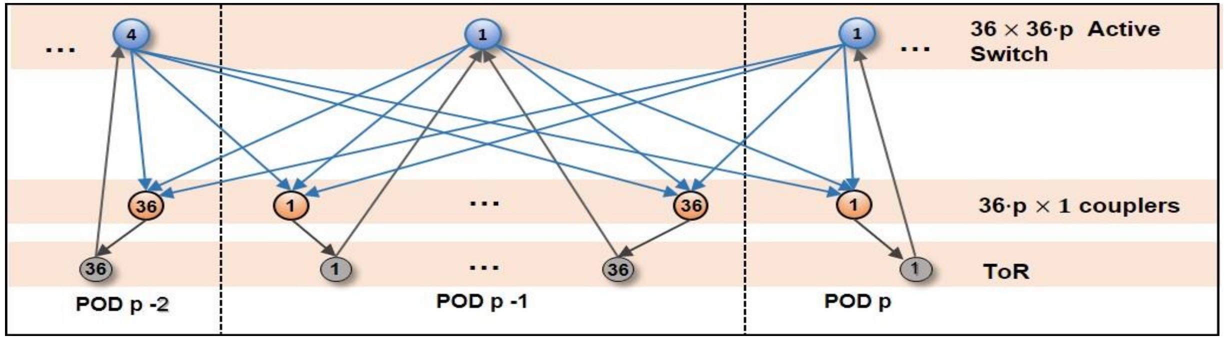
H2020 3PEAT project

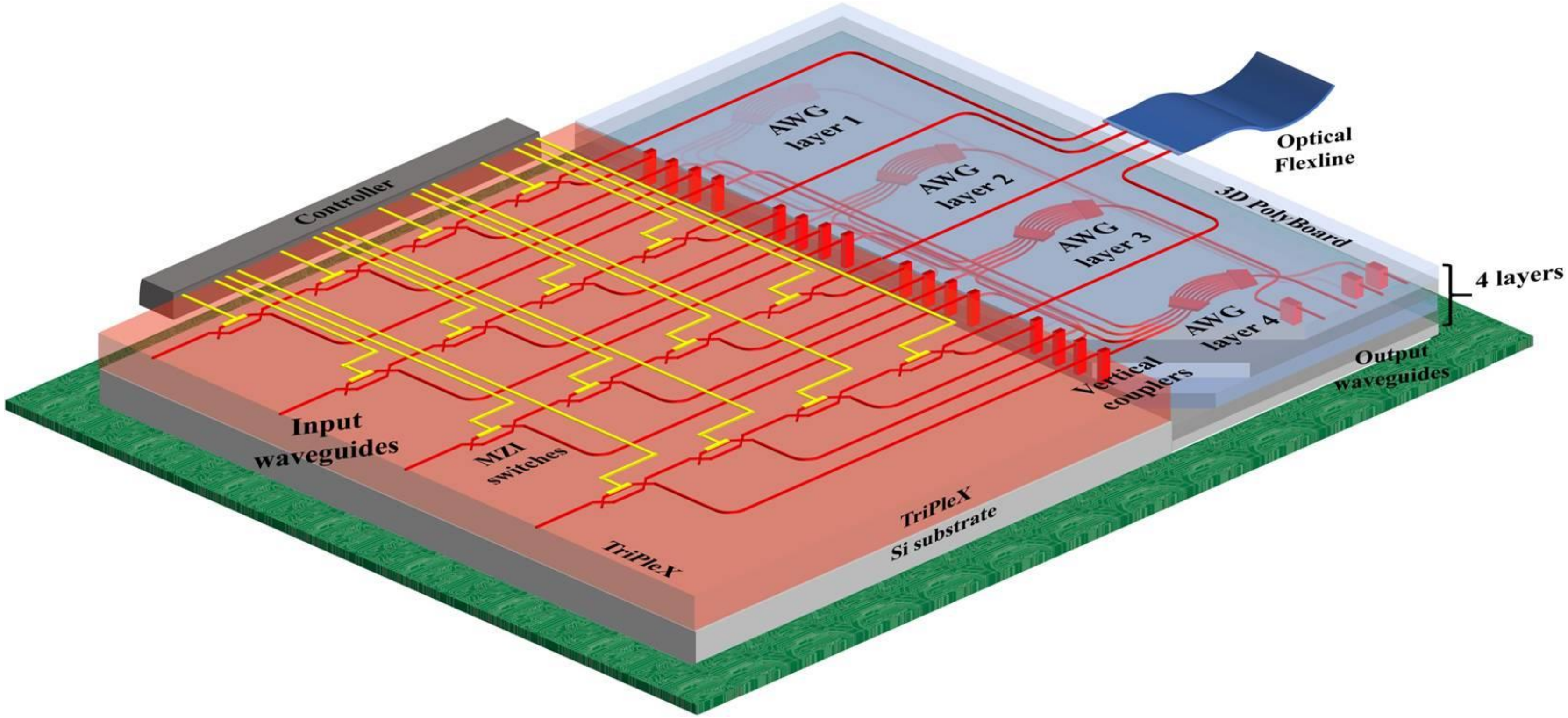
3D Photonic integration platform based on multilayer PolyBoard and TriPleX technology for optical switching and remote sensing and ranging applications

H2020 3PEAT project

Disrupting the application space

- 36 × 36 optical switch
- 20 ns reconfiguration time
- 1.44 Tb/s throughput
- Up to 95% cost savings





Conclusions - Outlook

- NEPHELE network architecture validated experimentally
- Different communication scenarios demonstrated
  - *intra-pod and inter-pod communication from electrical and optical ToR*
  - *error free operation for a wide range of received optical powers, with similar performance*
- Network control and management overarching framework under development
  - *fast and efficient TDMA scheduler*
  - *SDN controller and interfaces with cloud management platform*

Next Steps

- Fully integrate NEPHELE data plane and control plane
- Investigate more forward looking schemes leveraging progress in photonic integration

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 645212 (NEPHELE)

nephele

www.nepheleproject.eu

This presentation reflects only the speaker's view and the Commission is not responsible for any use that may be made of the information it contains.

Thank You

Visit us at booth 531

Mellanox® TECHNOLOGIES

Connect. Accelerate. Outperform.®