Switch architectures for Data Center networks

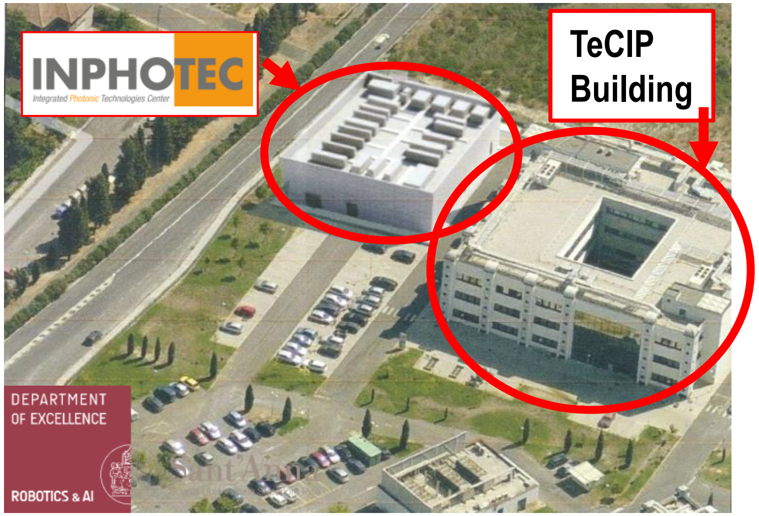
Piero Castoldi
TeCIP Institute, Scuola Superiore *Sant'Anna*

Joint ICTP-IAEA School on Systems-on-Chip Based on FPGA for Scientific Instrumentation and Reconfigurable Computing

November 28, 2023 – ICTP, Trieste

Scuola Superiore Sant'Anna

- **10** Research Institutes and 2 Dept. Of Excellence
- More than **500** undergraduate students
- More than **370** PhD students
- More than **1000** participants to life-long education programmes
- **1:7** teacher/student ratio
- **30%** of overseas students in PhD Programmes
- More than **30M€** of research grants in 2022
- **60+** Spin-off companies incubated and **159** patent families
- **150+** Faculty staff, **200+** Post-Doc Fellowships
- **200+** Administrative staff



TeCIP Institute (Director Prof. P. Castoldi)

TeCIP Building 4.500 m² total surface in the CNR Area of Pisa including

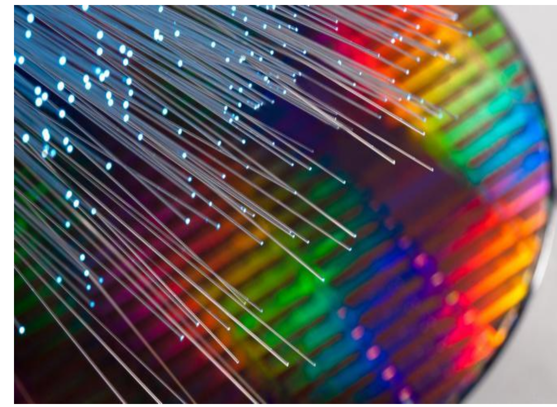
- **Telecommunications Lab, Photonics Lab, Cyber Physical Systems systems Lab**
- **Administration and Lecture Rooms**

colocated with CNIT Photonic Networks and Technologies Lab and Ericsson R&D Lab

Inphotec Clean Room 800 m^2 for PIC Fabrication, Packaging and Characterization

TeCIP Research Organization

The **Institute of Telecommunications, Computer Science and Photonics (TeCIP)** develops fundamental and applied research, education programs and technology transfer in three main disciplinary areas.



TELECOMMUNICATIONS

- Telecommunication Networks and Services
- Optical communication Theory and Techniques
- Optical Wireless and Communication systems

CYBER-PHYSICAL SYSTEMS

- Real-time software
- Cybersecurity and safety-by-design
- Predictable/Trustworthy AI
- Real-time Cloud Computing
- Modeling, optimization and control of industrial processes

PHOTONICS

- Remote Sensing and Microwave Photonics
- Integrated photonics

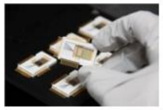
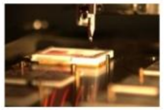
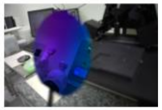
Impact on a variety of vertical industries: **digital healthcare, smart agriculture, space exploration, autonomous driving, navigation, metaverse**, and many other fields.

Teaching

Erasmus Mundus Master initiatives

2018-2022

Photonic Integrated Circuits, Sensors and NETworks – PIXNET Erasmus Mundus Joint Master Degree

(Consortium: Scuola Sant'Anna, OsakaU, AstonU, TUE)

2024-2027, in preparation proposal submitted

Master in phOtonic NetwoRking and clOud Engineering (MONROE)

(Consortium: Scuola Sant'Anna, Glasgow University, Budapest University of Technology and Economics)

Advanced Education Courses for honor students

- Photonic integrated circuits: design, fabrication & packaging
- Deep Learning and Neural Networks
- Advanced operating systems
- System-level security
- FPGA-based platforms and hardware accelerators
- Advanced networking techniques
- Optical communication systems
- Microwave Photonics

PhD in Emerging Digital Technologies (coordinator: Prof. Luca Valcarenghi)

It is a 3-year residential program with highly interdisciplinary connotation, involving structured courses and supervised research in our laboratories.

It includes three curricula targeting fundamental research or industrial research:

- **Photonic Technologies**
- **Embedded Systems**,
- **Perceptual Robotics** (supported by Institute of Intelligent Mechanics)

Seasonal School "Pervasive ARTIficial Intelligence for Next-G Softwarized Networks (ARTIST)" (coordinators: Prof. Piero Castoldi, Prof. Luca Valcarenghi)

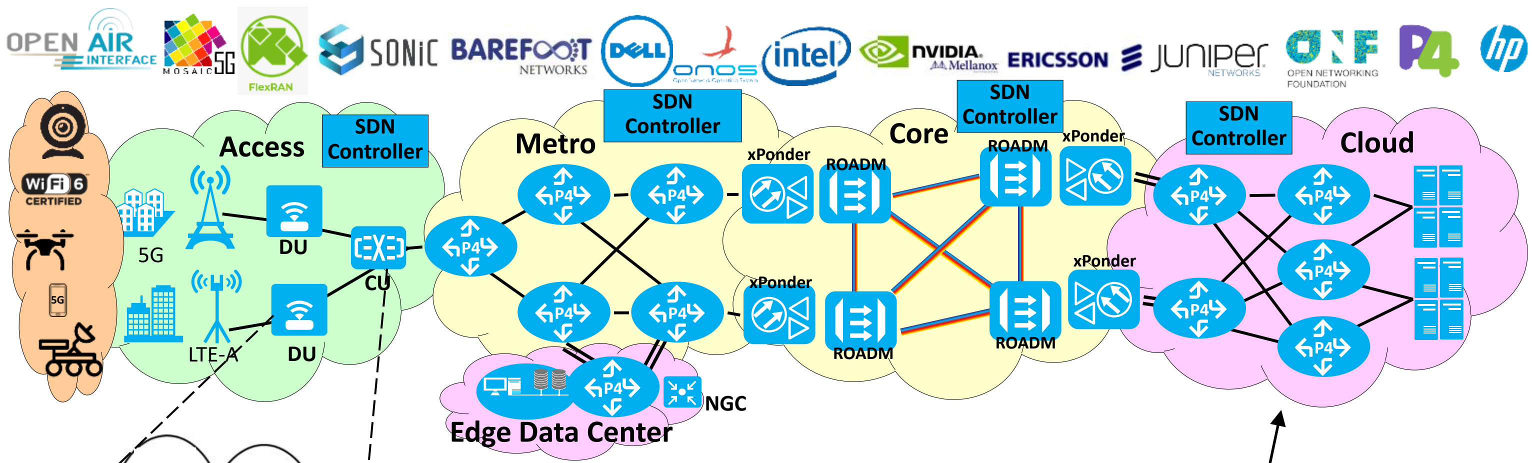
One-week – in presence – accommodation and canteen provided for free by the School

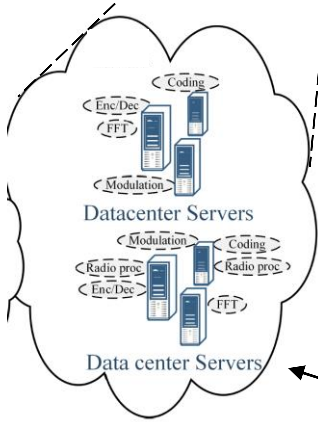
Period: 3rd-7th June 2024, Deadline for application: April 22nd, 2024 (application interface not yet open)

https://www.santannapisa.it/en/seasonalschool/artist

Data Center Networks … Motivations grounds

- The full picture of a network operator telecommunication infrastructure:





For services requiring real-time or low-latency in-network computation

Often owned by operators but mostly by the over-the-top

Data Centers exclusively devoted for mo-demodulation and separation of traffic coming from the mobile interface

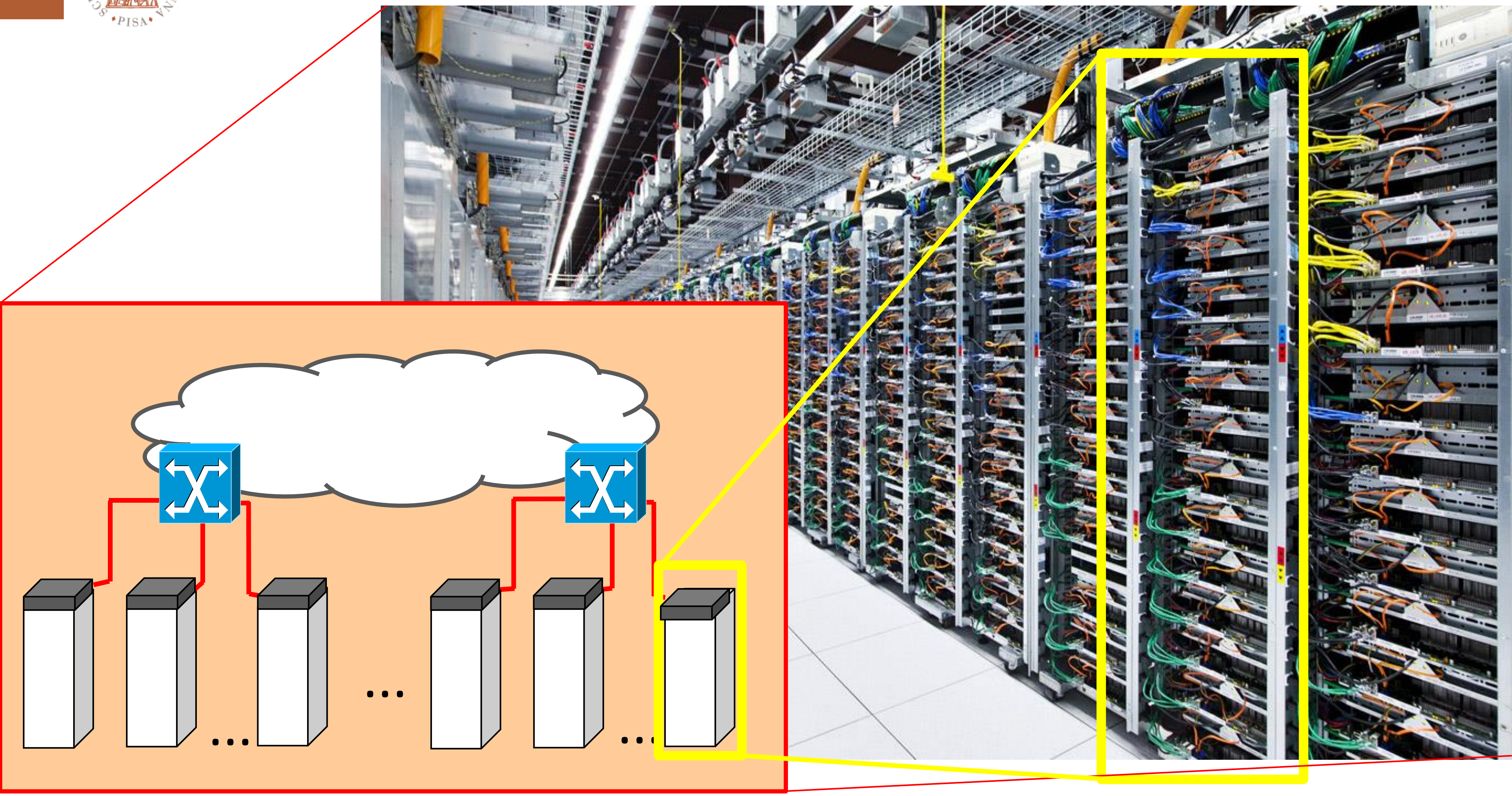
Outline

- Data Center Architectures

- Optical interconnection network based on space matrices
- Performance of optical space matrices

- Multi-MicroRing (MMR) optical interconnection network
- Scheduling in MMR
- FPGA-based MMR network validation
- Performance of MMR interconnection network

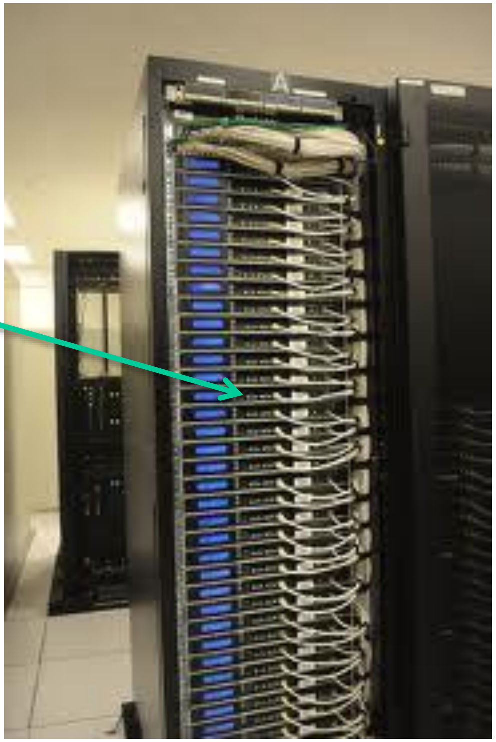
Server racks in Google DC



Each server rack has many blades (severs) that are interconnected through swiyches hierachically

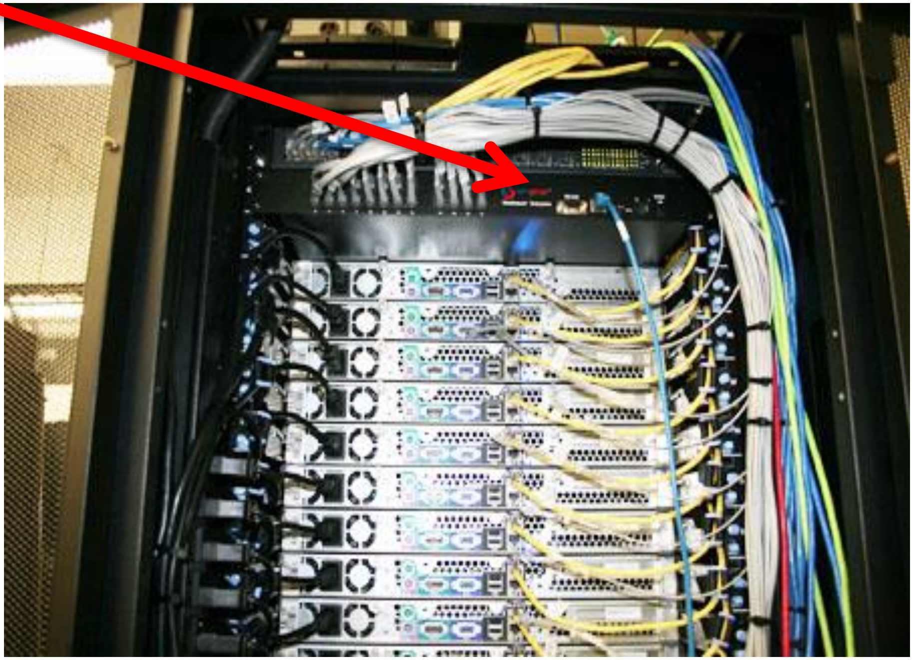
What is in a datacenter (network)?

- Servers (blades) organized in racks



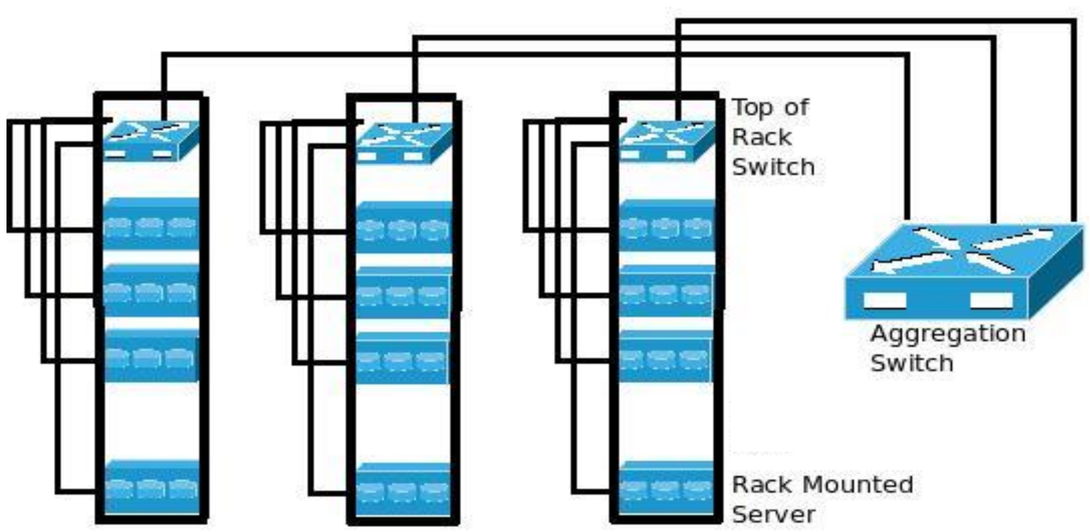
What is into a datacenter (network)?

- Servers (blades) organized in racks
- Each rack has a `Top of Rack' (ToR) switch



What is in a datacenter (network)?

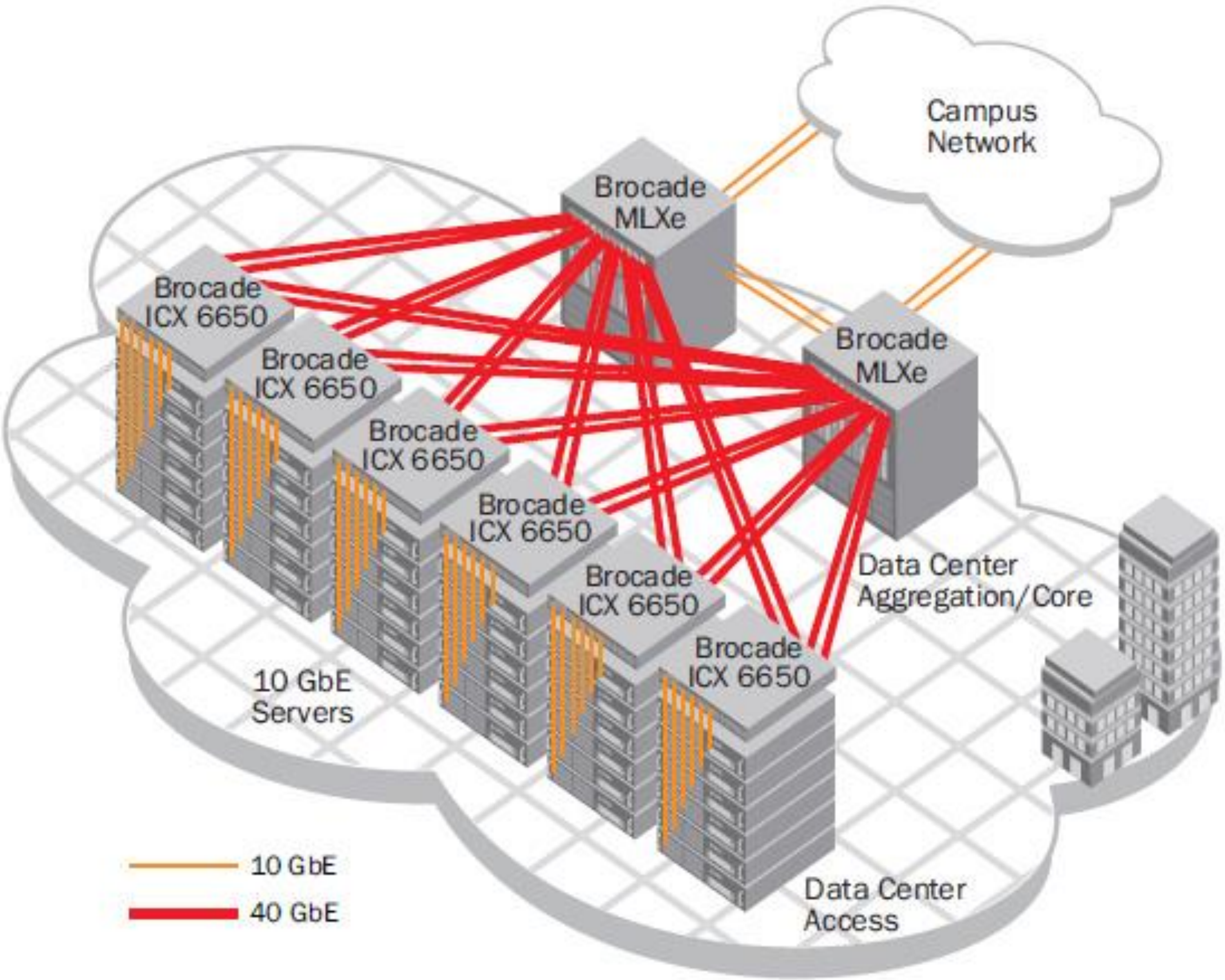
- Servers (blades) organized in racks
- Each rack has a `Top of Rack' (ToR) switch
- `Aggregation switches interconnect ToR switches



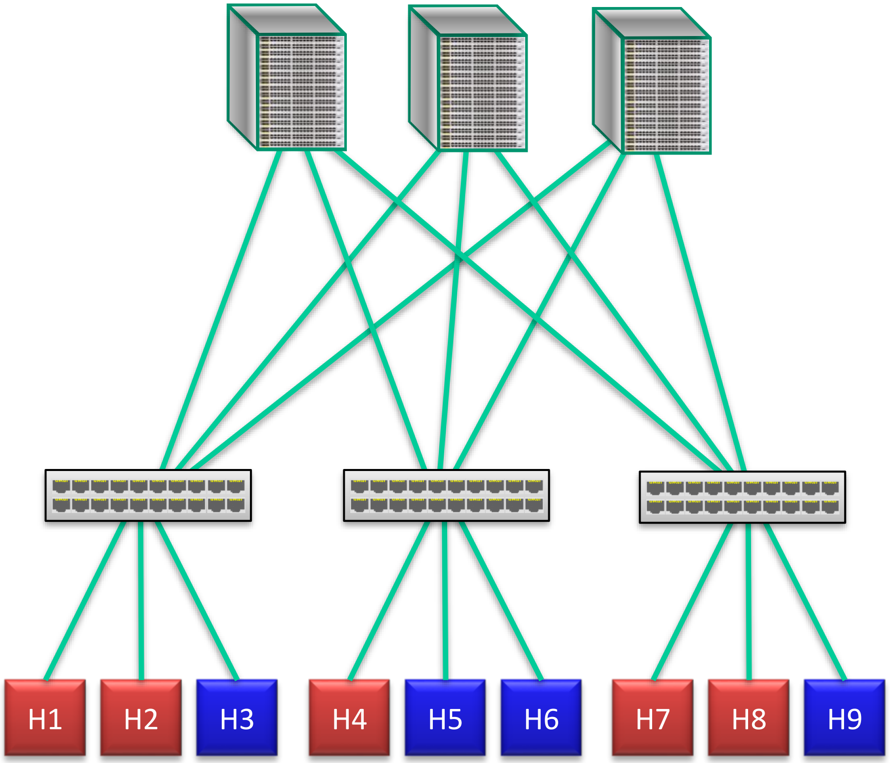
What is in a datacenter (network)?

- Servers (blades) organized in racks
- Each rack has a `Top of Rack' (ToR) switch
- `Aggregation switches interconnect ToR switches
- Connected to the outside via `core' switches
  - note: blurry line between aggregation and core
- With 2x redundancy for fault-tolerance

E.g., Brocade Reference Design

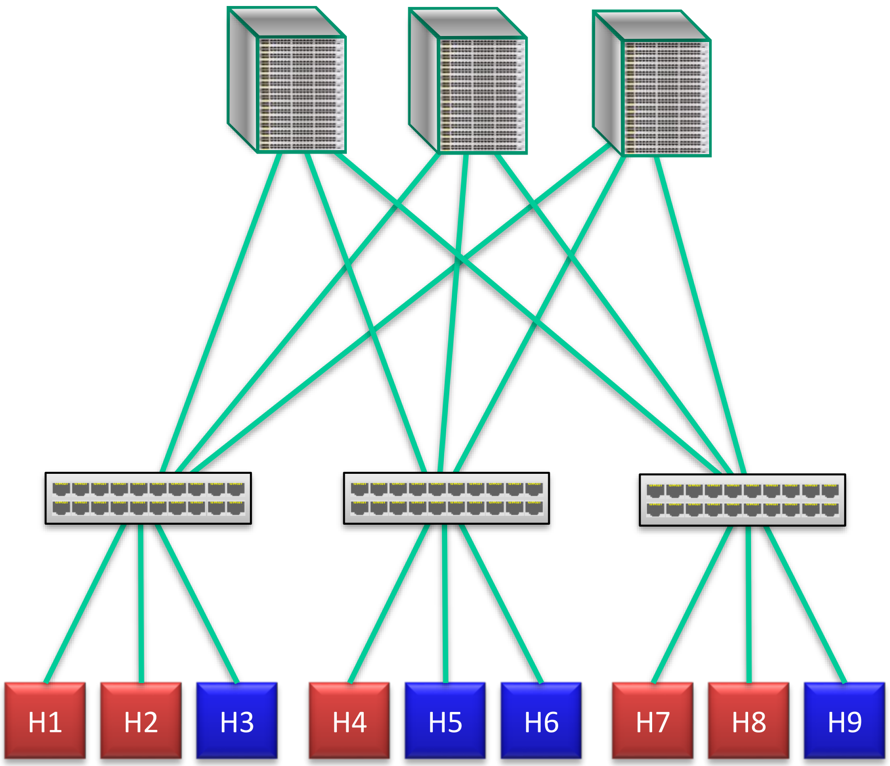


DC Network: Just a Giant Switch!



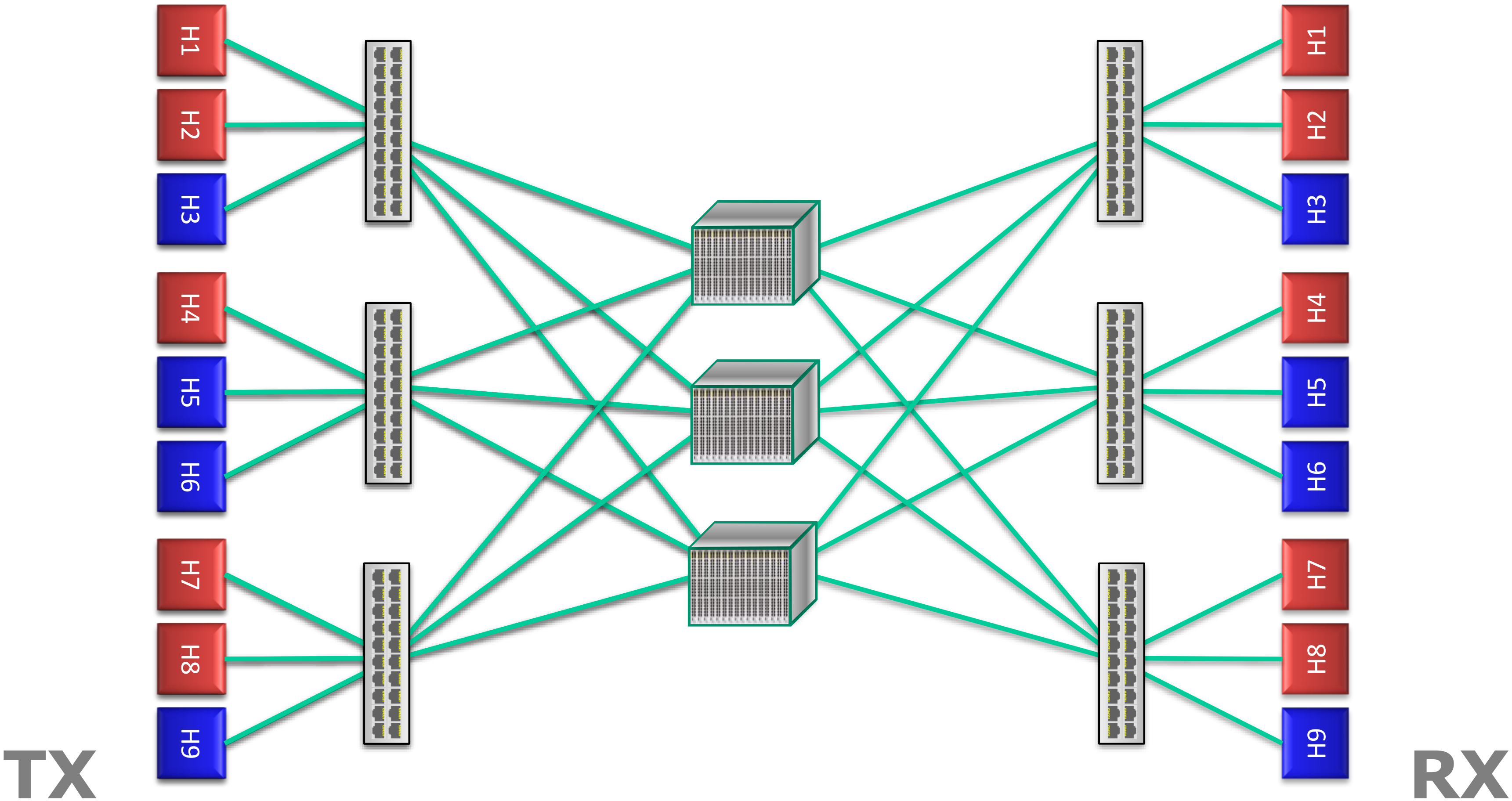
Slides from: Alizadeh, HotNets 2012

DC Network: Just a Giant Switch!



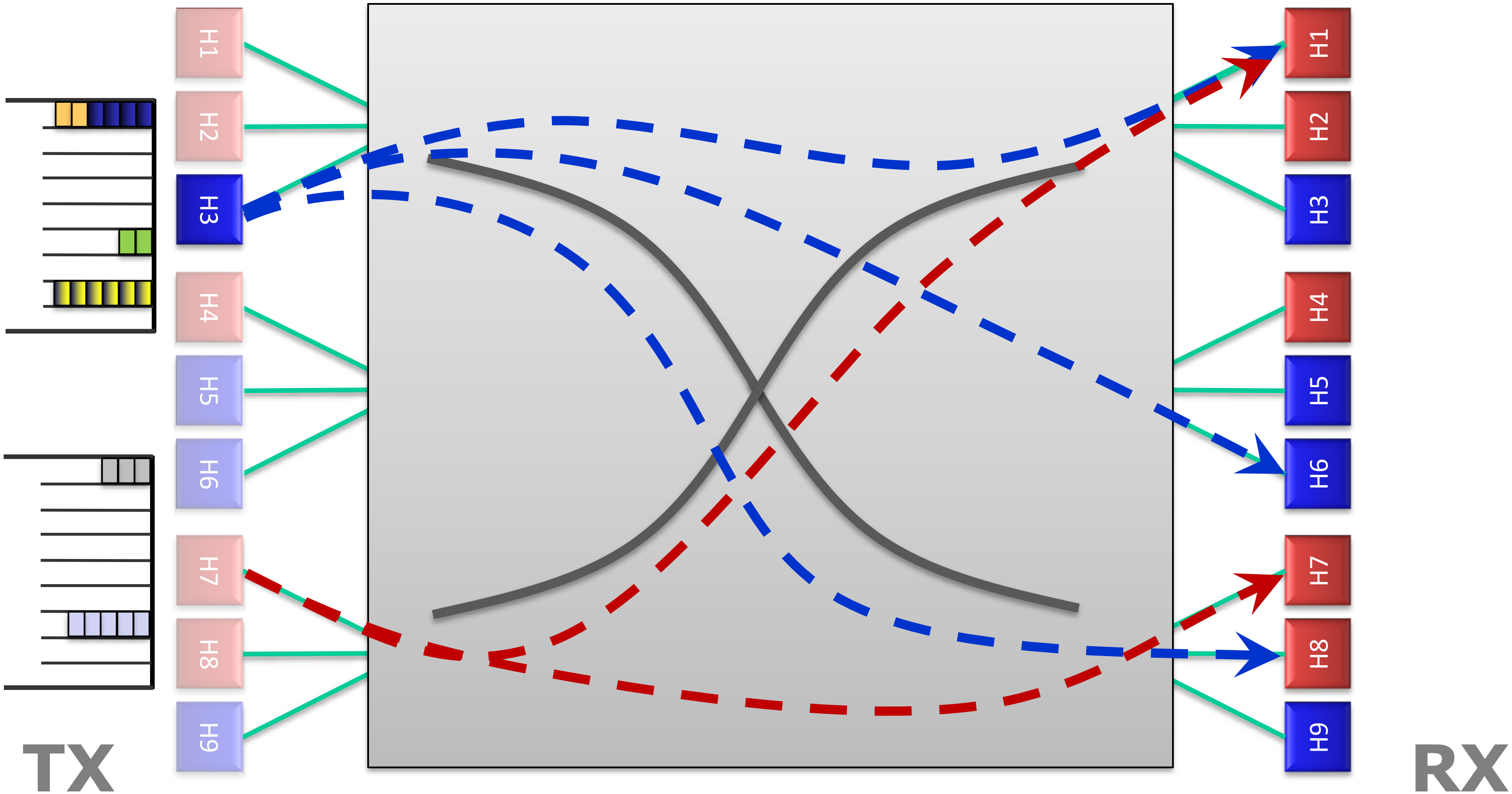
Slides from: Alizadeh, HotNets 2012

DC Network: Just a Giant Switch!



Slides from: Alizadeh, HotNets 2012

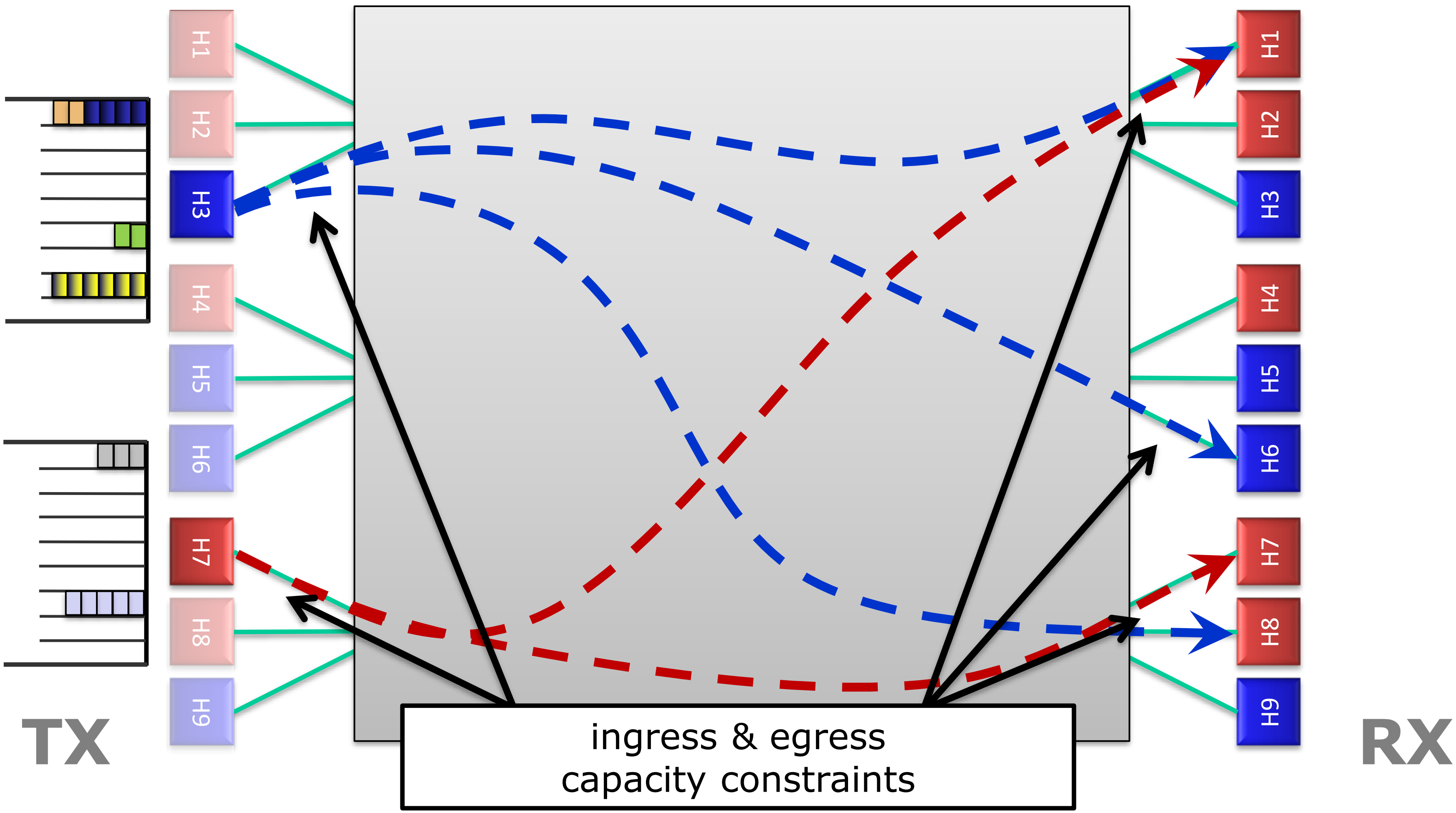
DC Network: Just a Giant Switch!



Slides from: Alizadeh, HotNets 2012

DC Network: Just a Giant Switch!

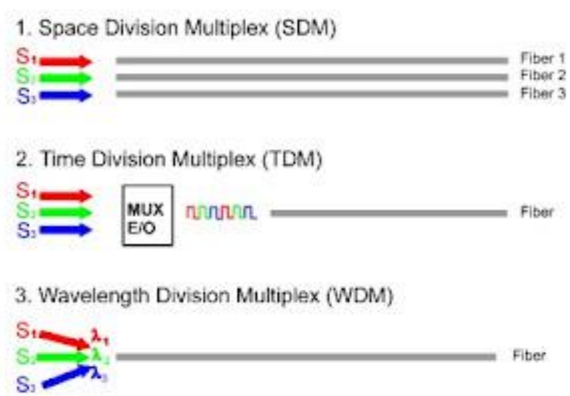
DC transport = Flow scheduling on giant switch

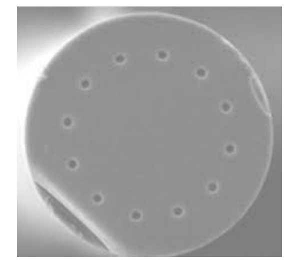


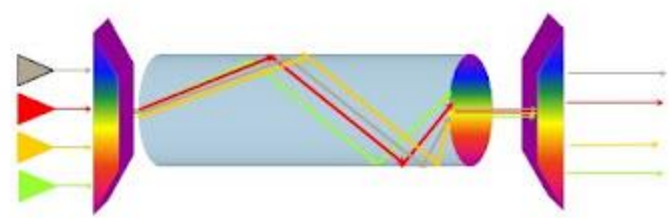
Slides from: Alizadeh, HotNets 2012

Optics can help scaling interconnection networks?

- **Photonic solutions** can contribute to alleviate limitations of current electrical interconnection networks
  - \+ Increased bandwidth at low power consumption
  - \+ No electromagnetic interference
  - \+ No delay variance
  - – No buffering
- Switching domains







- The most promising elements in term of scalability, integration capabilities and footprint are:
  - ➢ Microring resonators
  - ➢ Semiconductor optical amplifiers (SOAs)

18

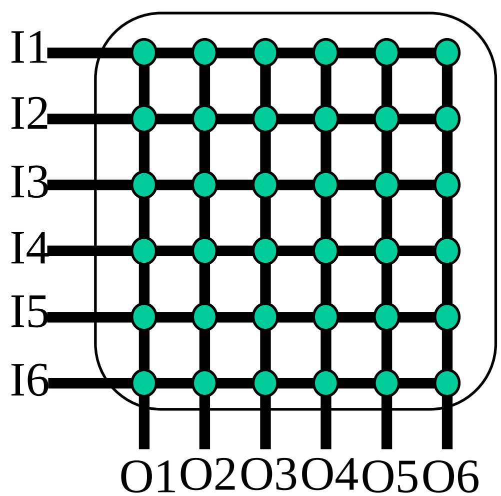
Semiconductor Optical Amplifiers and microring resonators as optical gating elements

SOAs

- ↓ More power hungry
- ↑ Amplification
- ↑ Fast switching time
- ↑ Wide-band

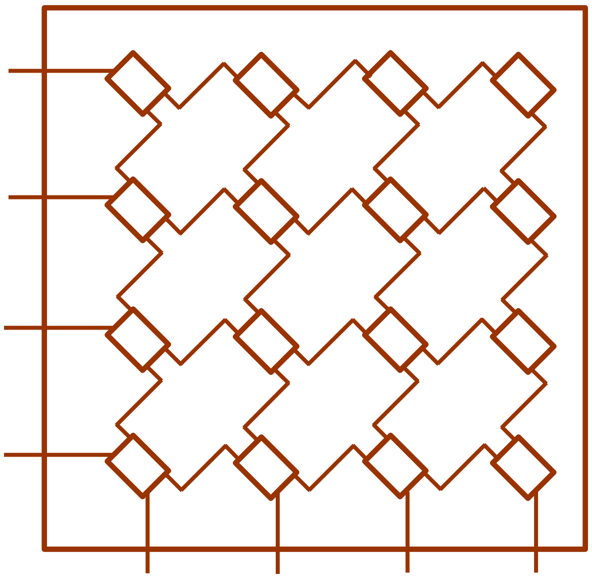
Microrings

- ↑ Low power consumption
- ↑ Small footprint
- ↓ Difference in attenuation between drop/through port
- ↓ Narrow-band



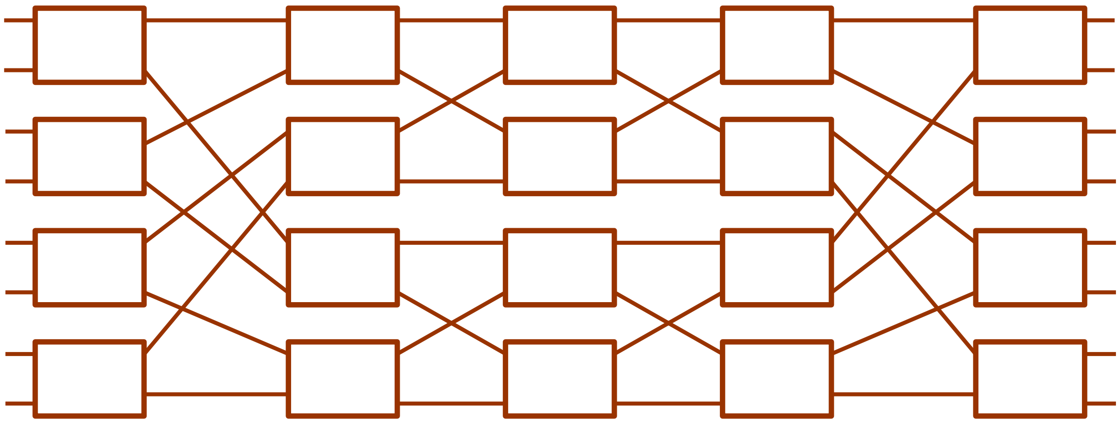
Space switching architectures

Crossbar



Non blocking

Benes

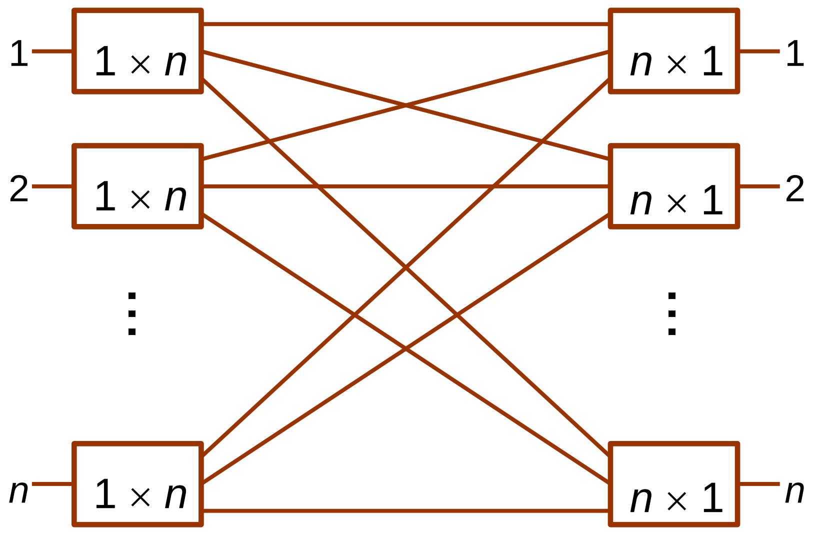


(Clos network where stages use 2 x 2 switches)

Rearrangeably Non blocking

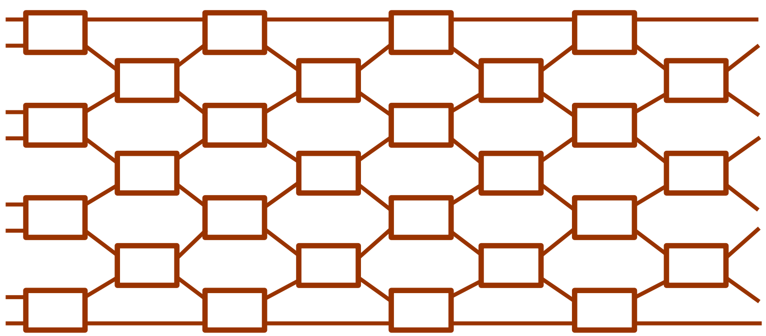
Space switching architectures /2

Spanke



- Non blocking
- Does not take advantage of 2x2 switching elements

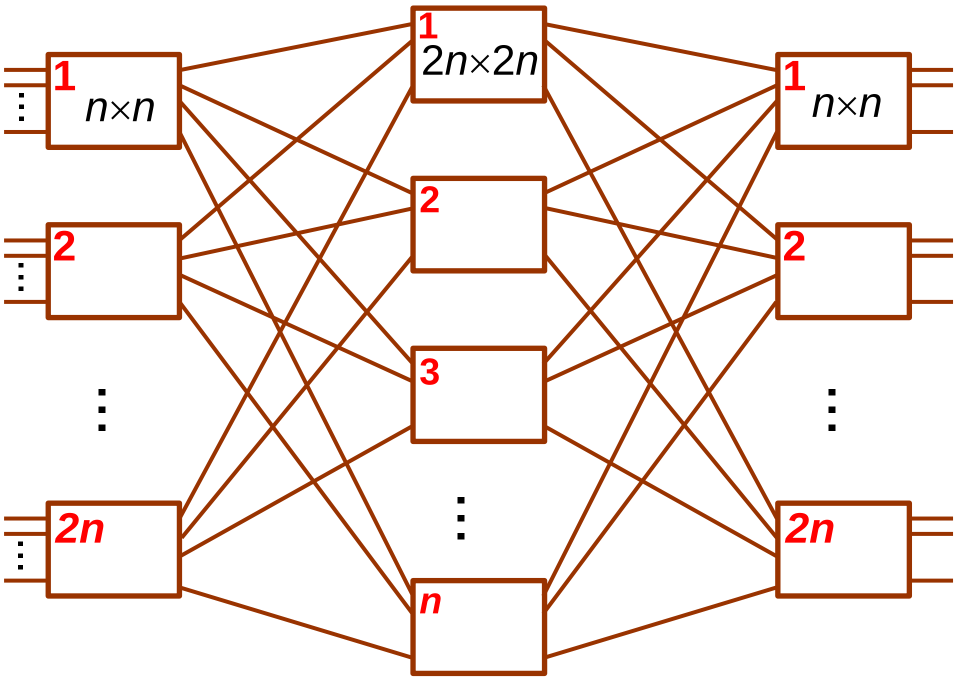
Spanke-Benes



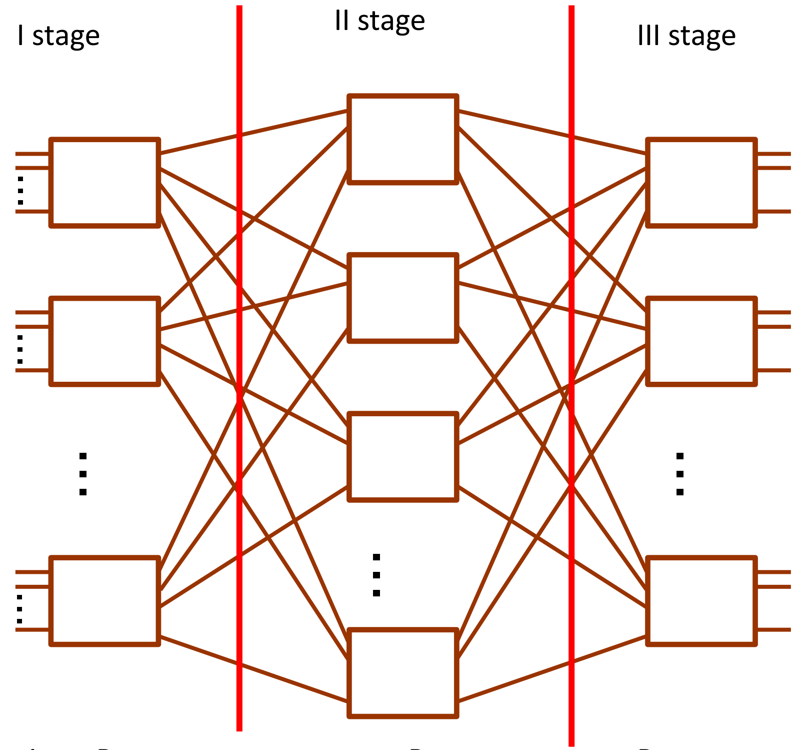
Also called N-Stage planar

Multi-stage architectures

Clos

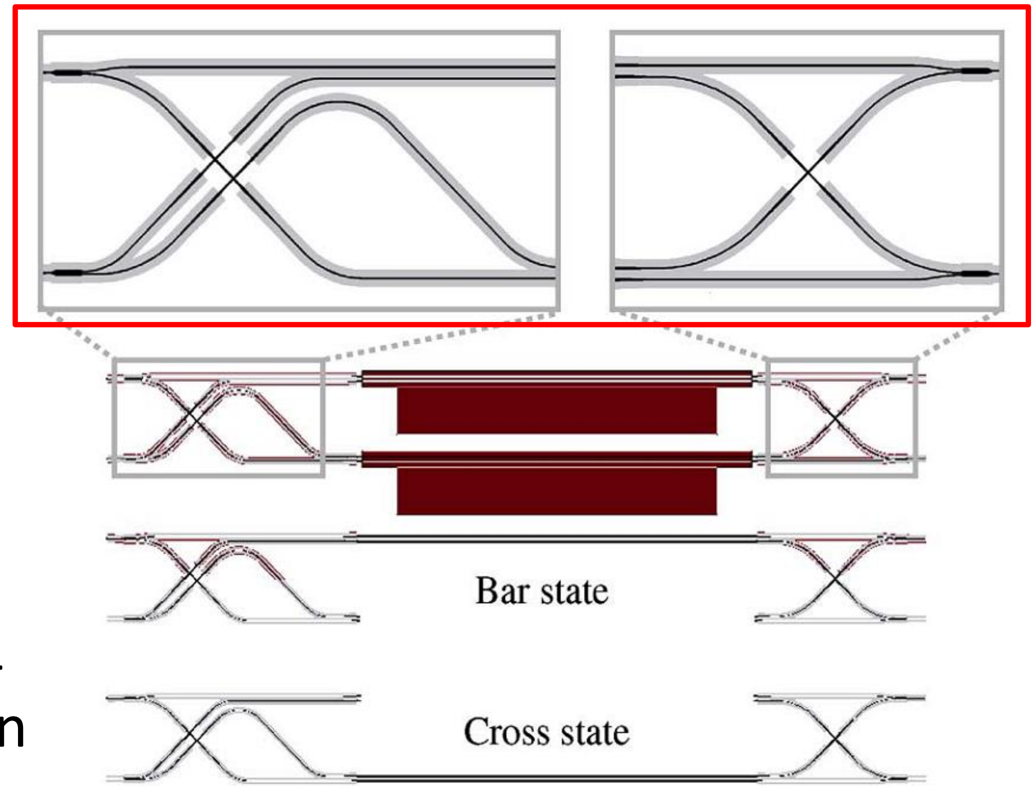


3-stages Clos network with minimum number of elements



	I stage	II stage	III stage
1.	Benes	Benes	Benes
2.	Spanke	Spanke	Spanke
3.	Benes	Spanke	Benes
4.	Spanke	Benes	Spanke

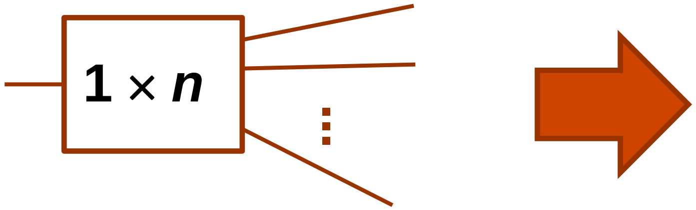
2×2 switching element

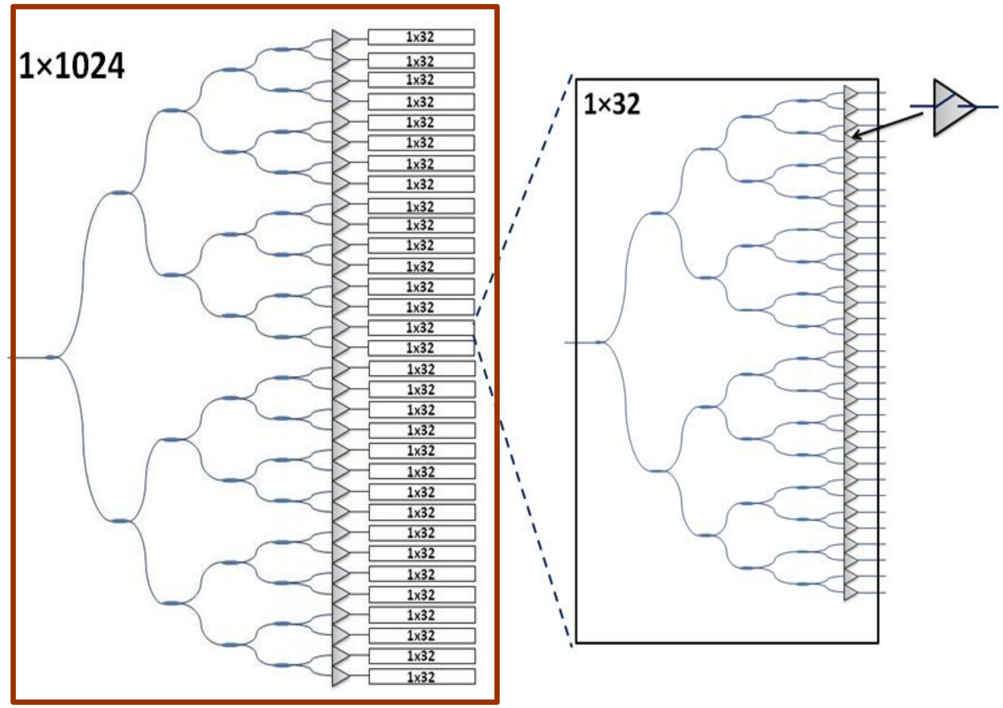


- The 2×2 switching element can be implemented by taking advantage of 2 SOAs
- In both cross and bar state 1 SOA is in active state and 1 in inactive state

Albores-Mejia, et al., "Monolithic Multistage Optoelectronic Switch Circuit Routing 160 Gb/s Line-Rate Data," *JLT,* vol.28, no.20, Oct.15, 2010

1×n switching element



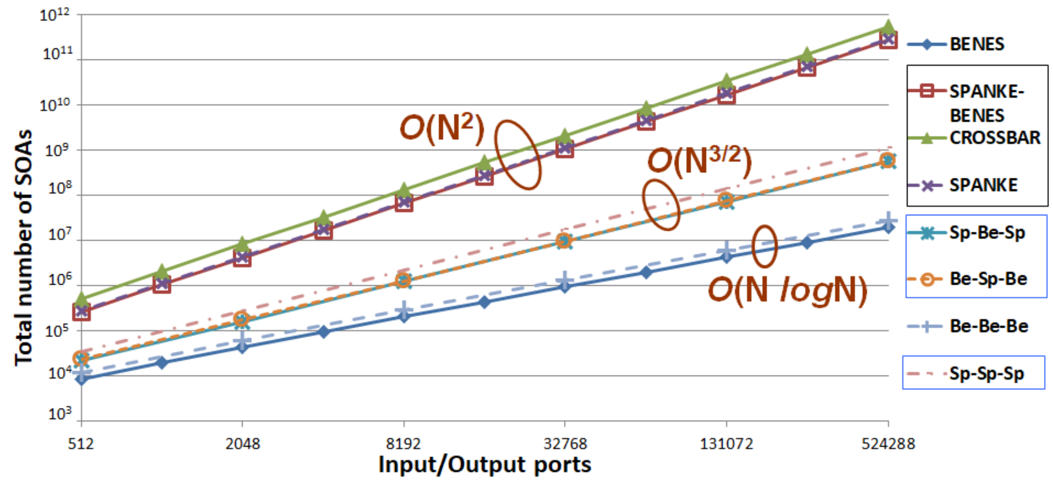


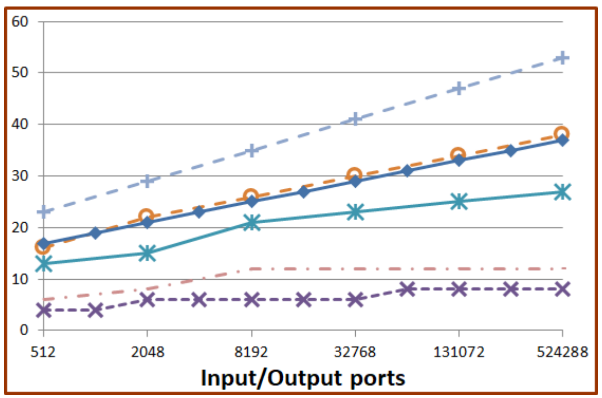
- The 1×n and n×1 switching in Spanke architecture can be implemented as binary trees, using SOAs as gates and amplifiers

O. Liboiron-Ladouceur, I. Cerutti, P.G. Raponi, N. Andriolli, P. Castoldi, "Energy-Efficient Design of a Scalable Optical Multiplane Interconnection Architecture," *JSTQE* , vol.17, no.2, Mar.-Apr. 2011

Take-aways on space-switching architectures

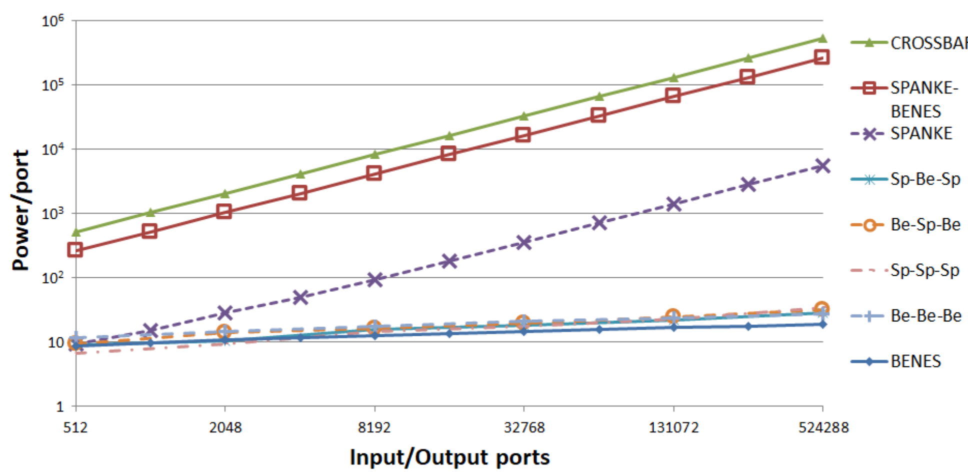
- Analyzing the total number of SOAs required, the best architectures are Benes and Be-Be-Be





- From the scalability (maximum number of elements crossed by a path) perspective the best solutions are Spanke, Sp-Sp-Sp

- Benes architecture shows the lowest power consumption per port



(Silicon Photonic) Multi-Microring Optical Interconnection Network

1. Photonic integration enabling optical interconnection networks [1]

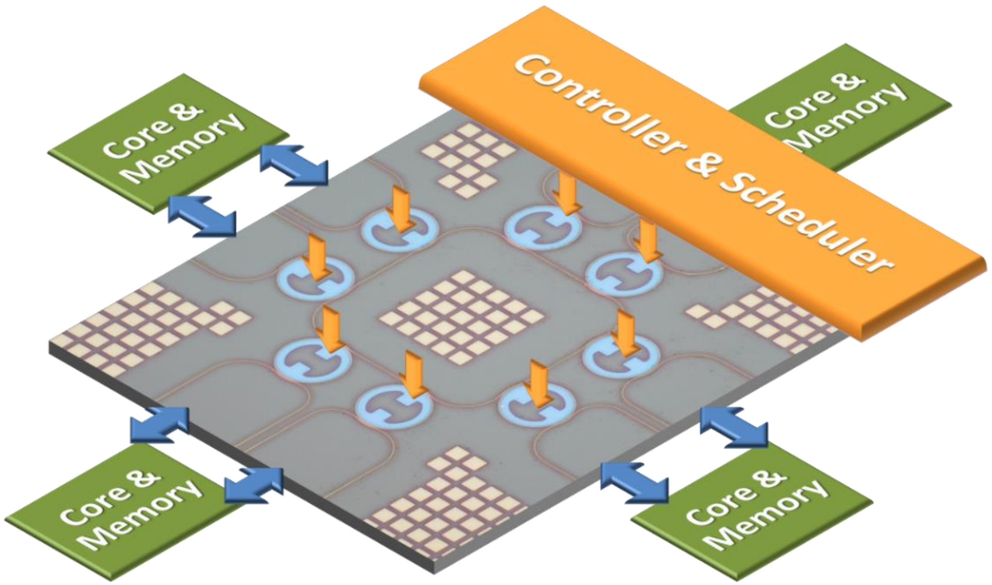
- Bandwidth density potential
- Compact footprint
- Leveraging CMOS infrastructure

2. Network with a ring topology [2]

- Avoids waveguide crossings
- WDM for concurrent transmission

3. Network dynamic control [3]

- High throughput
- Low queuing latency



[1] A. K. Ziabari et al., "Leveraging Silicon-Photonic NoC for Designing Scalable GPUs," in *Proc. ACM ICS 2015*.

[2] P. Pintus, P. Contu, P. Raponi, I. Cerutti, and N. Andriolli, "Silicon-based all-optical multi microring network-on-chip," *Opt. Lett.*, 2014.

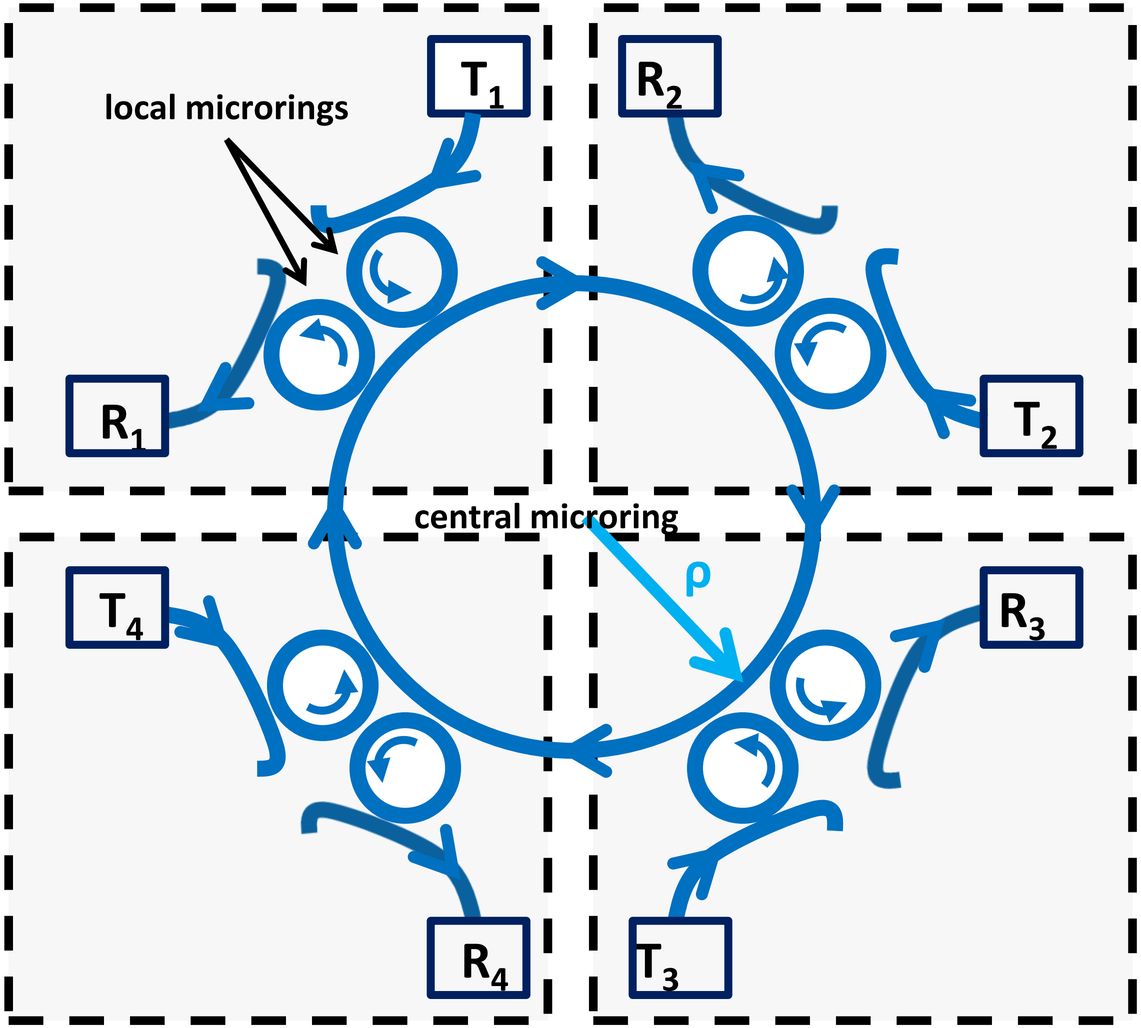
[3] I. Cerutti, N. Andriolli, P. Pintus, S. Faralli, F. Gambini, P. Castoldi, and O. Liboiron-Ladouceur, "Fast scheduling based on iterative parallel wavelength matching for a multi-wavelength ring network-on-chip," in *Proc. ONDM2015*.

Multi-MicroRing (MMR) optical interconnection network

- Central microring:
  shared waveguide
- Local microrings:
  used to either
  - add (from T_i) or
  - drop (to R_i)

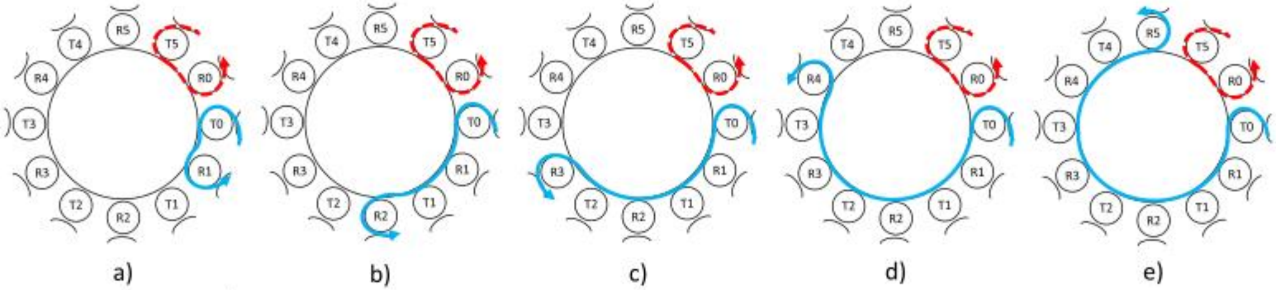
  the optical signals

- Simultaneous transmissions on the same wavelength are possible if their paths are disjoint
  → **Spatial reuse**
- Parallel transmissions on multiple wavelengths are possible by exploiting WDM → **Wavelength reuse**

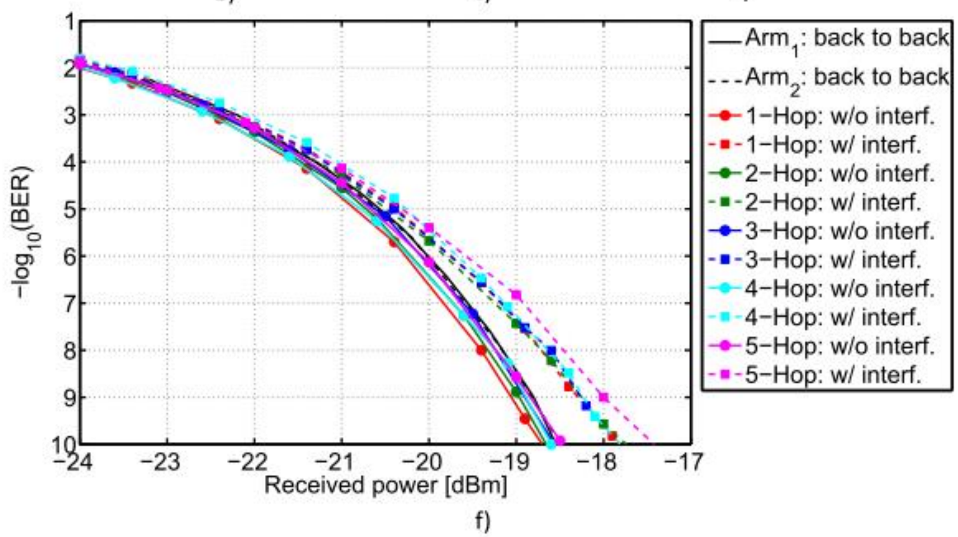


T_i : fixed or tunable transmitter
R_i : broadband photoreceiver

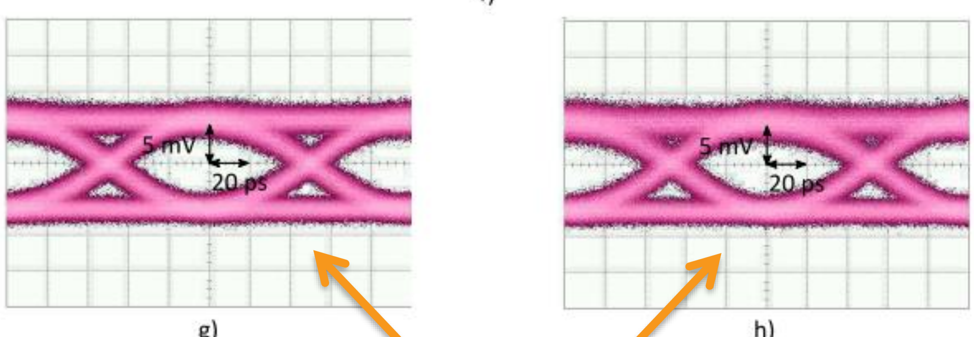
Two co-propagating transmissions



Intended transmission (blue) of different length, with the one-hop upstream interfering transmission (red) at the same wavelength



- The power penalty related to the upstream transmission is limited to 0.5 dB at BER = 10^{-9}
- Without the interfering transmission, the measured BER slightly outperforms the back-to-back BER due to the filtering effect of the rings, which act as adapted receivers [5]



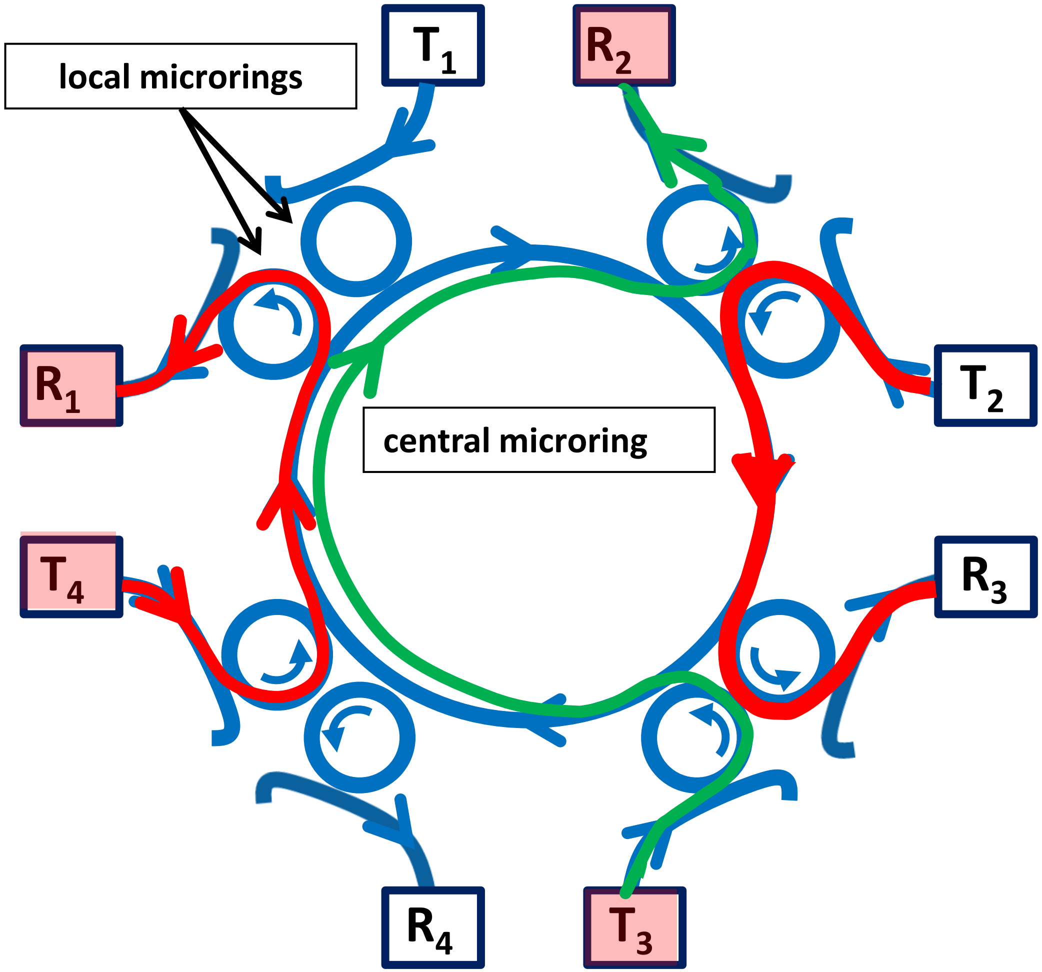
[5] A. Parini, et al., "BER evaluation of a passive SOI WDM router," IEEE Photon. Technol. Lett. 25(23) (2013).

Eye-diagrams for 1-hop and 5-hop transmissions in presence of the interfering transmission → No significant degradation

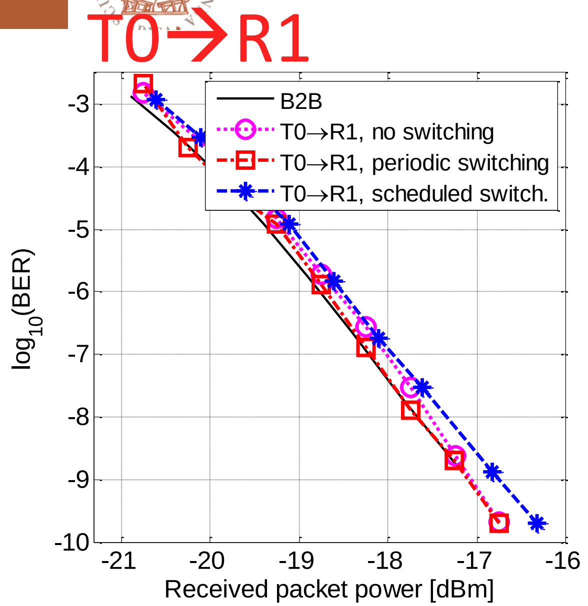
Scheduling in Multi-Microring (MMR) network-on-chip

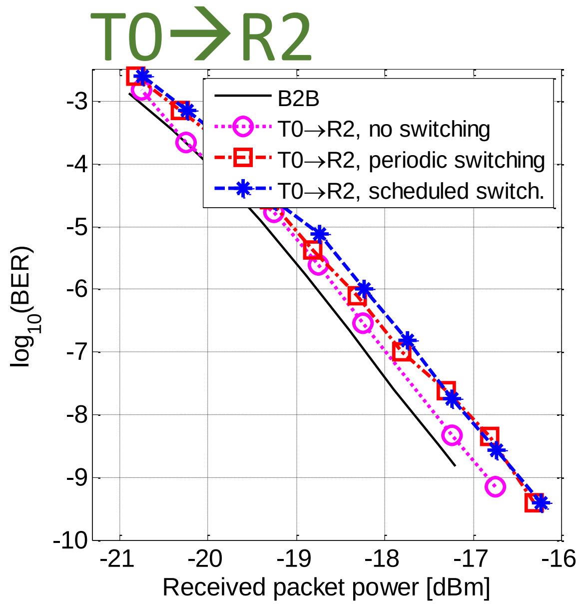
Required to avoid that two or more packets are simultaneously transmitting on the same wavelength and along the same link(s), leading thus to collisions

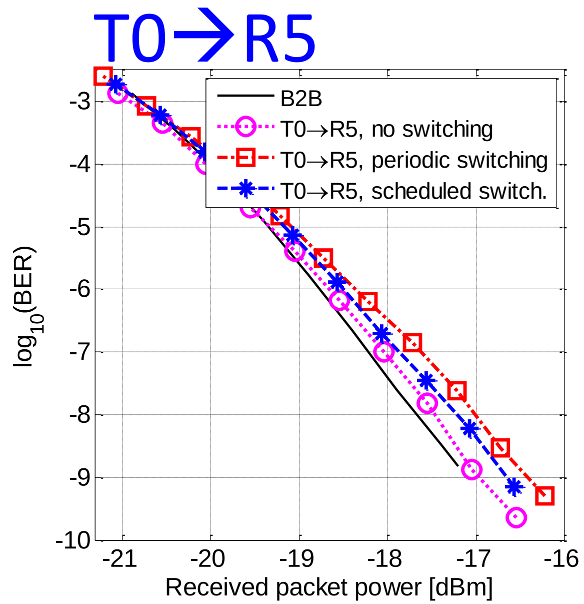
Since the interference among different transmissions is limited, spatial and wavelength reuse can be exploited



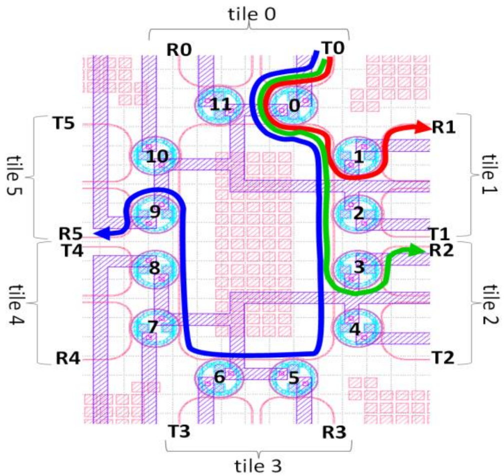
Dynamic Switching BER Performance



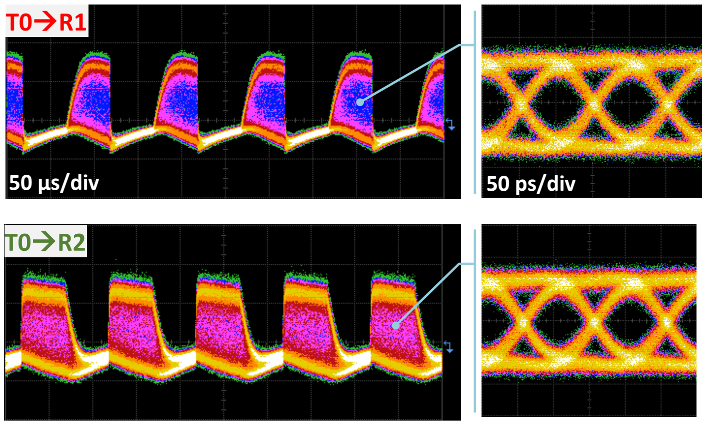




BER differences for all destinations < 0.5 dB at BER of 10^{-9}



- Data transmission: 10 Gb/s
- Time slots: 50 µs
- Guard time: 10 µs
- Network load: 90%
- Contention resolved by scheduler



Periodic packet switching

Y. Xiong, N. Andriolli, S. Faralli, F. Gambini, P. Pintus, M. Chiesa, R. Ortuño, O. Liboiron-Ladouceur, and I. Cerutti, «Demonstration of a Packaged Photonic Integrated Network on Chip controlled by an FPGA-based scheduler», in OFC2017 Tech. Dig.

Take-aways on microring architectures

- A photonic integrated multi microring (MMR) architecture can overcome the issues raised by current electrical interconnection network
- Photonic integrated circuits implementing the MMR have been designed, fabricated, packaged, and characterized in terms of spectral and BER performance
- The propogating signal has been tested in a real prototype operating at 10Gb/s per wavelengths
- Dynamic switching of packets has been demonstrated in a MMR controlled by an FPGA-based scheduler

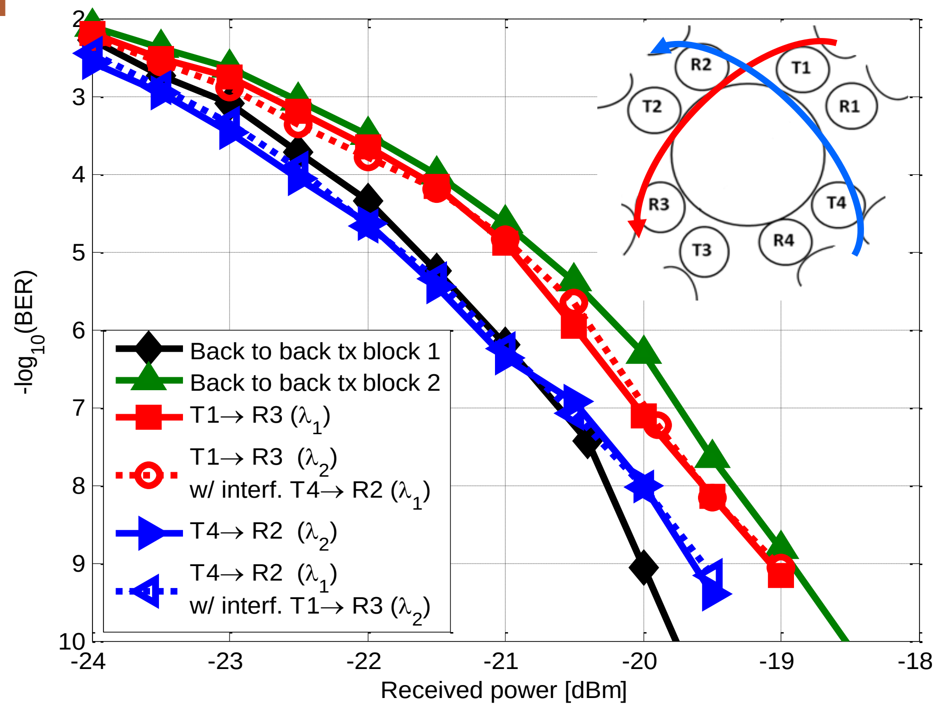
Selected Publications

- Journals
  - N. Andriolli, A. Giorgetti, P. Castoldi, G. Cecchetti, I. Cerutti, N. Sambo, A. Sgambelluri, L. Valcarenghi, F. Cugini, B. Martini, F. Paolucci, (2022). Optical networks management and control: A review and recent challenges. OPTICAL SWITCHING AND NETWORKING, vol. 44, ISSN: 1573-4277, doi: 10.1016/j.osn.2021.100652
  - Borromeo J. C., Cerutti I., Castoldi P., Reyes R., Andriolli N. (2021). FPGA-based implementation of two-step schedulers for modular optical interconnection networks. JOURNAL OF OPTICAL COMMUNICATIONS AND NETWORKING, vol. 13, p. 116-125, ISSN: 1943-0620, doi: 10.1364/JOCN.417897
  - Cerutti, I., Acmad, M. N. A., Reyes, R., Castoldi, P., Andriolli, N. (2018). Scheduling in multi-wavelength ring-based optical networks-on-chip. JOURNAL OF OPTICAL COMMUNICATIONS AND NETWORKING, vol. 10, p. 322-331, ISSN: 1943- 0620, doi: 10.1364/JOCN.10.000322
  - S. Faralli, F. Gambini, P. Pintus, M. Scaffardi, O. Liboiron Ladouceur, Y. Xiong, P. Castoldi, F. di Pasquale, N. Andriolli, I. Cerutti, (2016). Bidirectional Transmission in an Optical Network on Chip With Bus and Ring Topologies. IEEE PHOTONICS JOURNAL, vol. 8, p. 1-7, ISSN: 1943-0655, doi: 10.1109/JPHOT.2016.2526607
  - F. Gambini, P. Pintus, S. Faralli, M. Chiesa, G.B. Preve, I. Cerutti, and N. Andriolli, “Experimental demonstration of a 24-port packaged multi-microring network-on-chip in silicon photonic platform,” Opt. Express, vol. 25, no. 18, Sep. 4, 2017, pp. 22004-22016.

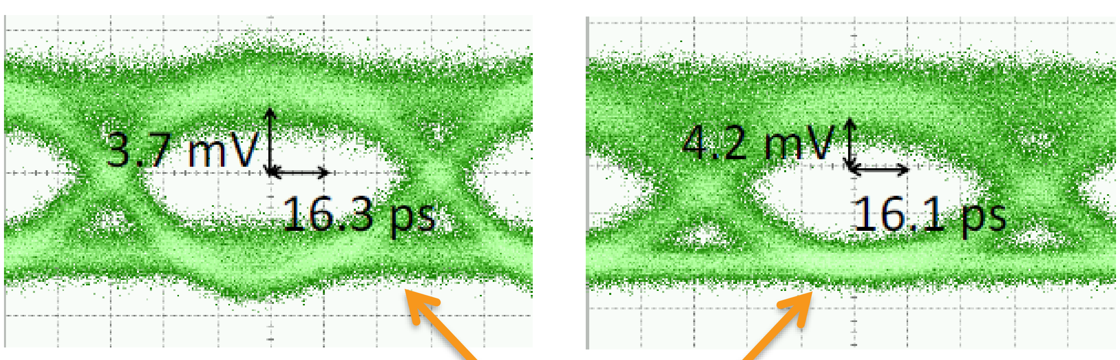
Thank you! Q&A

Email: piero.castoldi@santannapisa.it

Transmissions on different wavelengths



- Two concurrent transmissions on partially overlapping paths: T1→R3 on λ_1 and T4→R2 on λ_2
- Low crosstalk between transmissions on different wavelengths → BER of a given transmission is not affected by the presence of the other transmission

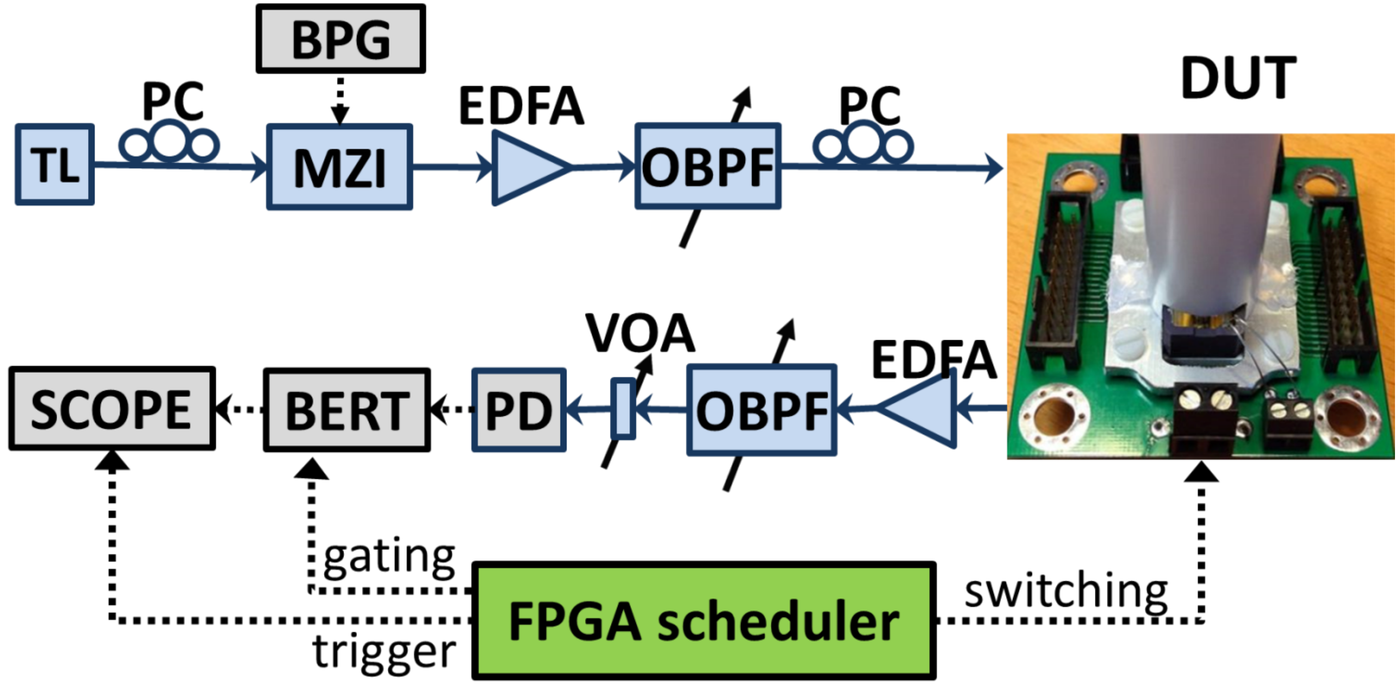


(c) (d)

F. Gambini, P. Pintus, S. Faralli, M. Chiesa, G.B. Preve, I. Cerutti, and N. Andriolli, “Experimental demonstration of a 24-port packaged multi-microring network-on-chip in silicon photonic platform,” *Opt. Express*, vol. 25, no. 18, Sep. 4, 2017, pp. 22004-22016.

Eye-diagrams of both intended transmissions in the presence of their respective interfering transmission

Validation test bed setup



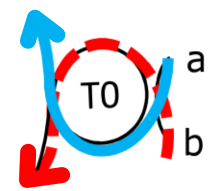
TL: Tunable laser; **PC**: Polarization controller; **BPG**: Bit pattern generator; **MZI**: Mach-Zehnder Interferometer modulator (10 Gbps); **EDFA**: Erbium-doped fiber amplifier; **OBPF**: Optical bandpass filter (1.3 nm BW); **DUT**: Device under test; **VOA**: Variable optical attenuator; **PD**: Photodetector; **BERT**: Bit error rate tester.

FPGA scheduler provides synchronized gating, switching and trigger signals for BER tester, optical switches and scope, respectively.

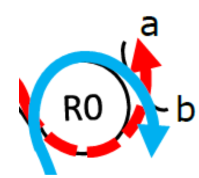
Investigated transmission scenarios

Transmissions on the same wavelength

1. Two co-propagating transmissions with increasing hop length
2. Multiple co-propagating transmissions
3. Two counter-propagating transmissions with shared-source ring
   – Counter-propagating intended and interfering data streams are transmitted from the same local ring, for increasing hop length



4. Two counter-propagating transmissions with shared-destination ring
   – Counter-propagating intended and interfering data streams are received at the same local ring for increasing hop length

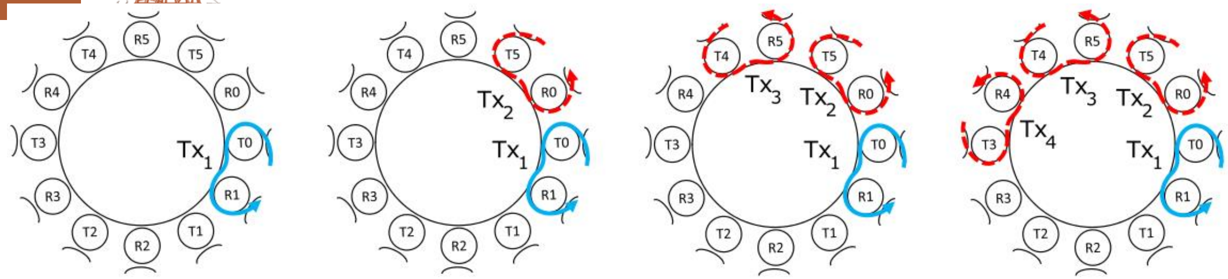


Spatial reuse

5. Transmissions on different wavelengths → **Wavelength reuse**

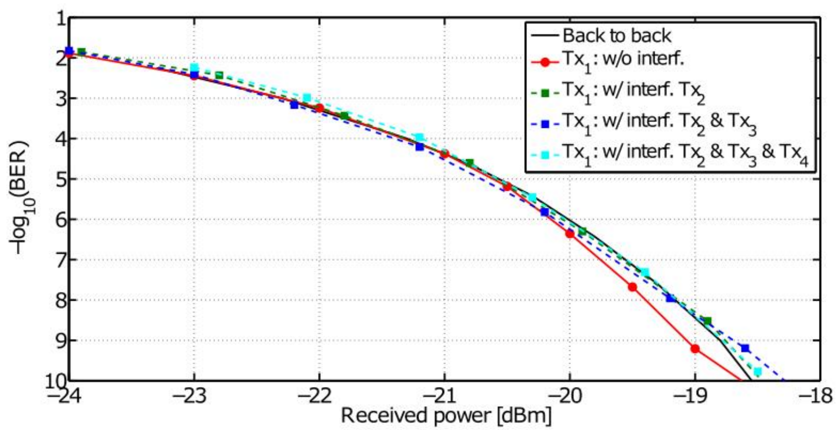
F. Gambini, P. Pintus, S. Faralli, M. Chiesa, G.B. Preve, I. Cerutti, and N. Andriolli, "Experimental demonstration of a 24-port packaged multi-microring network-on-chip in silicon photonic platform," *Opt. Express*, vol. 25, no. 18, Sep. 4, 2017, pp. 22004-22016.

2. Multiple co-propagating transmissions



a) b) c) d)

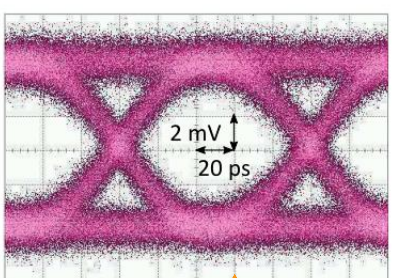
Intended transmission (blue) tested alone and with up to three upstream interfering transmissions (red) on the same wavelength



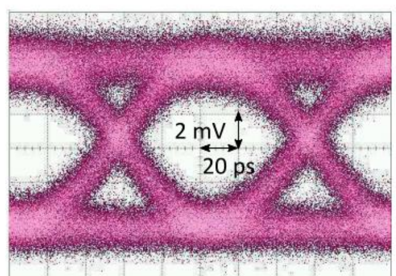
e)

No significant interference is caused by the upstream transmissions

- Power penalties less than 0.5 dB for a BER of 10^{-9}
- Eye diagrams still open



f)

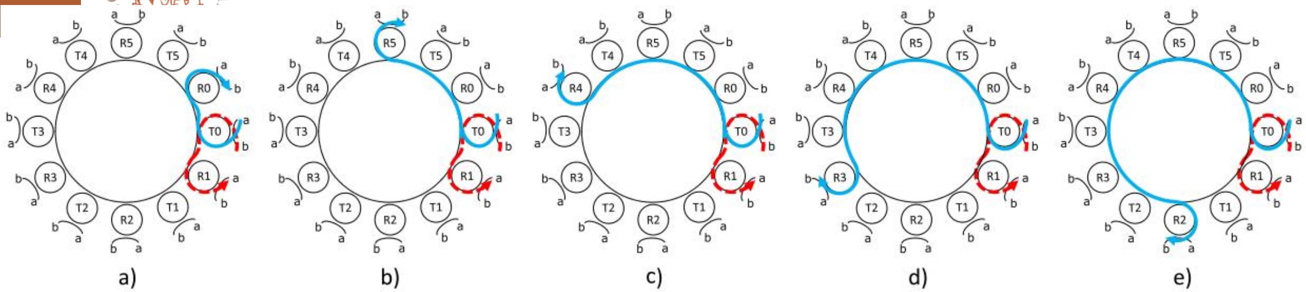


g)

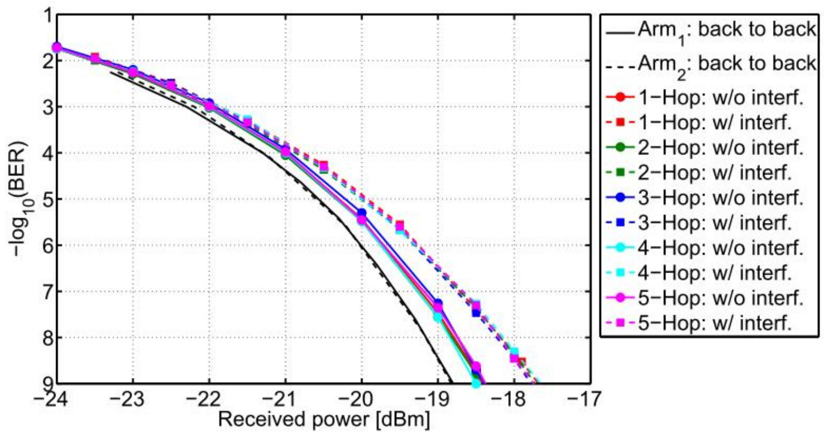
No interferer

3 interferers

3. Counter-propagating transmissions – Shared source ring

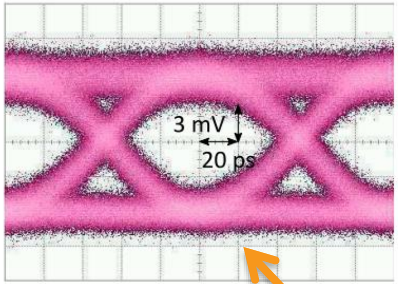


Intended transmission (blue) of different length, with the one-hop interfering transmission (red) at the same wavelength

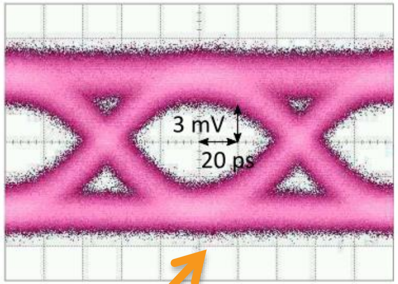


f)

- No correlation between the number of hops and the performance of the network
- The presence of the counter-propagating transmission causes a power penalty of about 0.8 dB at a BER of 10^{-9}
- No eye diagram degradation when increasing the traversed hops



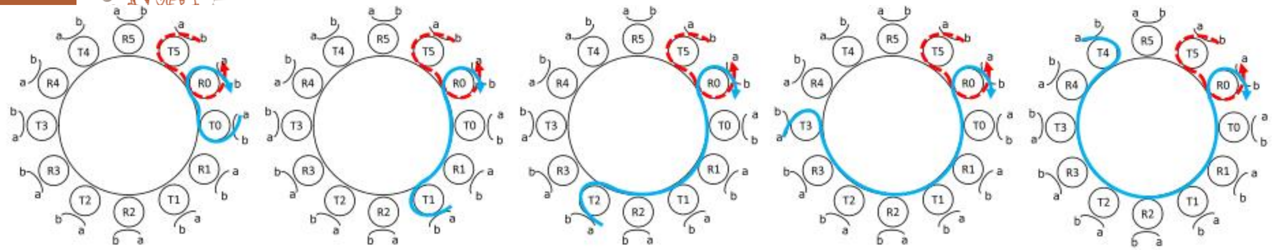
g)



h)

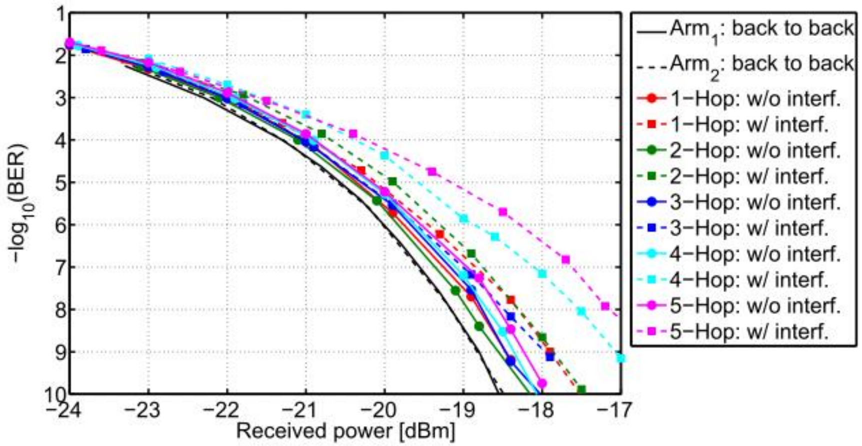
Eye-diagrams for 1-hop and 5-hop transmissions in presence of the interfering transmission

4. Counter-propagating transmissions – Shared destination ring



a) b) c) d) e)

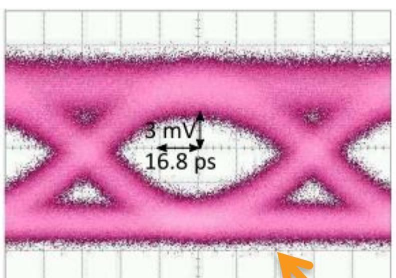
Intended transmission (blue) of different length, with the one-hop interfering transmission (red) at the same wavelength



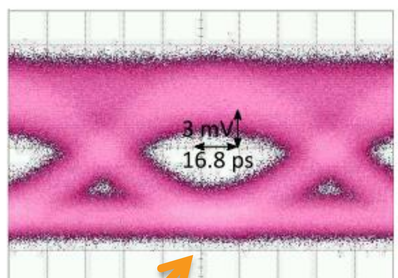
f)

In this scenario the performance depends on the path length → Slightly larger degradation on longer paths (maximum power penalty of about 2 dB)

– Due to the loss along the path of the intended transmission



g)



h)

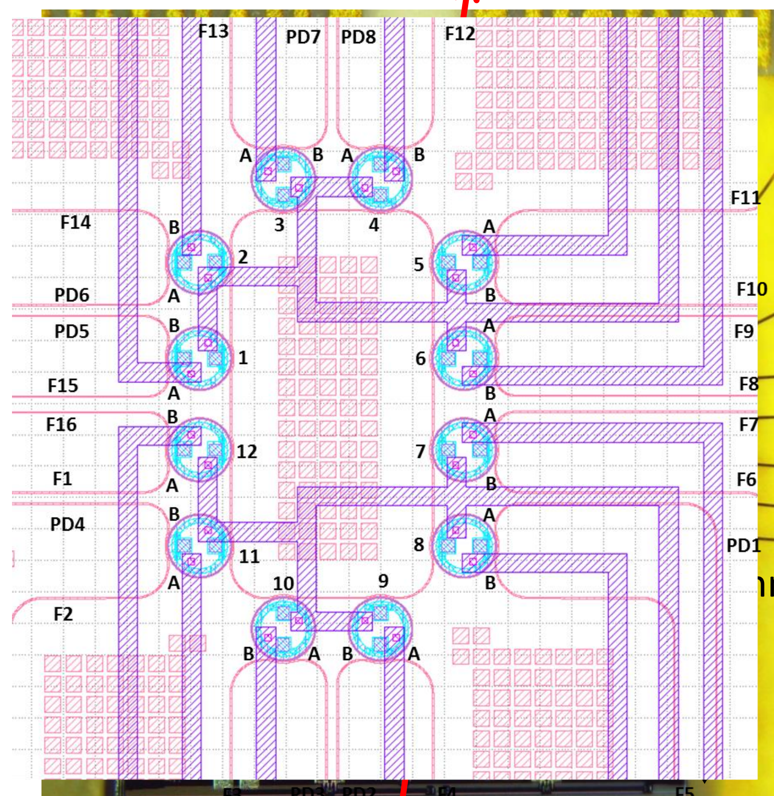
Eye-diagrams for 1-hop and 5-hop transmissions in presence of the interfering transmission →Longer transmission shows a worse eye

Silicon-on-Insulator based photonic integrated circuit

Packaged MMR interconnection network:

- 12 local rings for a total of 24 optical IOs
  - 8 ring IOs to monitoring PDs
  - 16 ring IOs to fiber array
- 26 electrical pads (signal + ground)
  - 12+2 for ring control heaters
  - 8+4 for monitoring PDs
- Die wirebonded to ceramic package and mounted on a custom PCB
- Controlled through GPIO
- 16-fiber array vertically mounted
- Temperature control through Peltier cell with heat sink underneath the package

MMR



Array of 16 grating couplers

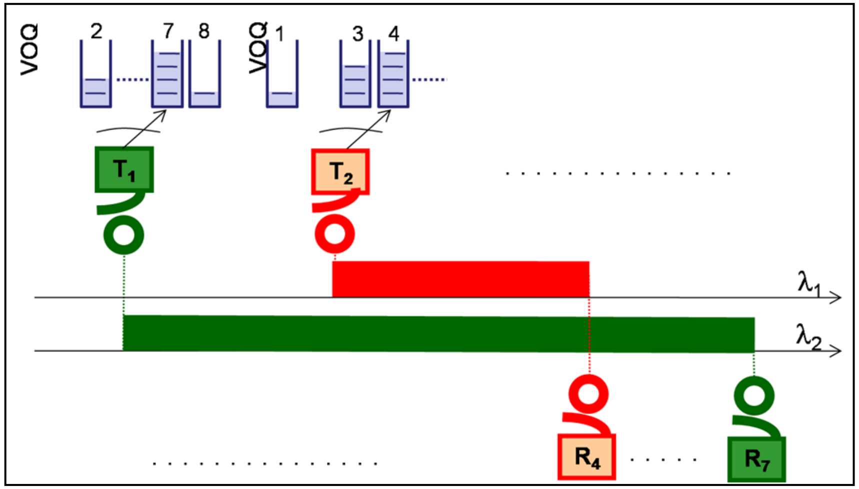
- Manufactured on a Silicon-on-Insulator platform
  - Multiproject wafer run by IME through CMC Microsystems
  - Packaging at INPHOTEC Center of SSSA, Pisa
- Thermal tuning is used to control the resonance of the local rings

Scheduler

Slotted system:
- Fixed-size packets
- Transmission synchronized at time slots

At each input port, an electronic buffer stores the incoming data
- Organized into virtual output queues (**VOQ**) according to the destination

Example of packet scheduling and wavelength occupancy during a time slot



Scheduling decisions must be taken:
- **At each time slot**, based on the information on VOQ occupancy
- **Centrally**, as wavelength usage information are required

Scheduling differs depending on the MMR type:
- **fixed**: Fixed lasers connected to transmitters and tunable receivers
- **tunable**: Tunable lasers connected to transmitters and tunable receivers