High Performance Optical DCN based on WDM Optical Cross-Connect Switches

Nicola Calabretta

Eindhoven University of Technology

n.calabretta@tue.nl

TU/e Technische Universiteit Eindhoven University of Technology

Lightness

COSIGN

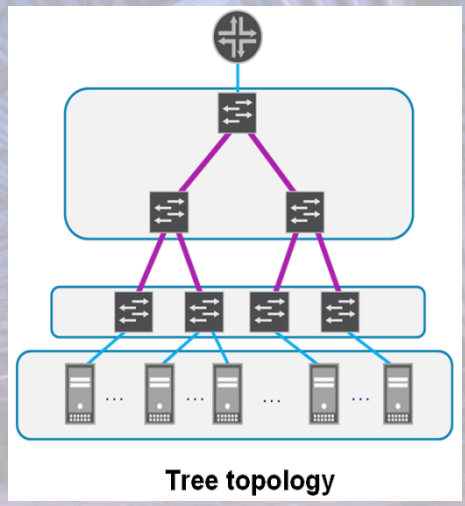
Materials Devices Systems

Outline

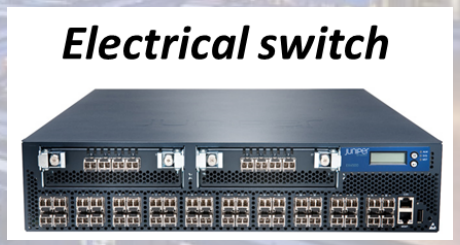
- **Scalable DCN architecture: bandwidth, latency, power consumption issues**
- **OPSquare DCN architecture based on distributed flow-controlled WDM cross-connect switches**
- **Photonic integrated cross-connect switch**
- **Conclusions**

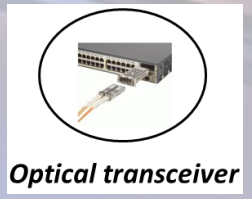
Wha the is ? Data Center Network (DCN)

- **Network archite**
  - *Bandwidth bottlenec*
  - *Large latency*
  - *Static*



- **Electrical switch**
  - *Limited bandwidth*
  - *Large latency*
  - *High cost and power*

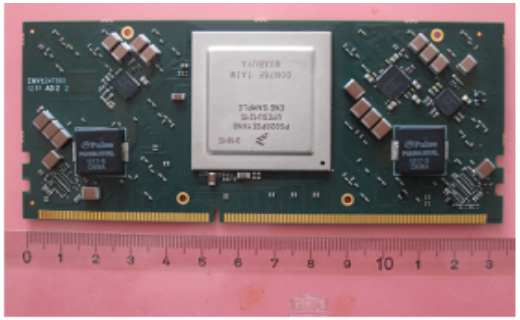




Mor

Scaling to Petabit/s interconnect networks

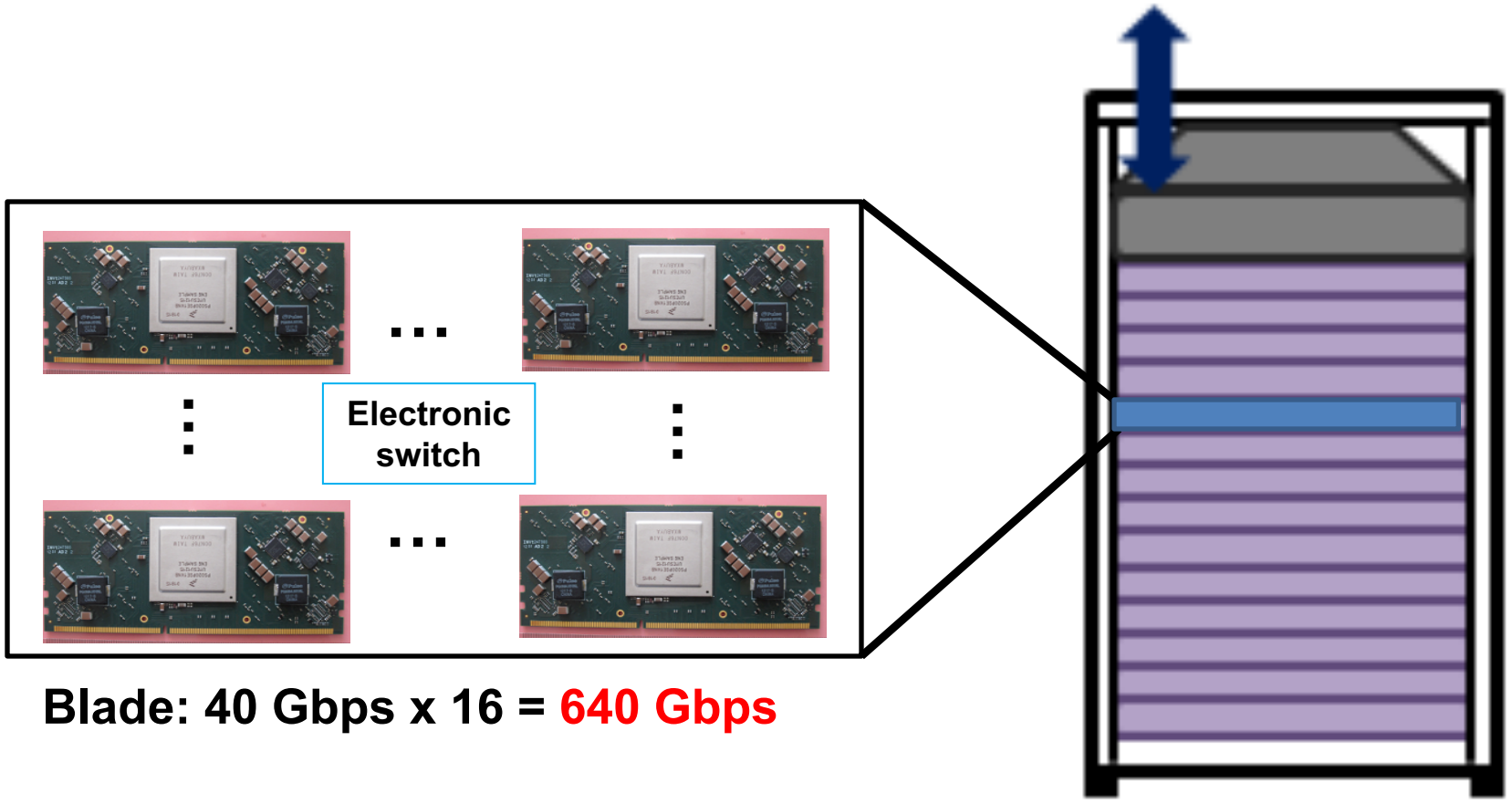
X86 Motherboard **30 cm x25 cm**
40 Gbps Ethernet



IBM Microserver **14 cm x 5.5 cm**
40 Gbps Ethernet

AcQ Microserver **10 cm x 15 cm**
40 Gbps Ethernet
24 Gbps PCIe3

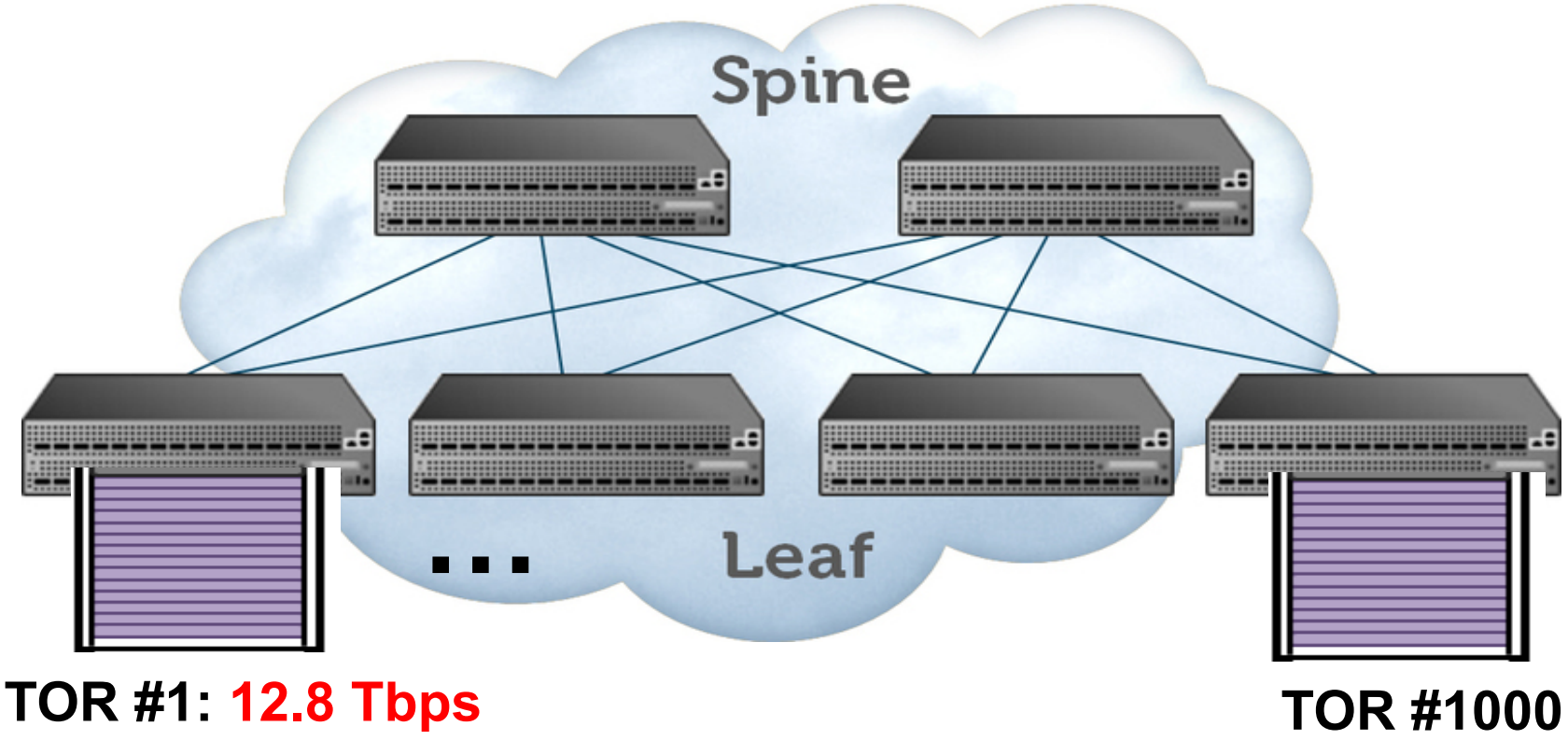
Scaling to Petabit/s interconnect networks



Blade: 40 Gbps x 16 = 640 Gbps

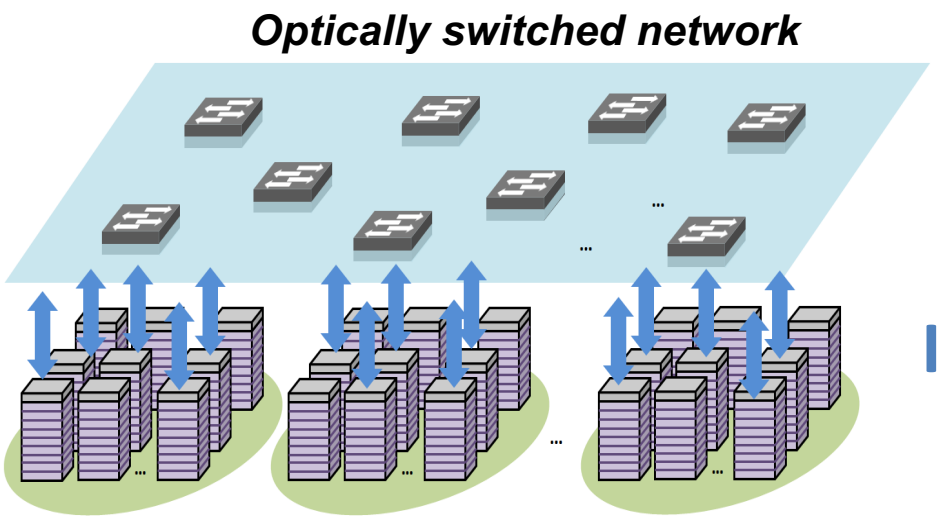
TOR: 20 x 640 Gbps = 12.8 Tbps

Scaling to Petabit/s interconnect networks



Scalability and capacity limited by the switch radix and port bandwidth

High capacity DCN based on fast controlled optical switches



One flat network

Overcome the bandwidth bottleneck

Low latency

Flat any-to-any connectivity

Introducing optical switching

Transparency to data rate/format

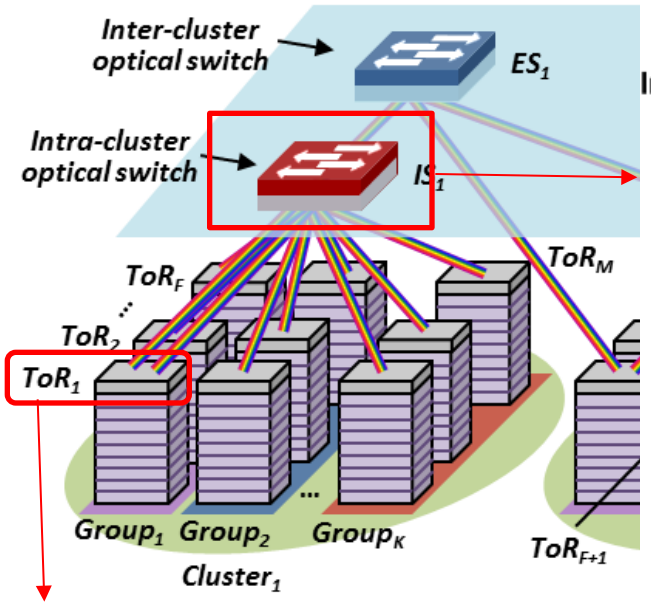
Elimination of O/E/O conversions

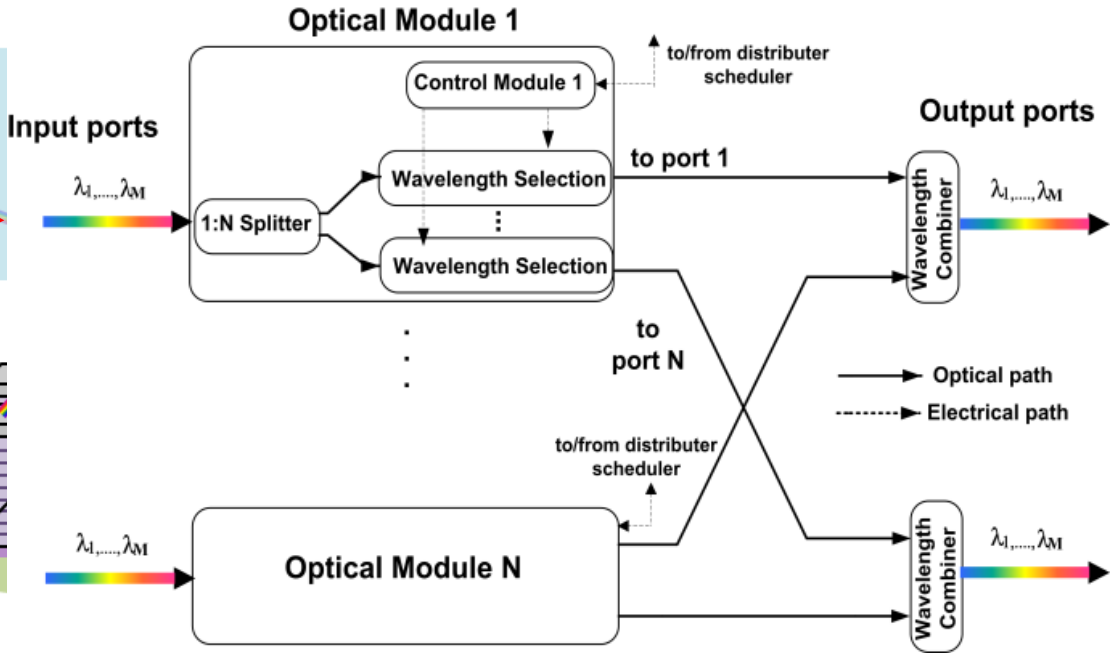
Fast control ?

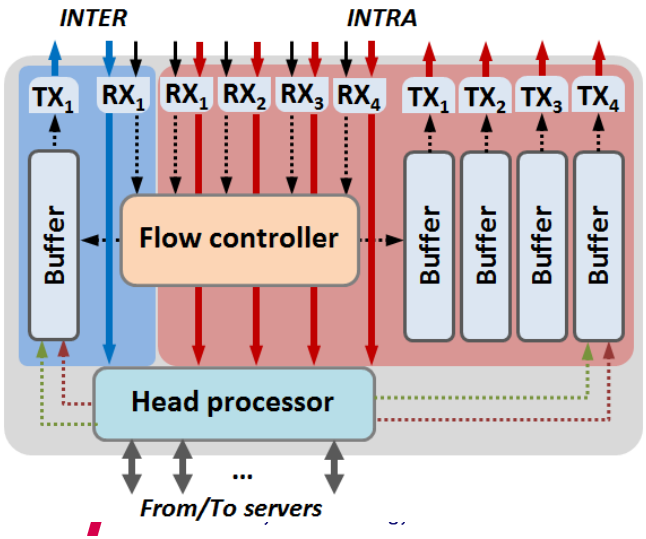
Buffering issue?

Connectivity?

OPSquare DCN architecture







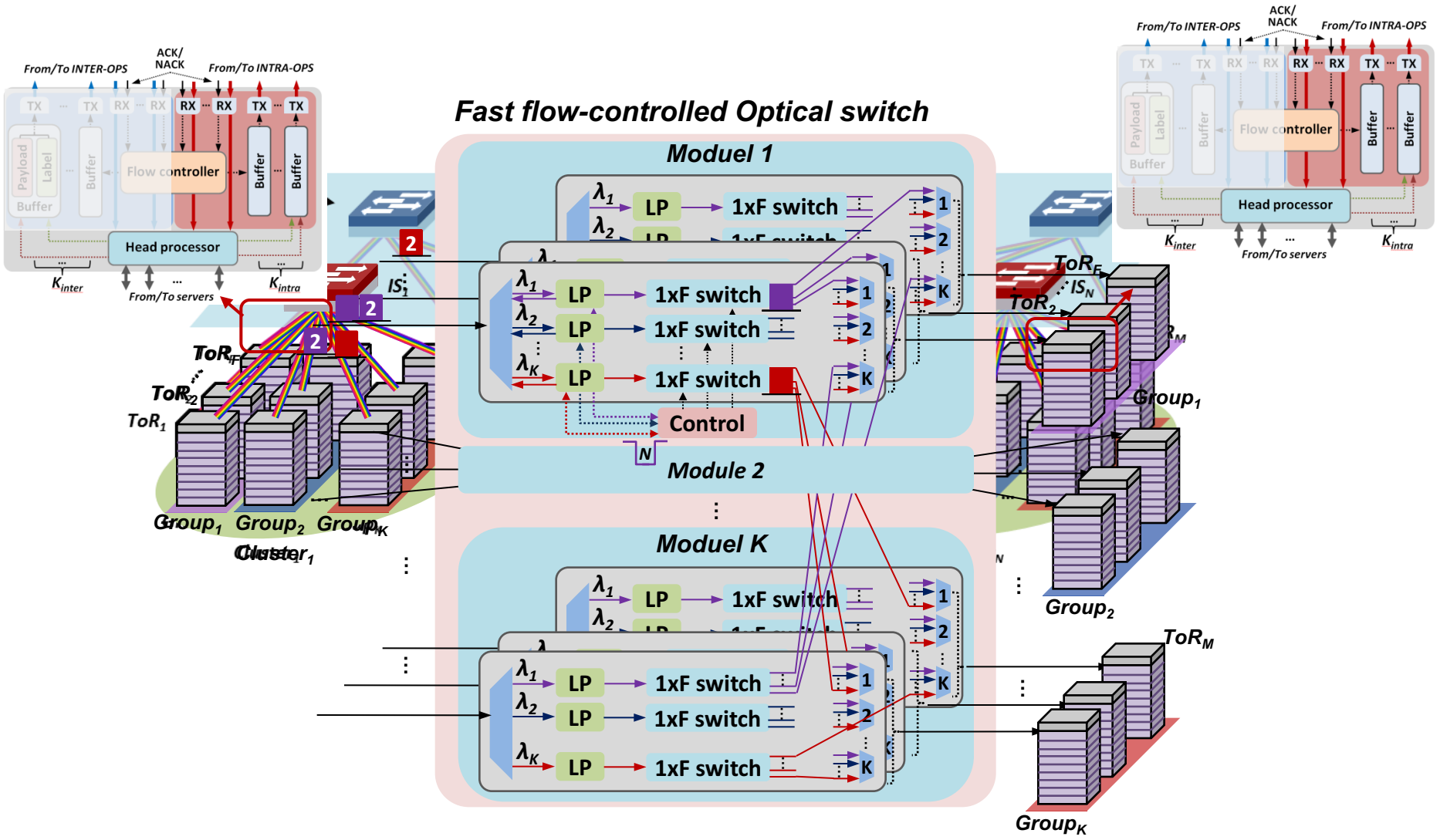
Optical label processing → Nanoseconds switch control

Optical flow control → Buffer-less operation

Nanoseconds reconfiguration → statistical multiplexing

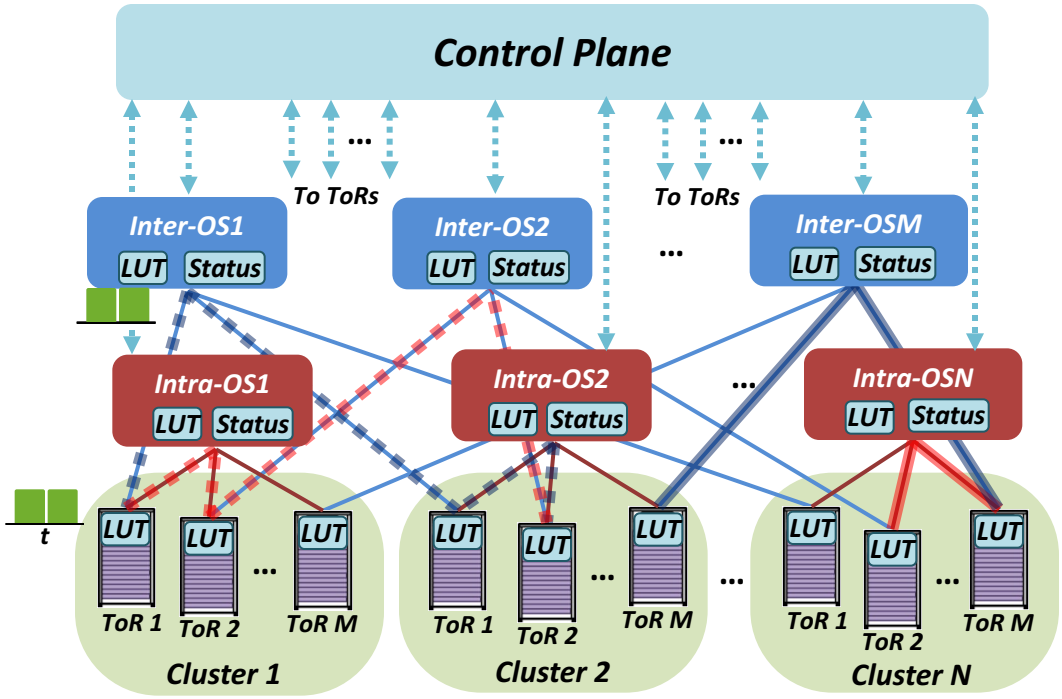
Single stage architecture → Scalable

Fast controlled optical switch



- *Fast parallel processing of the label*
- *On-the-fly distributed control*
- *KxF connectivity achieved by 1xF switches*
- *Fast flow control and retransmission*

Dynamic virtual DCN reconfiguration



LUT – look-up table

1. ***Network provisioning***

   - Milliseconds operation

2. ***Fast switching***

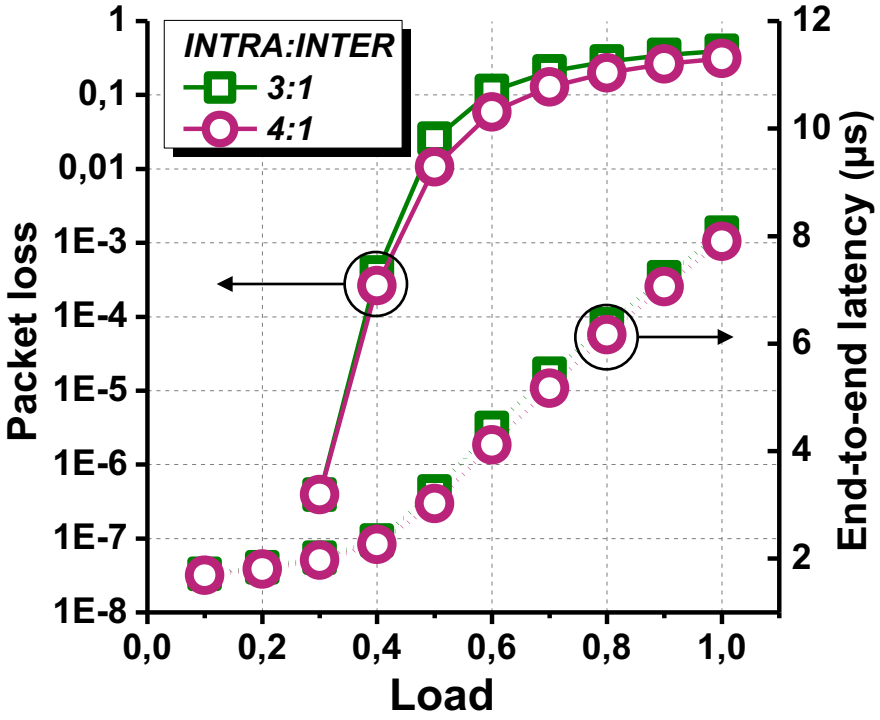
   - nanoseconds operation

Decoupling of control plane (ms) and data plane (ns) operation

Numerical performance (I)

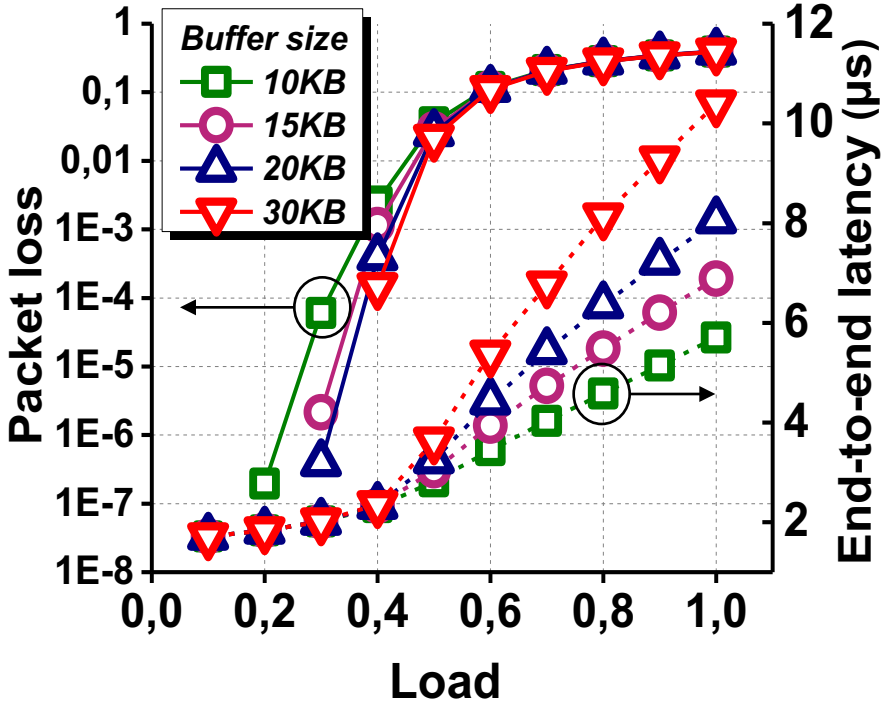
Different intra-/inter-cluster traffic ratio

- Buffer size is 20KB per transceiver



- ***Packet loss <1E-6***
- ***End-to-end latency <2µs***

- ***Buffer dimensioning***



Larger buffer improve packet loss but for load > 0.5 latency performance increases

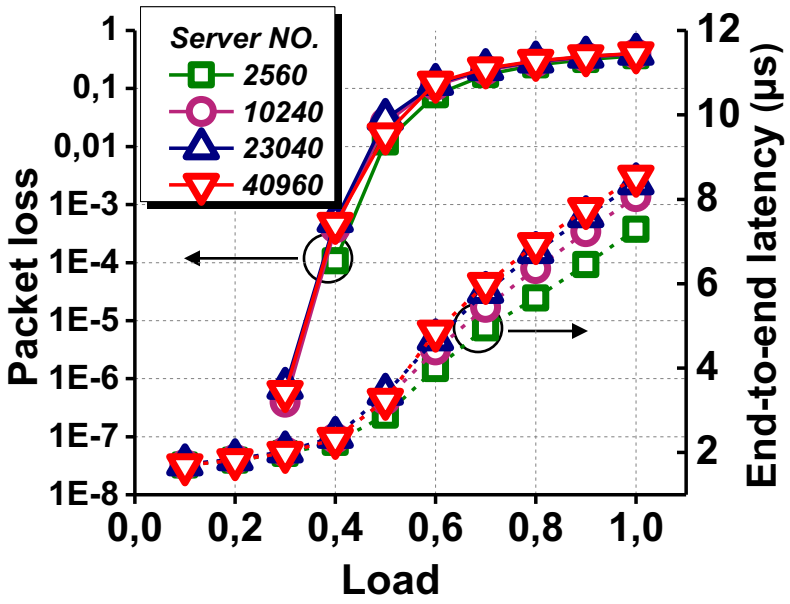
Numerical performance (II)

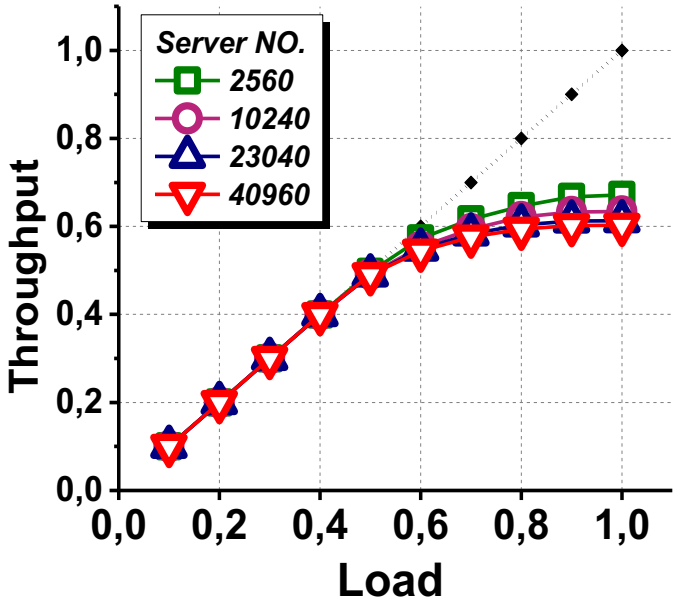
Scaling to larger number of servers:

- Network size varies from 2560 to 40960 servers

- Optical switches have radix of 8x8 to 32x32

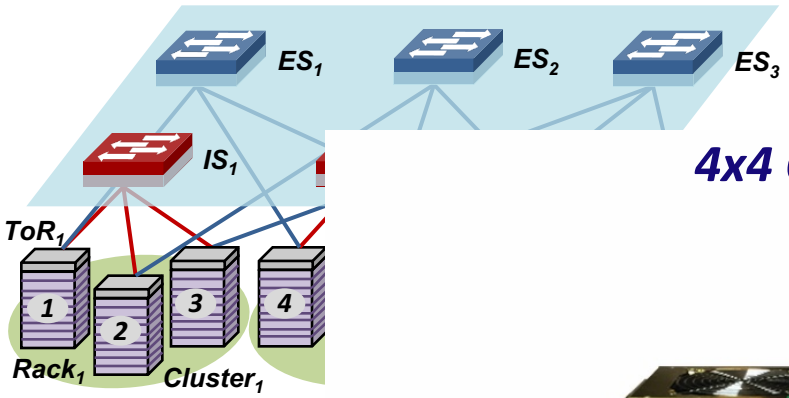
- 20KB buffer per transceiver



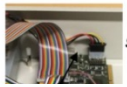


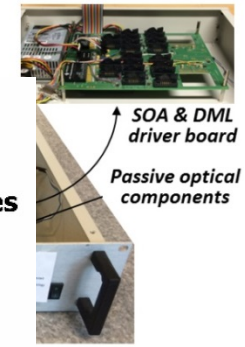
- ***Larger scales perform similarly with limited degradation***
- ***Network saturates for load higher than 0.6***

Experimental investigation

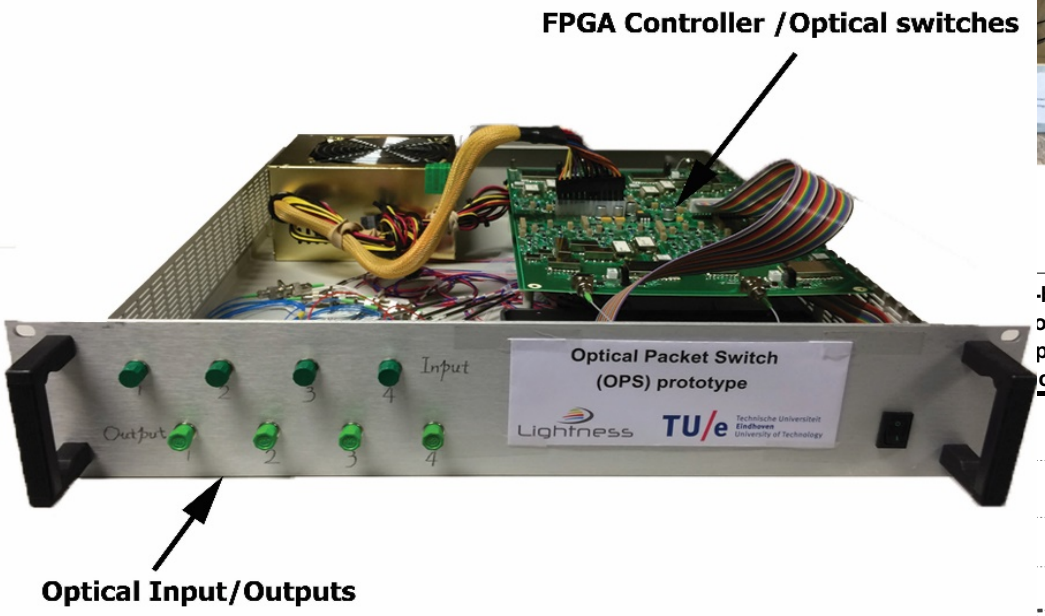


Optical switch prototype

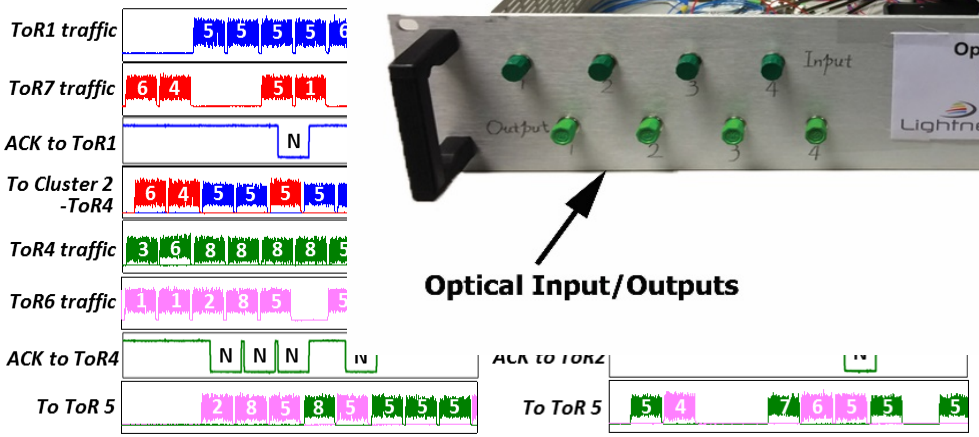


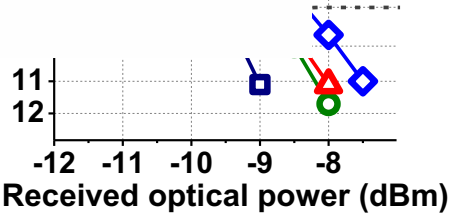


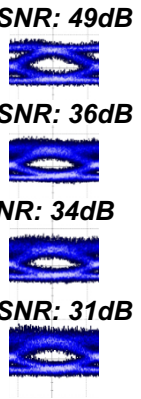
4x4 OPS prototype



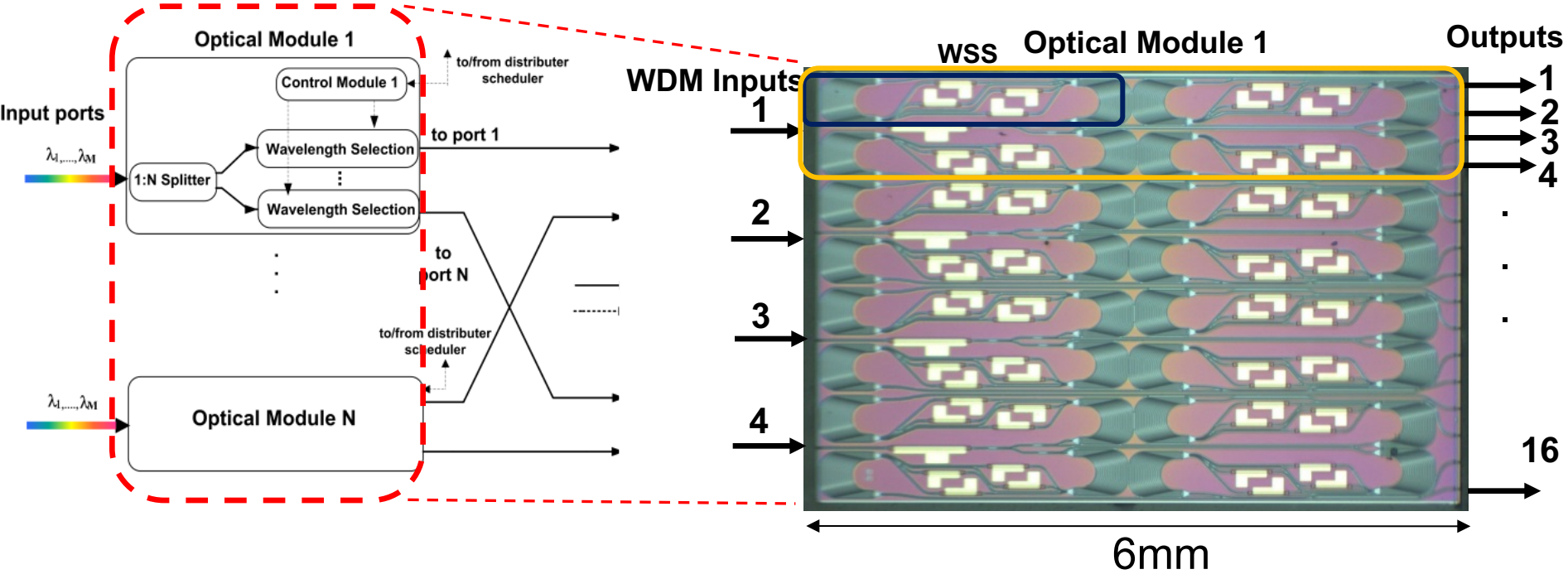
Path







Towards Photonic Integration



Multi-level modulated traffic

Large input power dynamic range

Small footprint

Low power consumption

High bandwidth density

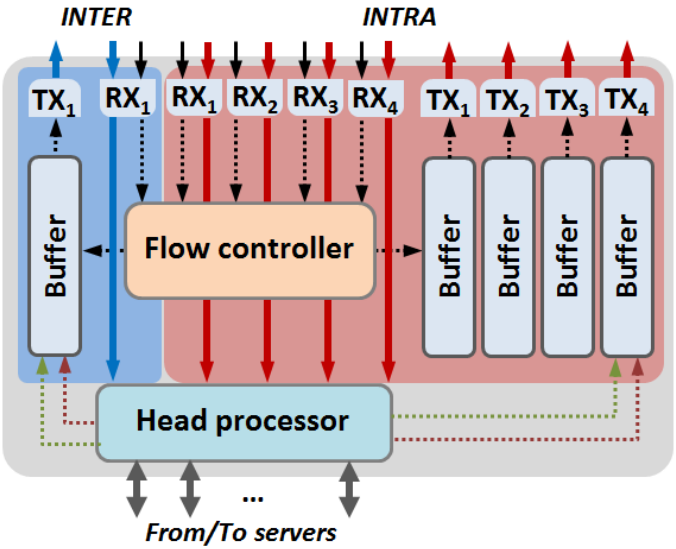
N. Calabretta et al, ECOC 2016 and OFC 2017

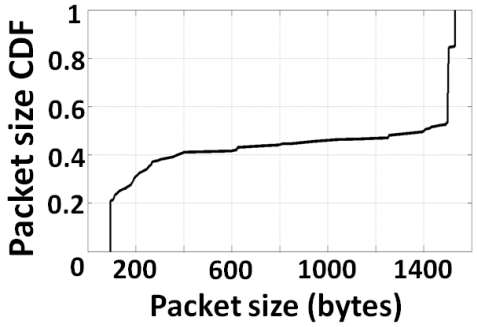
Conclusions

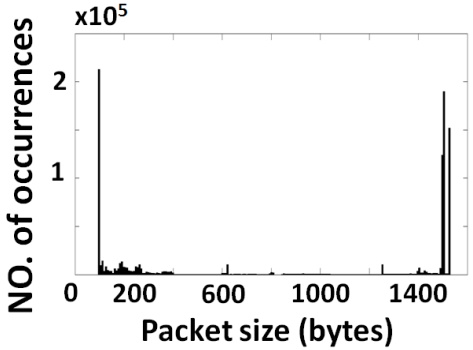
- OPSquare scalable parallel flat intra-/inter-cluster data center networks architecture based on distributed buffer-less optical
- Fast flow control and label processing enable nanoseconds control of the optical switches → statistical multiplexing
- Optical switch transparency → data rate and format independent
- WDM TRXs at the TOR → improve DCN capacity and the feasibility of the optical switch (lower B&S splitting losses)
- Assessments with realistic traffic in OMNeT++
  - Packet loss <1E-6 and latency <2μs at 0.5 load
  - up to 40960 servers with limited degradation
- Photonic integration enables large port count WDM cross-connect switches with reduced power and costs

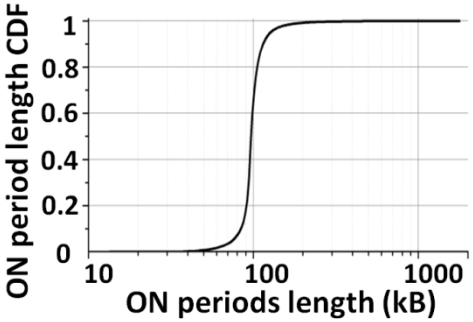
Acknowledgements

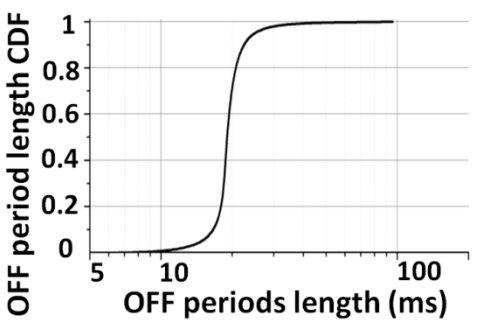
Numerical performance investigation











- *40-server (10Gb/s) rack*
- *Data center-like traffic pattern**
- *K_{intra}=4 and K_{inter}=1*
- *Transceivers operating at 50Gb/s*
- *560ns round trip time*

*T. Benson et al., “Network Traffic Characteristics of Data centers in the Wild,” Proc. ACM, 2010.