NOKIA Bell Labs

“VERNE”: New Packet-Optical Network for Optically Transparent and Lossless Data Centers

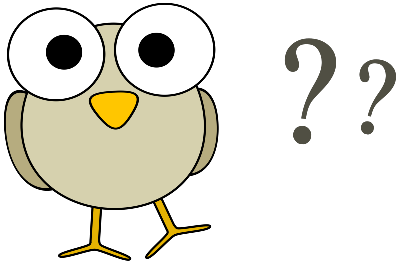
ONDM 2017

- Bogdan USCUMLIC, Dominique CHIARONI
- Nokia Bell Labs, Paris Saclay, France

- 17-05-2017

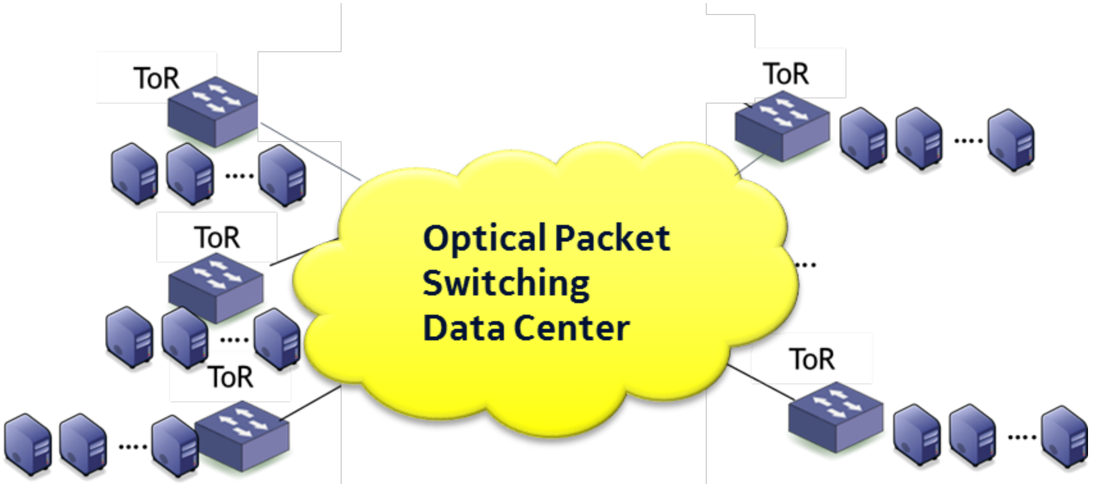
- Ethernet is the key technology inside of Data Centers

- Can an OPS network provide a cost efficient and a performant alternative?



Key requirements for data centers

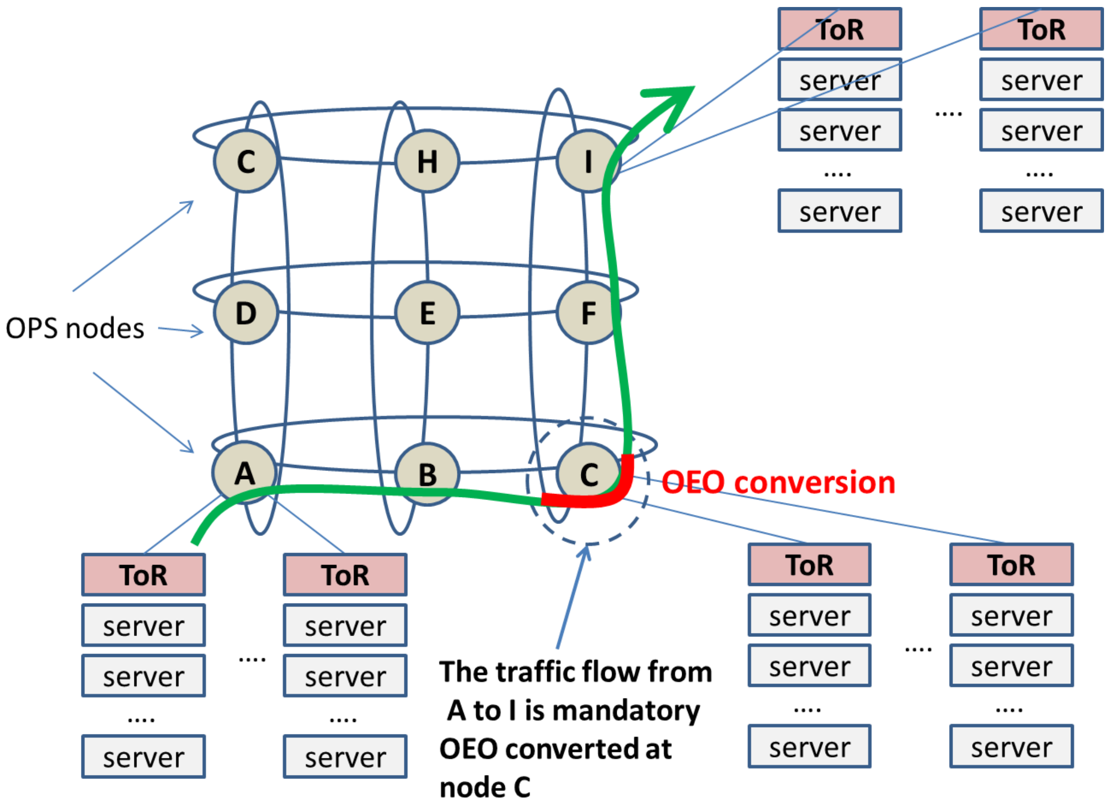
- Data Centers requirements:
  - Cost
  - Scalability
  - Latency
- Optical Packet Switching benefits:
  - Traffic grooming & statistical multiplexing in the optical domain
  - Optical transparency of transit traffic
- Question:
  <u>What are the challenges for OPS networks in data centers???</u>



NOKIA Bell Labs

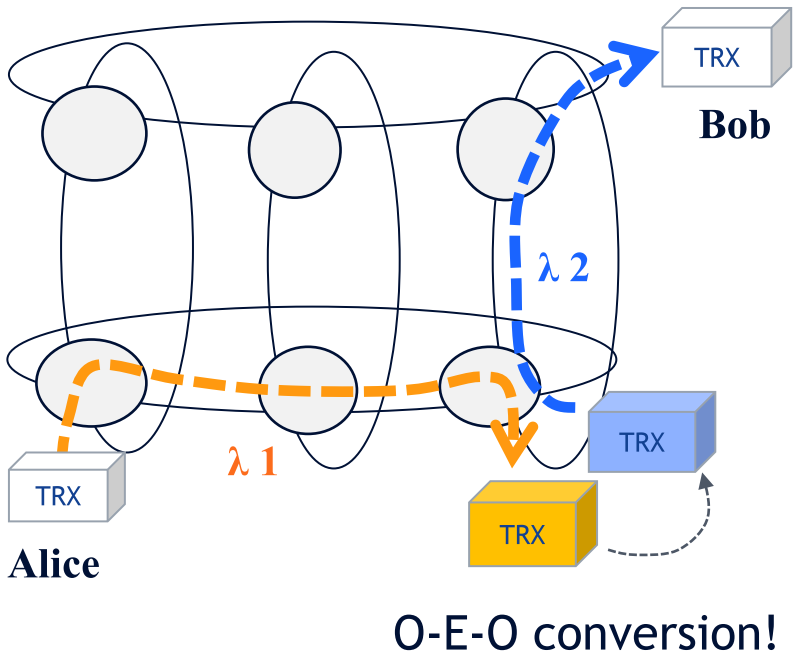
Problem of OEO conversion

- **2D torus** optical packet switching network
- **OEO conversion not completely removed!**



OPS 2D torus: OEO conversion bottleneck!

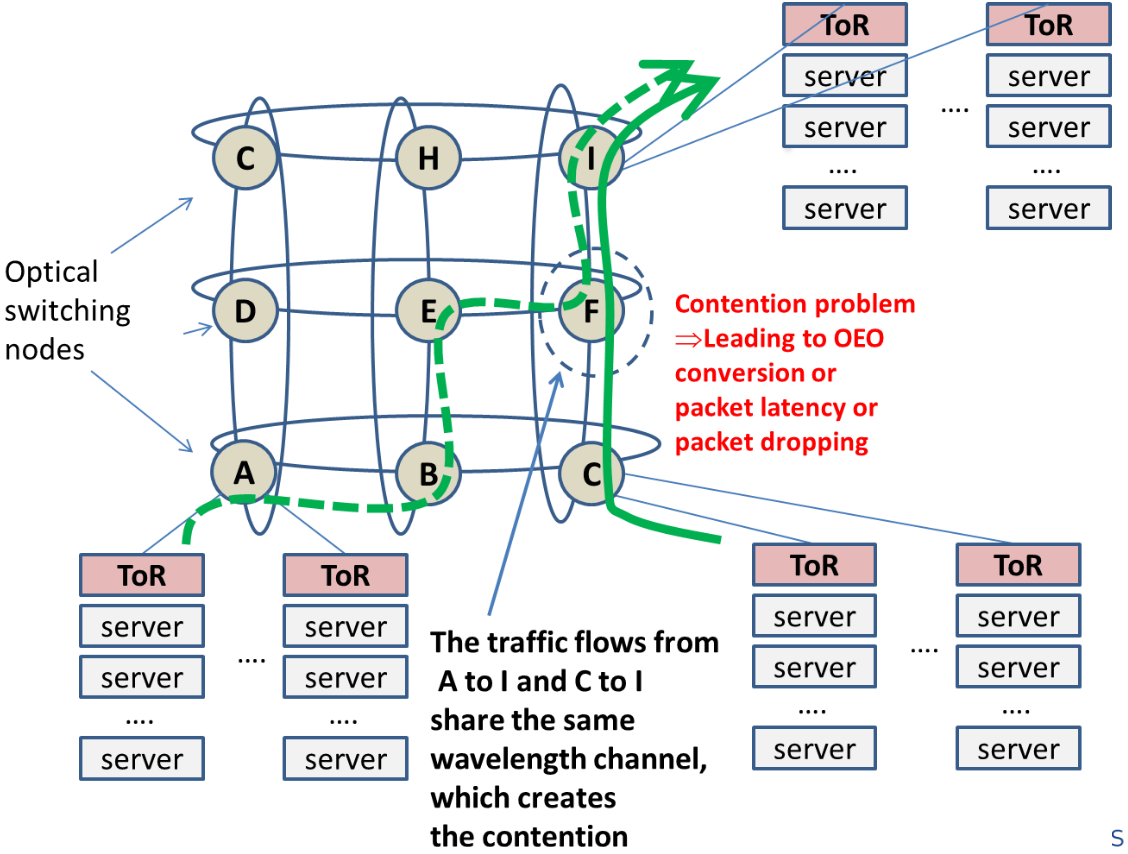
- It would be good to **avoid the OEO** conversion!



Problem of contention

- E.g. when sharing the wavelength between several flows

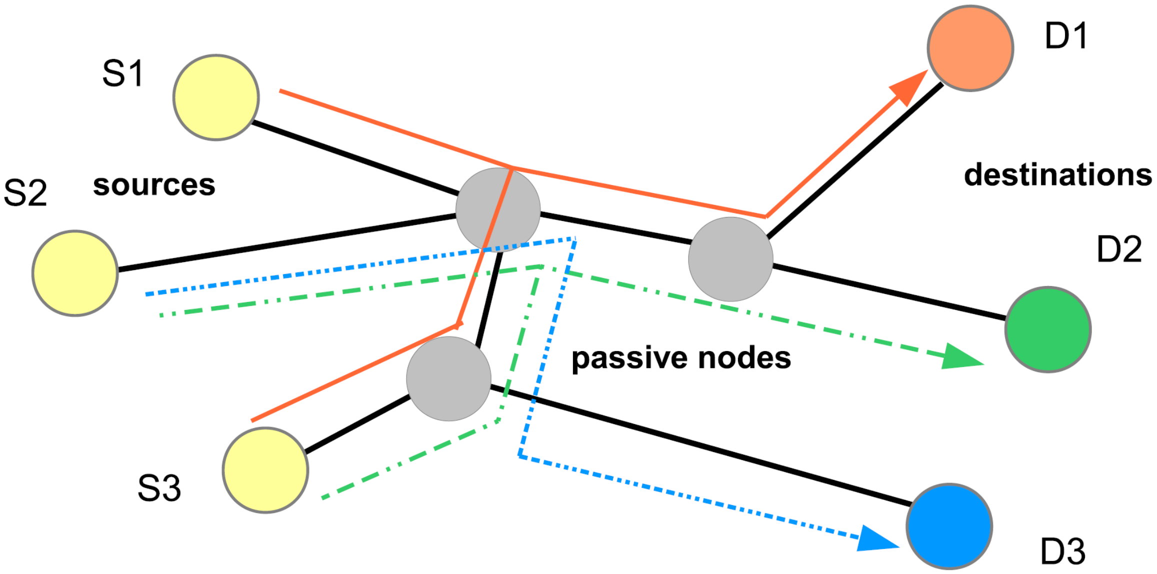
- Contention problem leads to:
  - **Packet rerouting**
  - **Packet delay**
  - **Packet loss!**



Problem of scalability

- TWIN network: Time-Domain Wavelength Interleaved Network

- The **scalability** issue due to wavelength addressing



NOKIA Bell Labs

Technical challenge

- Build an OPS network that is:
  - **Optically transparent**
  - **Lossless (contention free)**
  - **Scalable**

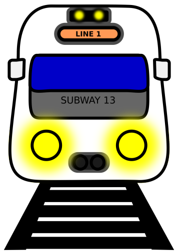
- Is this possible?
  - YES!

NOKIA Bell Labs

VERNE network

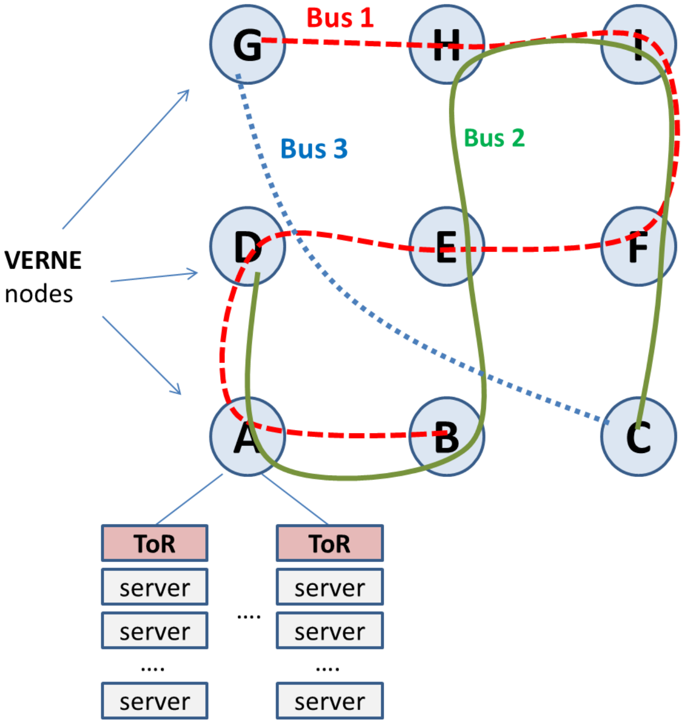
- VERNE = "Virtual, fully transparent, cost and enERgy efficient NEtwork"

- VERNE network analogy in real life:
  System of metro lines, without the intermediate stop



VERNE network architecture

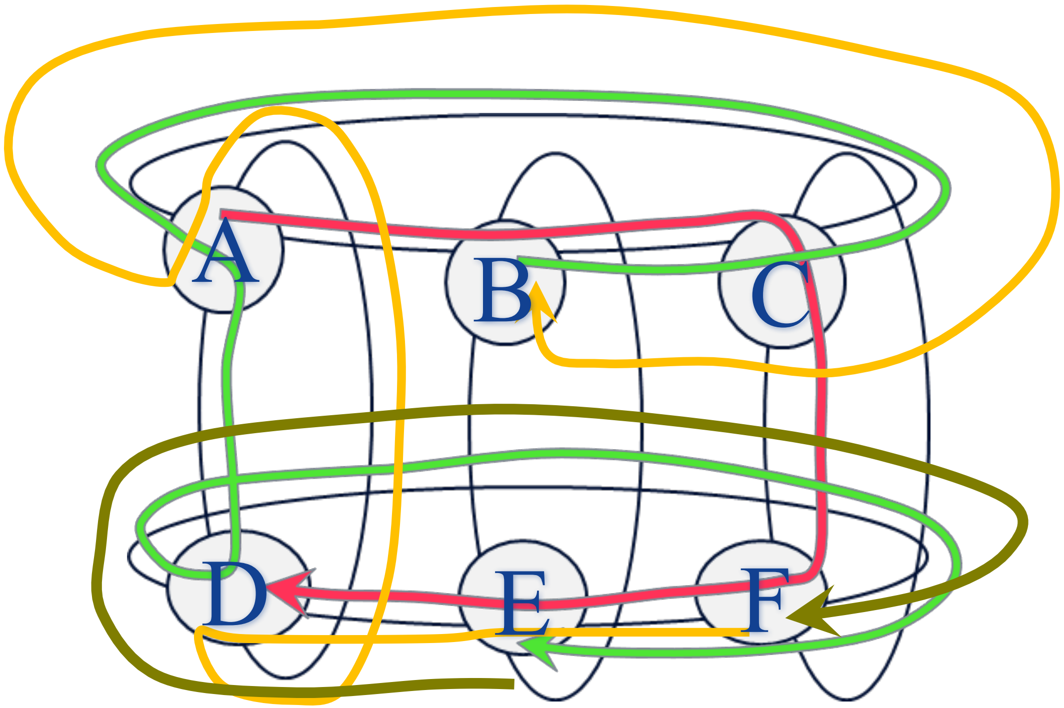
- VERNE, a network that reaches **<u>the lower bound on TRX</u>**
- Network is covered by a number of **<u>"optical buses/rings"</u>**, that are disjoint sets of wavelengths /fibers
- Any destination can be reached via a single optical hop!
- **<u>Time slotted, control channel attributed to each optical bus/ring</u>**
- No O-E-O conversion on the intermediate nodes!
- No intermediate queueing => lower latency!



The example of VERNE network interconnection, when reach limit is set to 7 nodes

VERNE network example

Interconnecting the network with the virtual optical buses



A is connected to everybody

B, C are connected to everybody

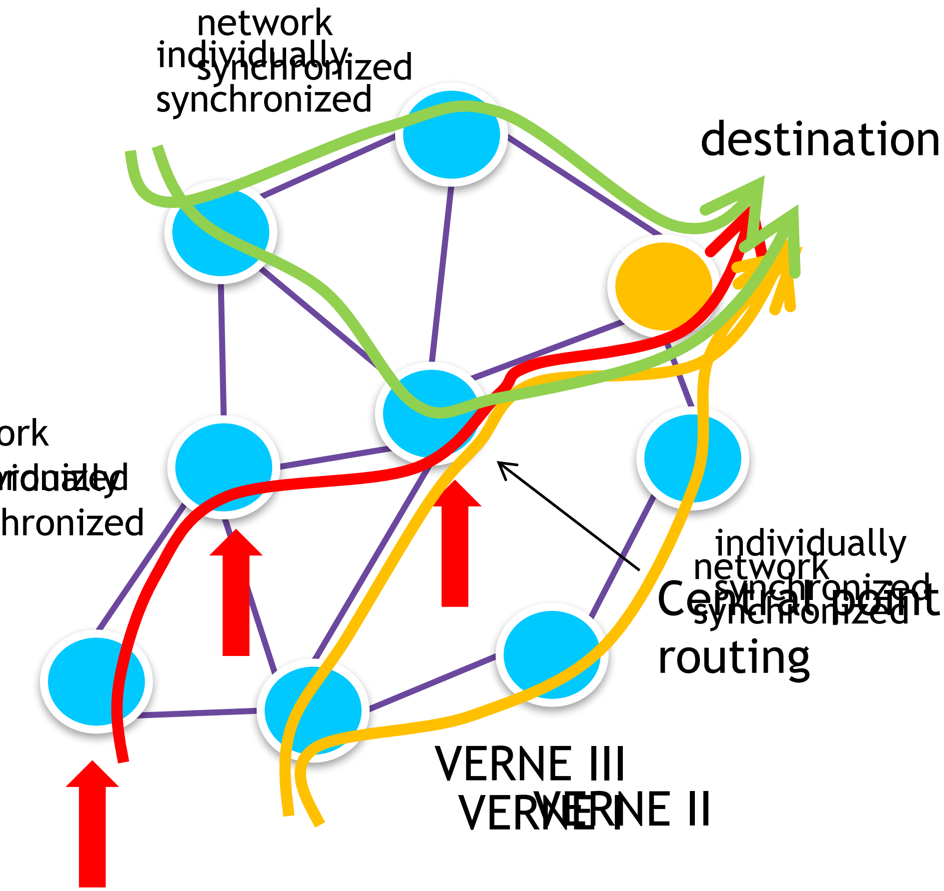
D, F are connected to everybody

E is connected to everybody

Different variants of VERNE network

- From point of view of the synchronization:
  - VERNE I or "synchronous"
  - VERNE II or "fully synchronous«
  - VERNE III or "asynchronous«

- From the point of view of the scheduling:
  - VERNE I: Scheduling per bus
  - VERNE II: Centralized or opportunistic scheduling
  - VERNE III: Centralized scheduling with central routing point



NOKIA Bell Labs

Comparison of different VERNE variants

VERNE III
asychronous

VERNE I
Per bus

VERNE II
Entire network

low

high

Synchronization complexity

VERNE I
Per bus

VERNE II
centralized or
opportunistic

VERNE III
Centralized with central
routing point

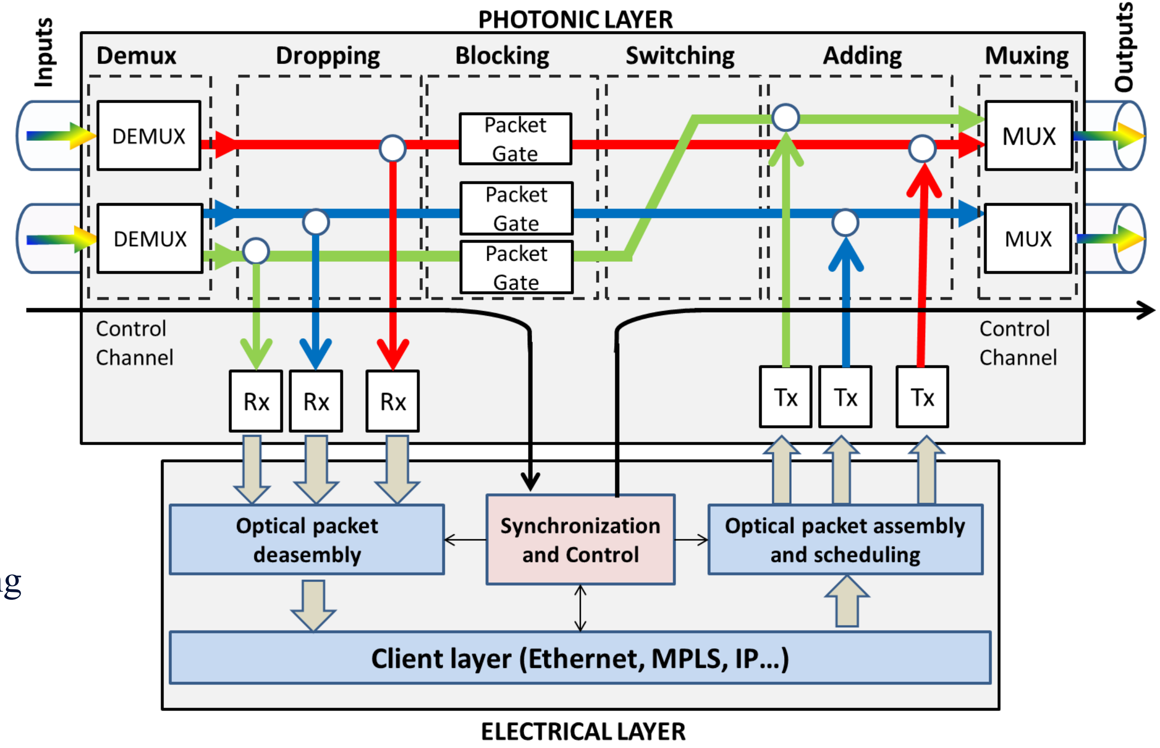
low

high

Scheduling complexity

Architecture of the VERNE node

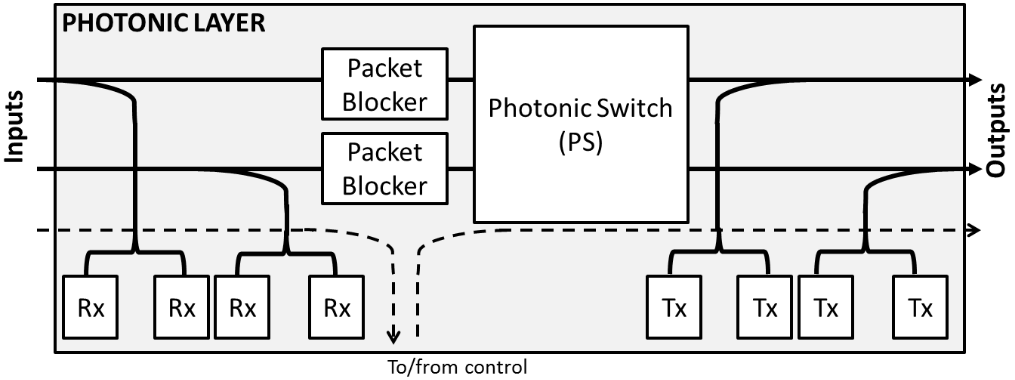
- Node composed of:
  - Photonic layer
  - Electrical layer
- Basic blocks:
  - Demultiplexing & Dropping
  - Blocking
  - Switching
  - Adding & Multiplexing
  - Optical packet deasembly
  - Synchronization and control
  - Optical packet assembly and scheduling
  - Client adaptation layer



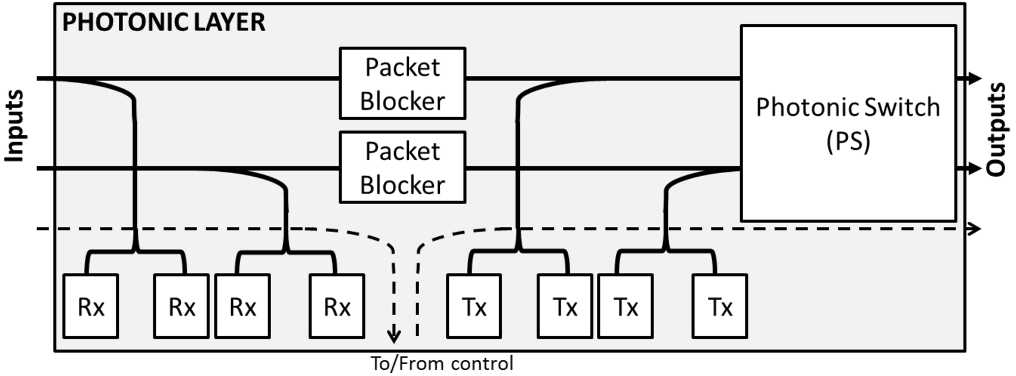
NOKIA Bell Labs

Photonic layer realization of the VERNE node

- Dropping by optical splitters
- Demultiplexing, packet blocking and multiplexing by packet blockers
- Switching by Photonic Switch
- Adding by couplers



- Different configurations in term of Photonic Switch position
- Control channel for carrying the OAM, synchronization and scheduling related information



NOKIA Bell Labs

Numerical results

- **Evalute savings in: #TRX, TRX cost, latency, scalability**
- Benchmark for the study:
  - Ethernet Fat Tree
  - Ethernet 2D torus
- Simple dimensioning algorithms
- TRX: 100 Gbit/s

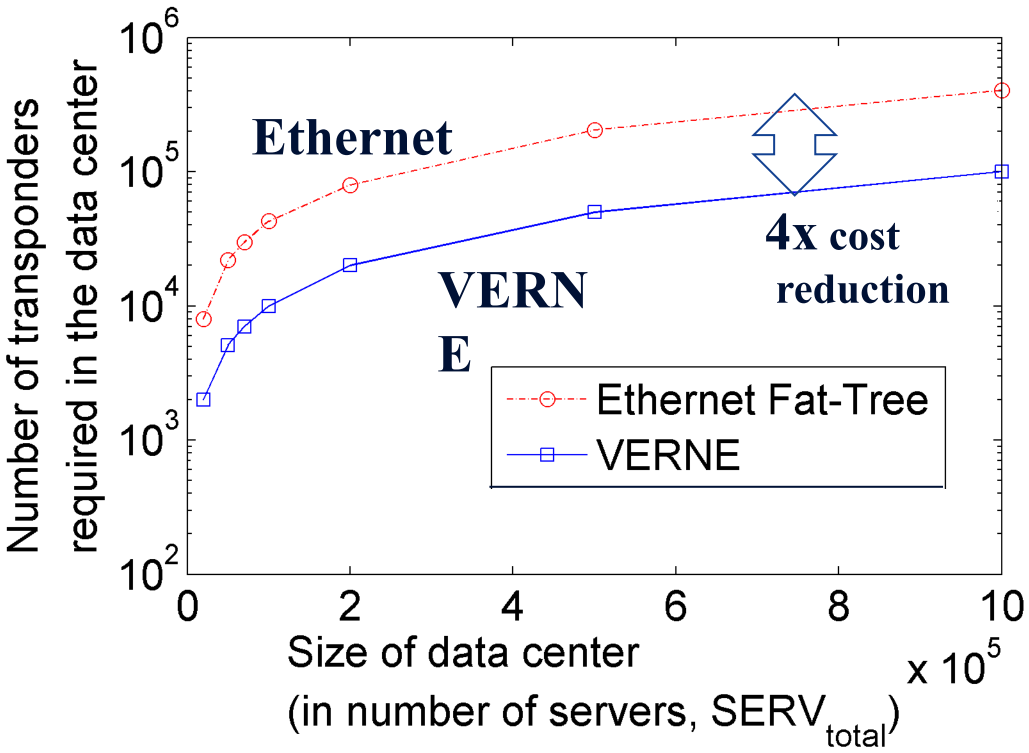
NOKIA Bell Labs

VERNE vs Ethernet Fat-Tree

TRX cost comparison

- Scenario 1:
  - VERNE vs Ethernet Fat-Tree data center
  - No oversubscription
  - In VERNE, TRXs allocated per optical buses (no impact of scheduling or synchronization)

- **<u>VERNE saves TRX cost up to x4 times</u>**

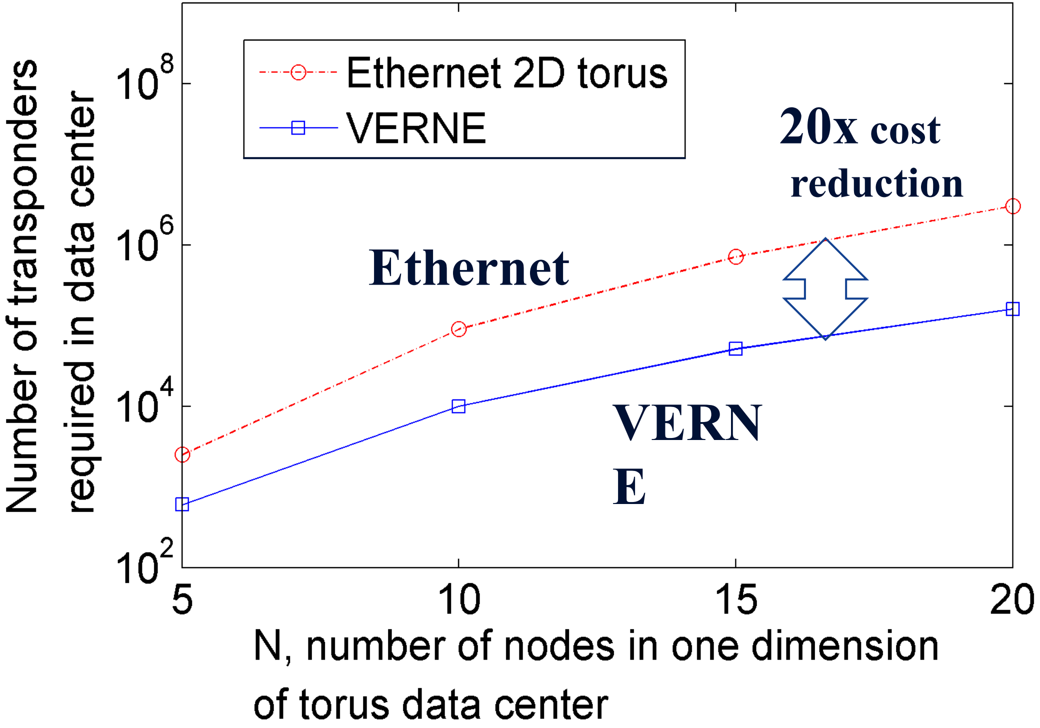


VERNE vs Ethernet 2D torus

TRX cost comparison

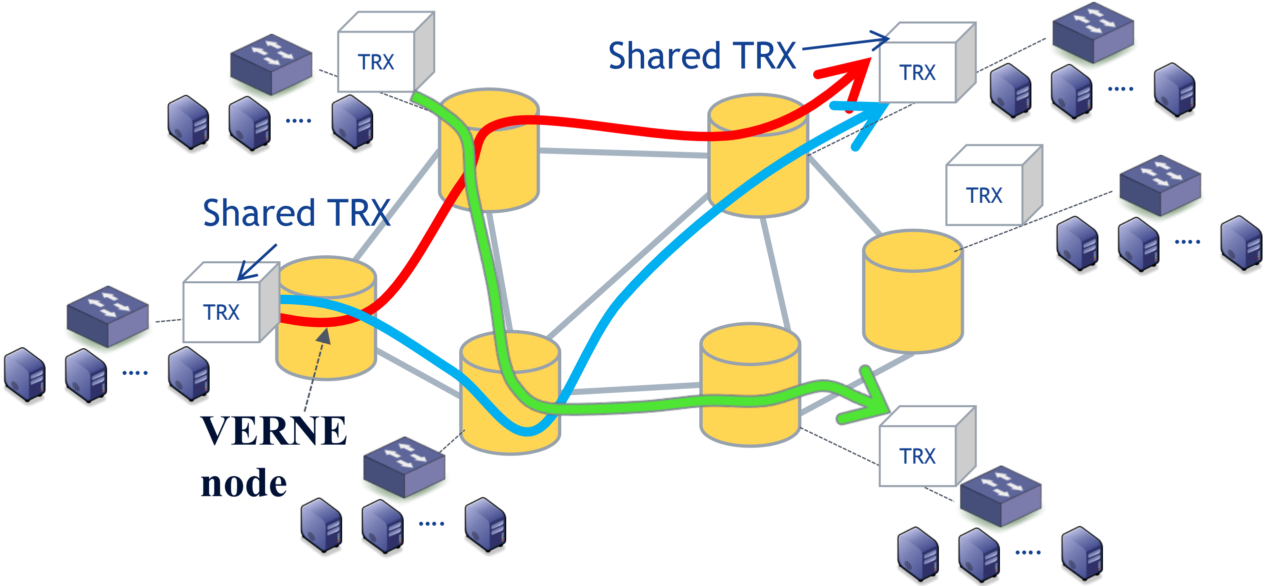
- Scenario 2:
  - VERNE vs Ethernet 2D torus data center
  - Symmetric torus (of dimension N)
  - Shortest path routing for Ethernet
  - In VERNE, TRXs allocated per optical buses (no impact of scheduling or synchronization)

- **VERNE achieves significant savings in TRX cost (up to 20 times)**



NOKIA Bell Labs

Ease of the network virtualization



- VERNE designed for **network virtualization and reconfigurability**

- Optical Packet Switching => natural support for the network virtualization

NOKIA Bell Labs

Latency in VERNE

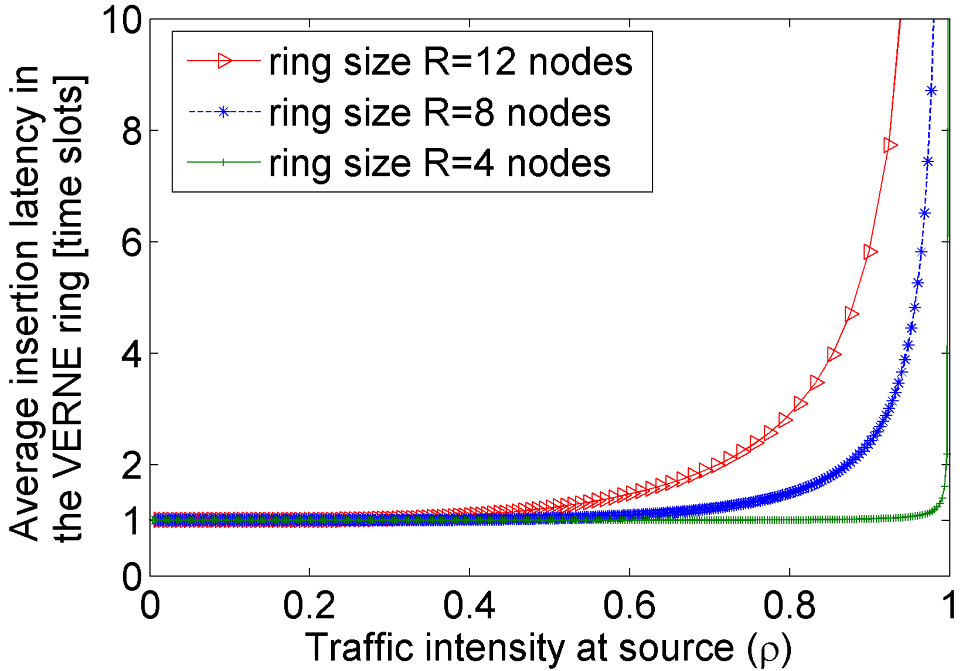
Sources of latency and jitter

	Ethernet	VERNE
Traffic insertion	**YES**	**YES**
Traffic extraction	**YES**	**YES**
Traffic transit (Eth.)	**YES**	n.a.
Traffic transit over the same bus/ring	n.a.	**NO**
Traffic transit when changing the bus/ring	n.a.	**NO**

- Insertion process latency in VERNE can be efficiently reduced by proper network dimensioning

An example of VERNE network with simple scheduling

- A ring VERNE network
- TRXs share the same wavelengths at the reception => Simple FIFO scheduling
- Geo/Geo/1 queueing model => average insertion latency in function of traffic intensity
- **Latency limited to few time slots**



NOKIA Bell Labs

VERNE: a scalable solution

- **<u>A scalable solution:</u>**
  - Can be mapped/installed over any physical topology
  - Node size is reduced, and number of TRX is at minimum
  - Packet blockers optional or not needed

Use of Packet Blockers

	VERNE I	VERNE II	VERNE III
YES			
NO			X
OPTIONAL	X	X	

- E.g. VERNE is more sca
  - The optical bus in VERNE can be mapped to a wavelength, but also to a waveband, a fiber or a fiber core

Conclusions

- We have shown that VERNE is a **good alternative to the Ethernet** Fat-Tree and 2D torus data centers

- VERNE network is focused on:
  - Reducing the **TRX number** to its minimum
  - Reducing the **network cost**
  - Improving the **network latency**
  - Improving the **network scalability**

NOKIA Bell Labs

THANK YOU !

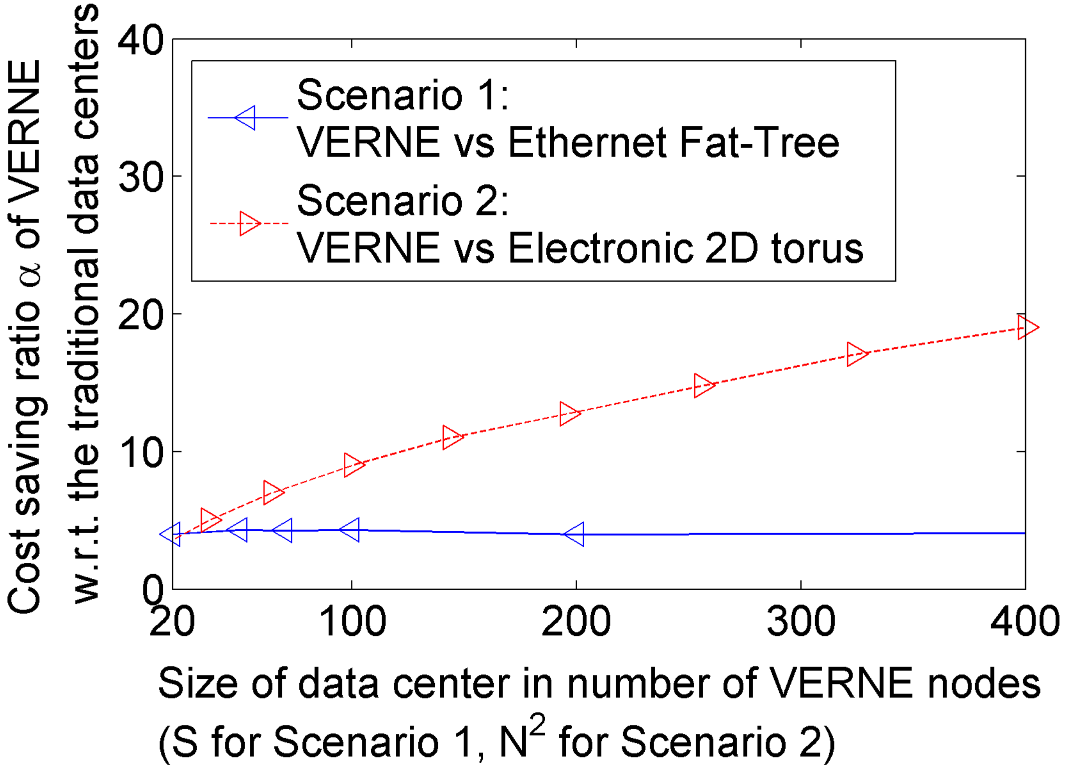
TRX cost savings of VERNE network

- Definition of cost saving ratio α :

$$\alpha = \frac{\#TRX\ (ETH)}{\#TRX\ (VERNE)}$$

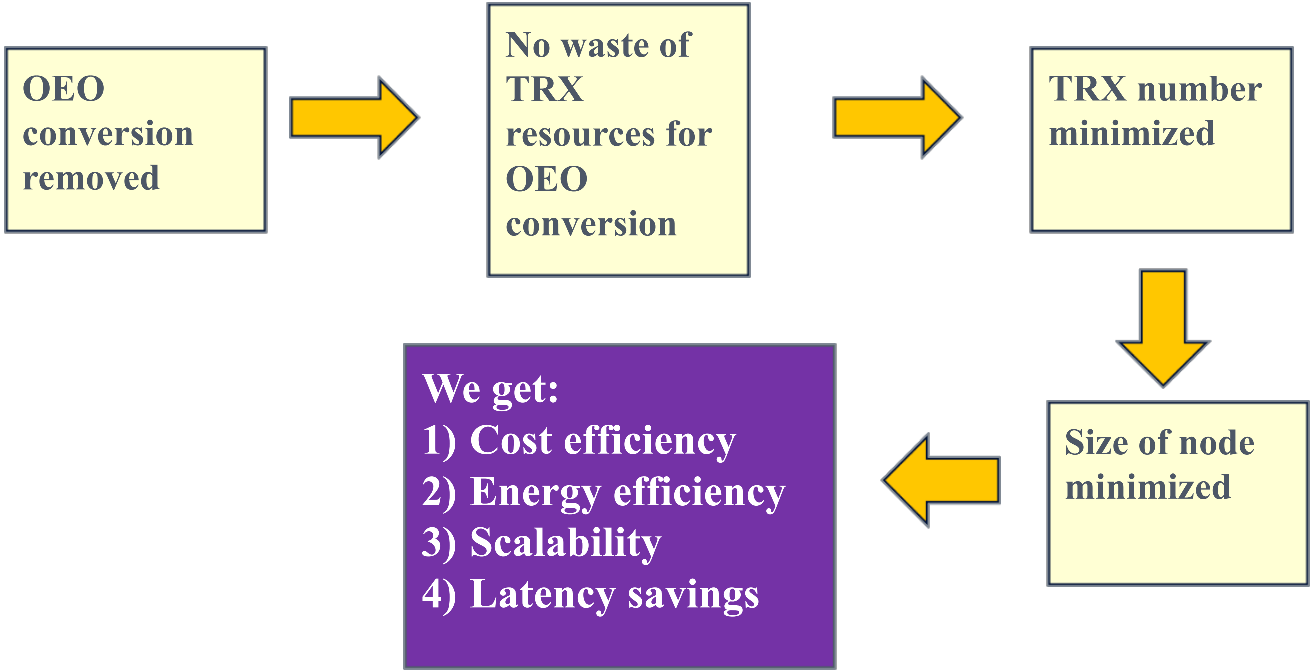
- TRX @ 100 Gbit/sec used both for VERNE and Ethernet

- **Cost savings for torus: x20 times**
- **Cost savings for Fat-Tree: x4 times**



NOKIA Bell Labs

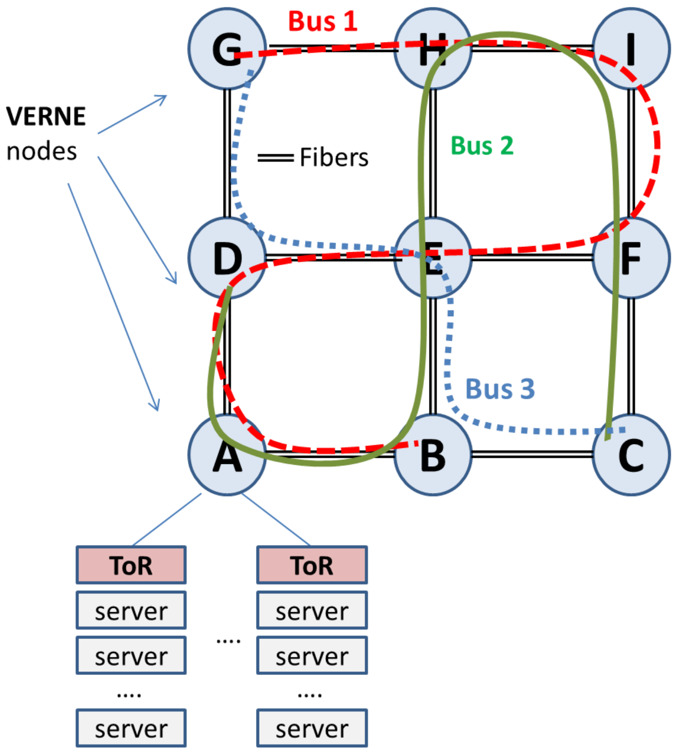
Key idea: remove the OEO conversion from the network



Reconfigurability of the VERNE network

- VERNE can be installed over a pre-existing physical fiber topology

- A key property of VERNE is the **reconfigurability**



The example of VERNE network interconnection, when reach limit is set to 7 nodes

NOKIA Bell Labs

VERNE III: the principle of centralized scheduling

- The **SDN controller** is responsible for the scheduling of the traffic on any "optical bus"

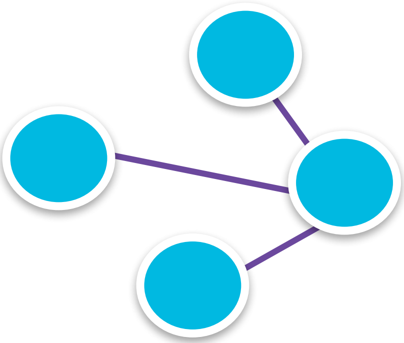
- The scheduling is based on the **dynamic control plane exchange** between the SDN controller and the source nodes

- To reach 100% throughput, it is essential to be in the case of **a separable graph** (containing a central point or a ring)

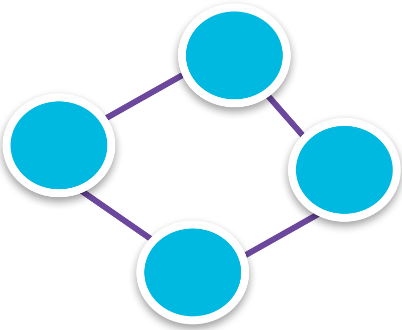
NOKIA Bell Labs

On the separability of the graphs and its impact on the network operation

- Imagine the entire OPS network as a cross bar electronic switch, with N inputs and N outputs
- For the separable and F-separable network topologies, 100% throughput scheduling is possible *
- Separability means that the delay from any source i to any destination j can be expressed as a sum of two delays u(i) + v(j), where u(i) is propagation delay to a central hub, and v(j) the propagation delay from it



Central point graph (star)



Ring graph (F-separability, where F=RTT)

*Keslassy, I.; Kodialam, M.; Lakshman, T.V.; Stiliadis, D., "Scheduling schemes for delay graphs with applications to optical packet networks," *HPSR 2004*