Enabling Low Latency at Large-Scale Data Center and High-Performance Computing Interconnect Networks Using Fine-Grained All-Optical Switching Technology

Author:

Nan Hua, *Zhizhen Zhong, Xiaoping Zheng*

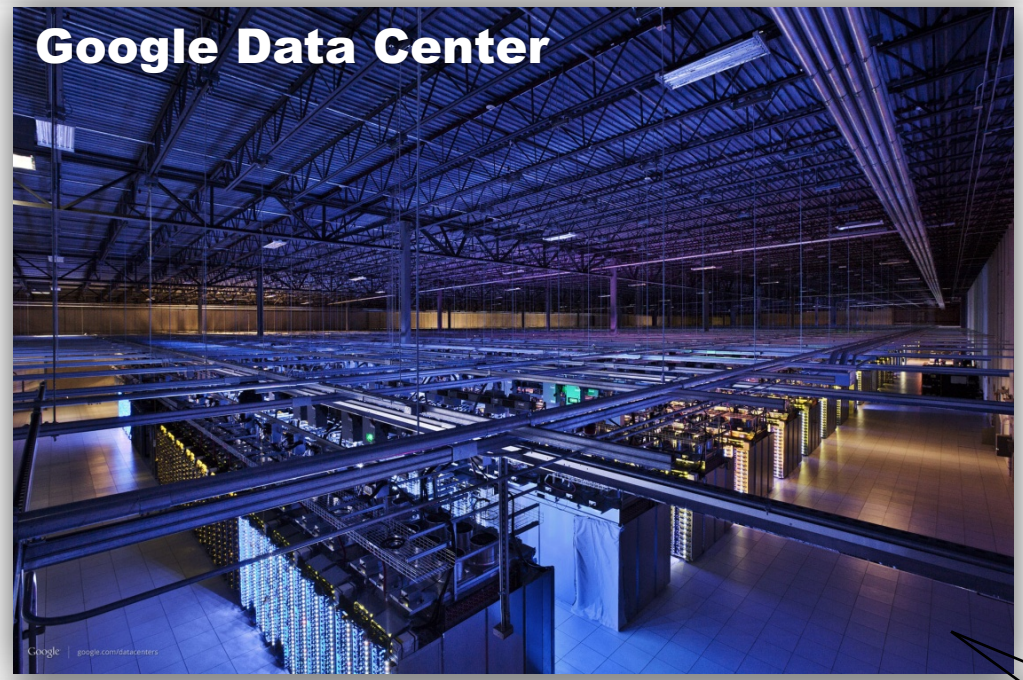
(huan@mail.tsinghua.edu.cn)

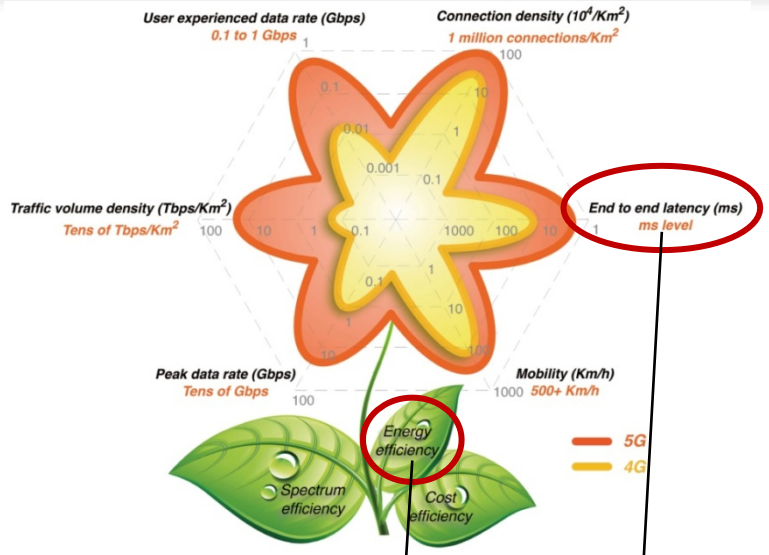
Tsinghua University, China

Outline

- **Motivation**
  challenges for large-scale DC and HPC networks
- **Can optical switching reduce latency?**
  scale, latency, throughput and granularity
- **Temporal-domain optical switching**
  optical time slice switching (OTSS)
- **Simulation and experimental results**
- **Conclusions**

Motivation - *challenges for large-scale DC and HPC networks*





5G Key Capabilities

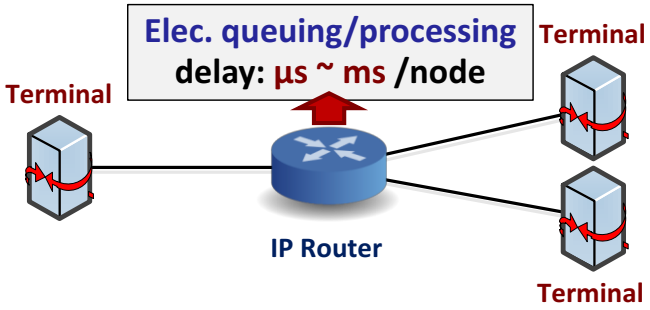
Image Source: IMT-2020 White Paper on "5G Vision and Requirements"

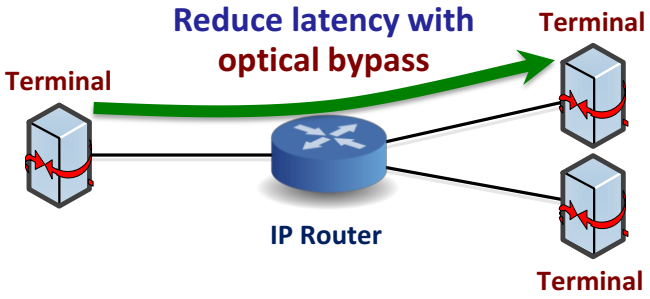
Energy

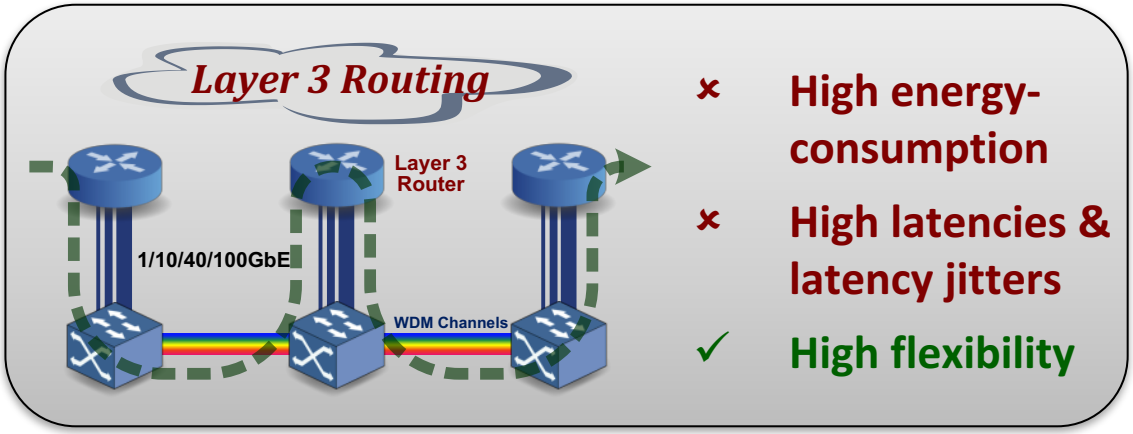
Latency

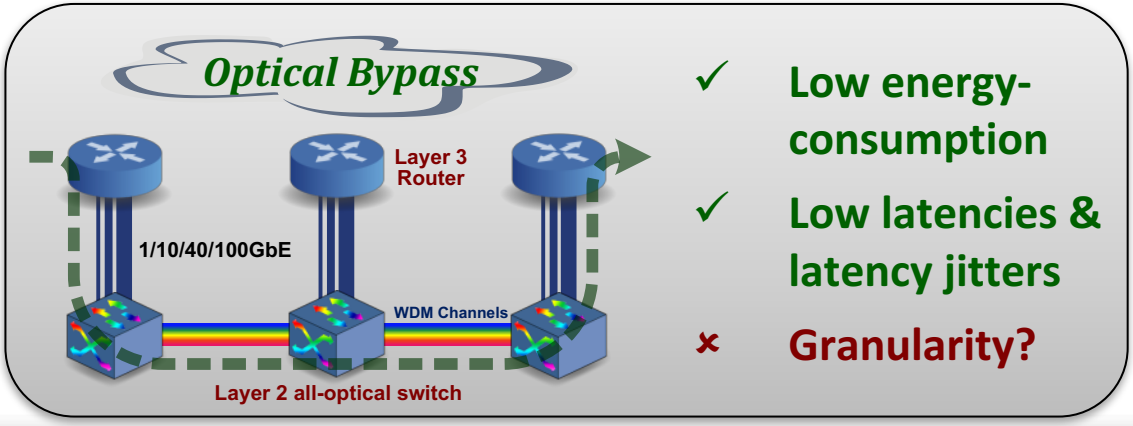
Motivation - *reduce latency by introducing optical bypass*

- **Electronic queuing and processing** delay increase **end-to-end latency**
- Introducing **all-optical switching** technology may reduce latency









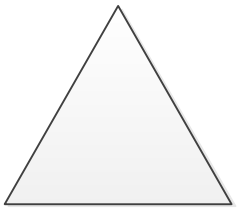
Outline

- **Motivation**
  challenges for large-scale DC and HPC networks
- **Can optical switching reduce latency?**
  scale, latency, throughput and granularity
- **Temporal-domain optical switching**
  optical time slice switching (OTSS)
- **Simulation and experimental results**
- **Conclusions**

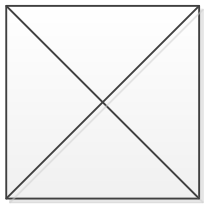
Can optical switching reduce latency?

- End-to-end latency can be reduced by setting up **direct optical connections**, but the number of optical **channels** is **limited**!

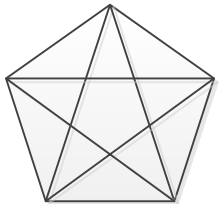
P2P direct connections → **What if 10,000+ nodes?**



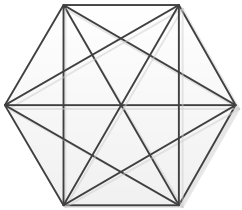
3 nodes



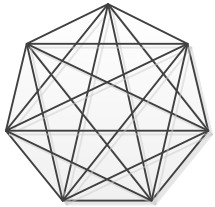
4 nodes



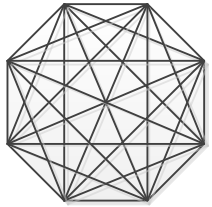
5 nodes



6 nodes



7 nodes

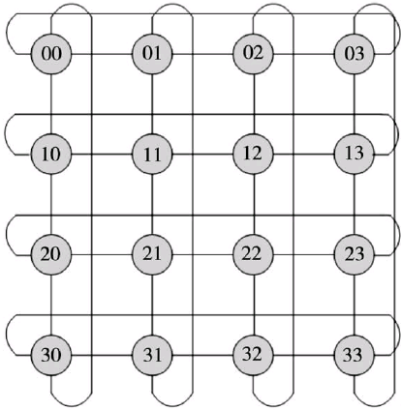


8 nodes

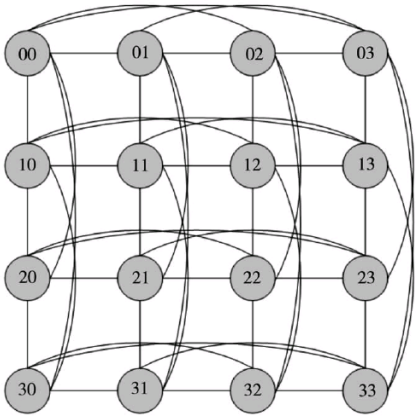
Can optical switching reduce latency?

❖ DC and HPC topologies

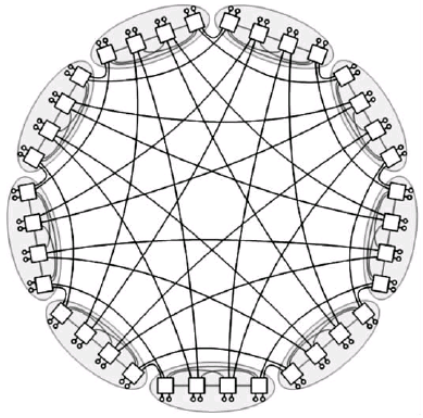
Direct



2-D Torus (d=4)

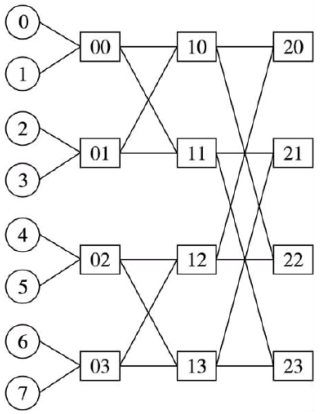


Flattened Butterfly

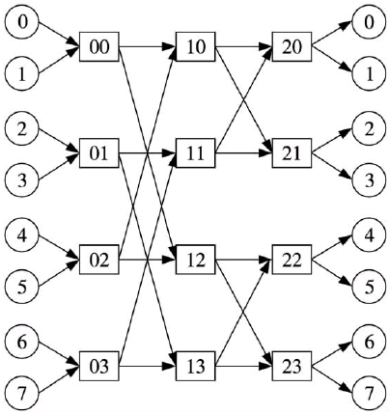


Dragonfly

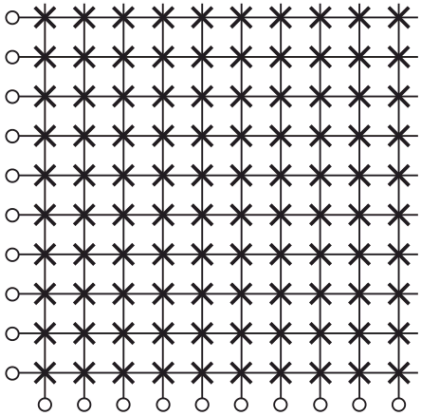
Indirect



Fat Tree (d=4)



Butterfly (d=4)

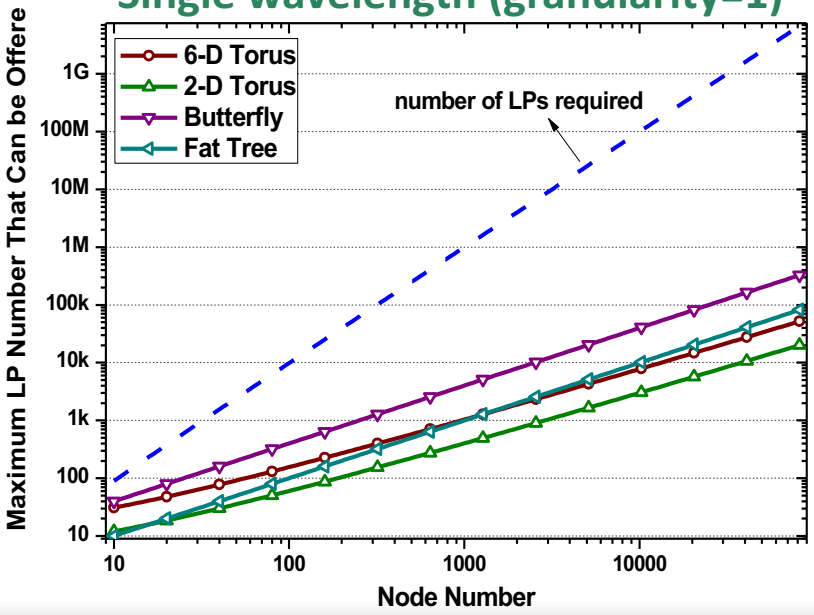


Crossbar

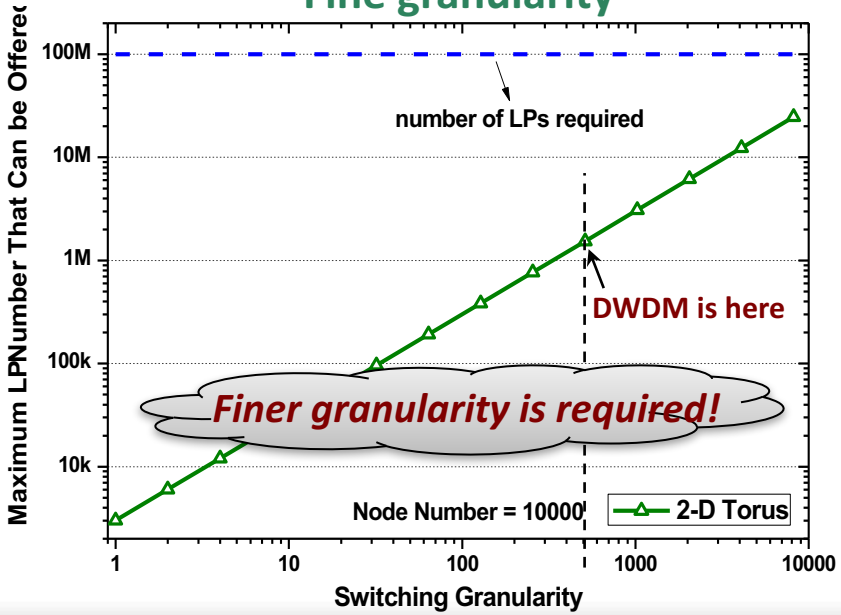
Can optical switching reduce latency?

- Number of LPs **required vs.** that **can be offered**
  - All-to-all communication
  - Set up **Dedicated end-to-end lightpath** for each node pair
  - **Topologies**: 2-D/6-D torus, Butterfly, Fat tree
  - Nodal degree: d= 4 (for 6-D torus: d=12)

Single wavelength (granularity=1)



Fine granularity

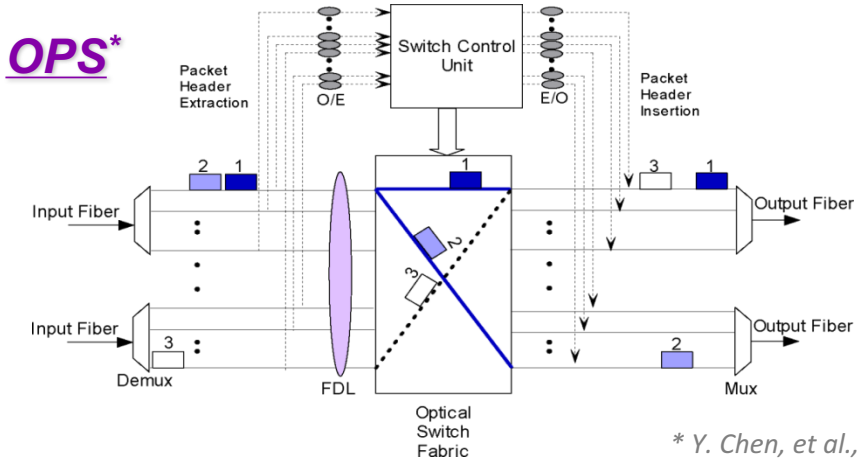


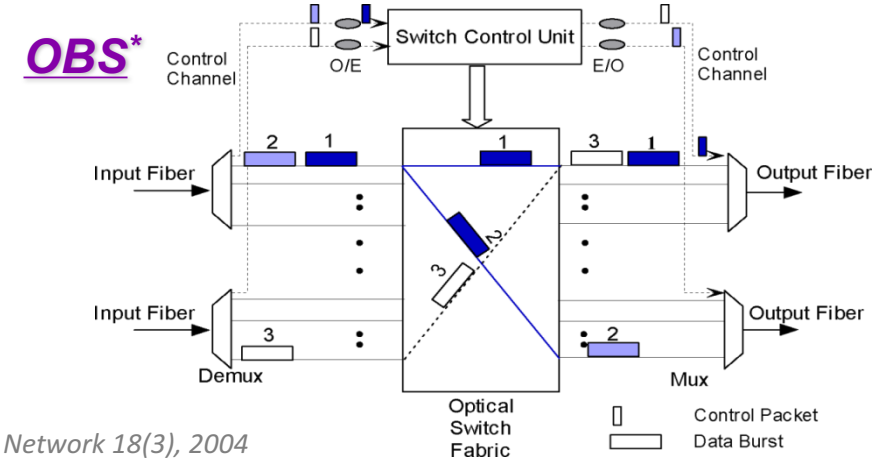
Outline

- **Motivation**
  challenges for large-scale DC and HPC networks
- **Can optical switching reduce latency?**
  scale, latency, throughput and granularity
- **Temporal-domain optical switching**
  optical time slice switching (OTSS)
- **Simulation and experimental results**
- **Conclusions**

Temporal-domain optical switching

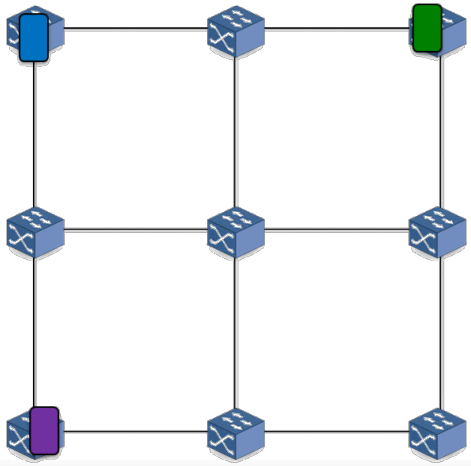
Collisions in OPS/OBS networks





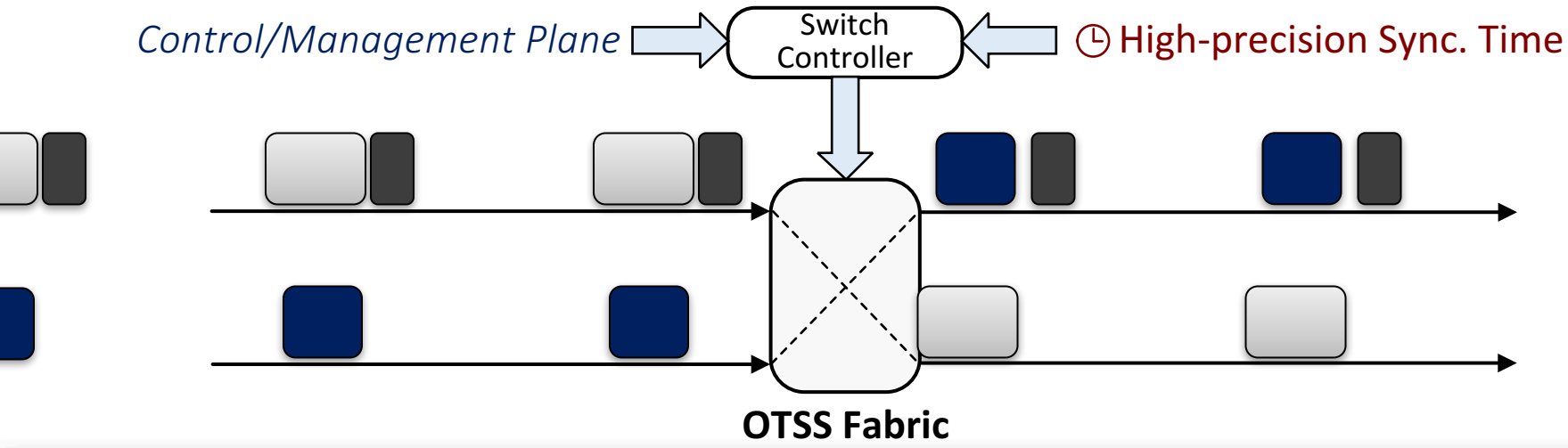
* Y. Chen, et al., IEEE Network 18(3), 2004

- Temporal-domain switching techniques, such as **OPS** and **OBS**, can provide **much finer granularity** than frequency-domain switching techniques
- However, **OPS** requires **optical buffering** (not mature) to avoid collisions.
- For **OBS**, **burst loss rate** could be **very high** at a heavy load without buffer



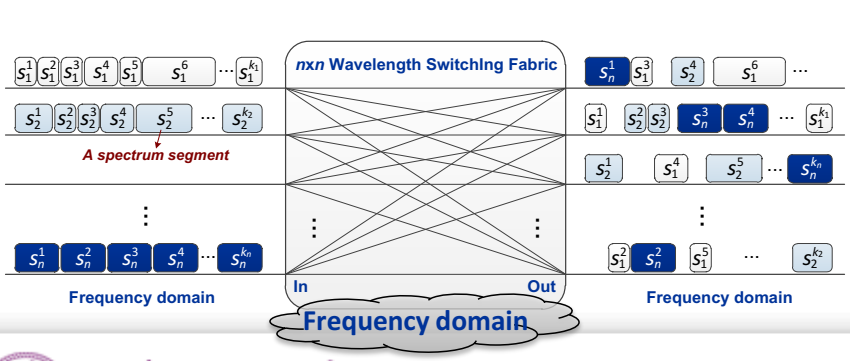
Temporal-domain optical switching

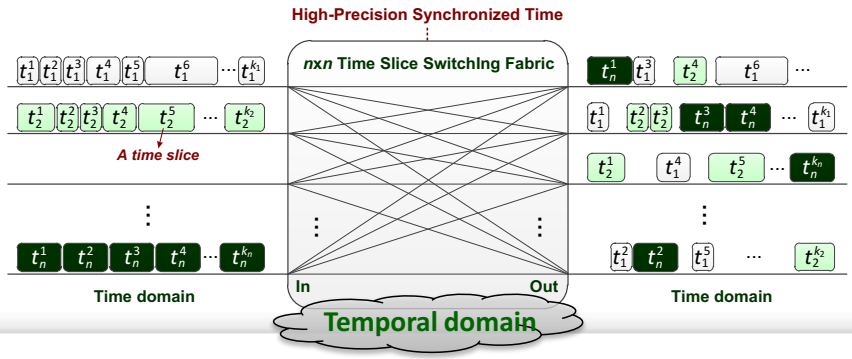
- **Optical time slice switching** (OTSS):
  - Realize **collision-free** fine-granularity optical switching **without buffer**
  - Optical channels are organized into **repetitive OTSS frames** in time domain
  - Each OTSS frame contains **variable-length time slice(s)**
  - The switch controller sends (periodic) control signals to the OTSS fabric **at the precise time** to direct the time slice to the expected output port.
  - **Time synchronization** is required



Temporal-domain optical switching

- **Optical time slice switching** (OTSS):
  - Realize **collision-free** fine-granularity optical switching **without buffer**
  - Optical channels are organized into **repetitive OTSS frames** in time domain
  - Each OTSS frame contains **variable-length time slice(s)**
  - The switch controller sends (periodic) control signals to the OTSS fabric **at the precise time** to direct the time slice to the expected output port.
  - **Time synchronization** is required
  - OTSS is **more easy** to provide **small and flexible slot(s)** compared with (frequency-domain) spectrum-flexible wavelength switching, depending only on switch speed, timing precision and control delay





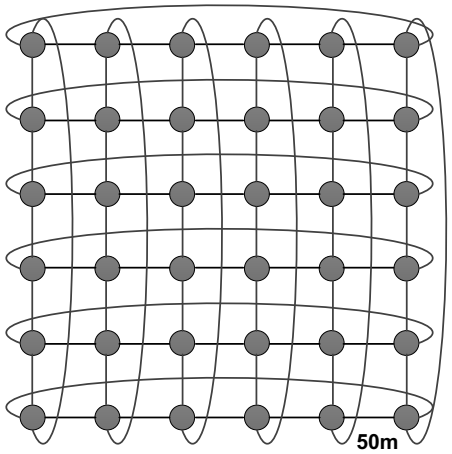
Outline

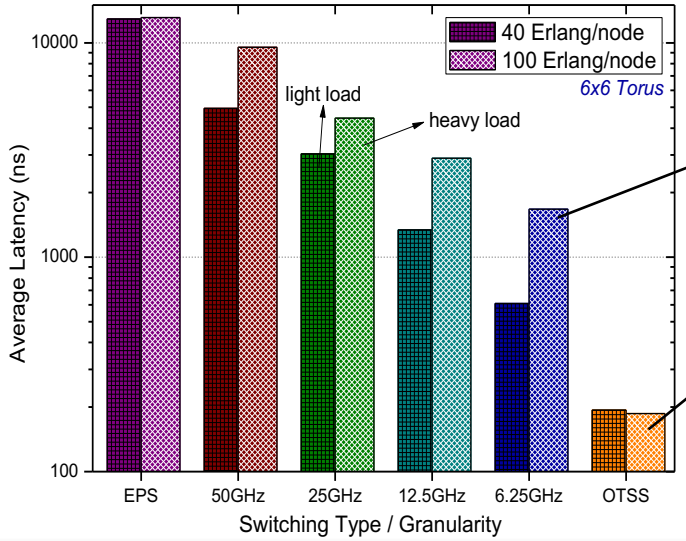
- **Motivation**
  challenges for large-scale DC and HPC networks
- **Can optical switching reduce latency?**
  scale, latency, throughput and granularity
- **Temporal-domain optical switching**
  optical time slice switching (OTSS)
- **Simulation and experimental results**
- **Conclusions**

Simulation and experimental results

Simulation setup and result

- **6x6 2-D Torus** topology (50-m fiber link between adjacent nodes)
- All-to-all Traffic, generated between all node pairs uniformly
- **WS**: **spectrum slot sizes** - 6.25GHz, 12.5GHz, 25GHz and 50GHz
- **Queuing and processing delay** at electronic switch: **5μs**
- **OTSS**: **time slot size** - **100ns**, **frame length** - **20μs**

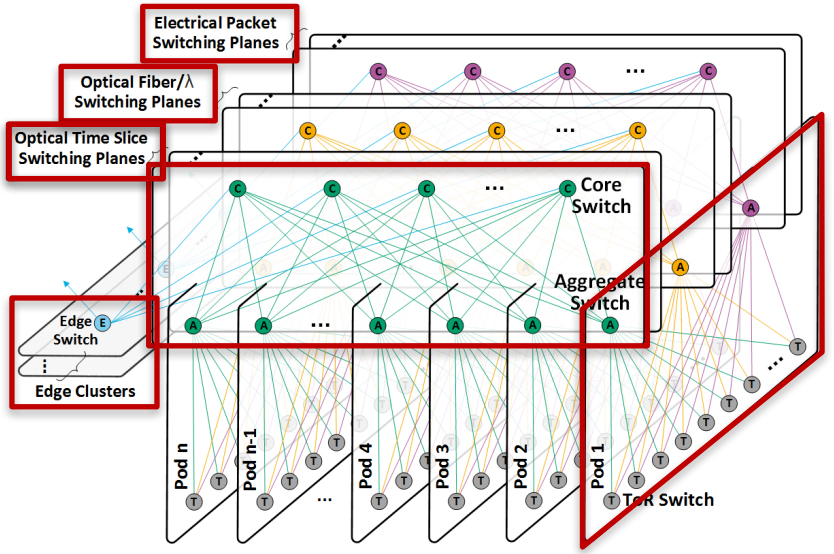


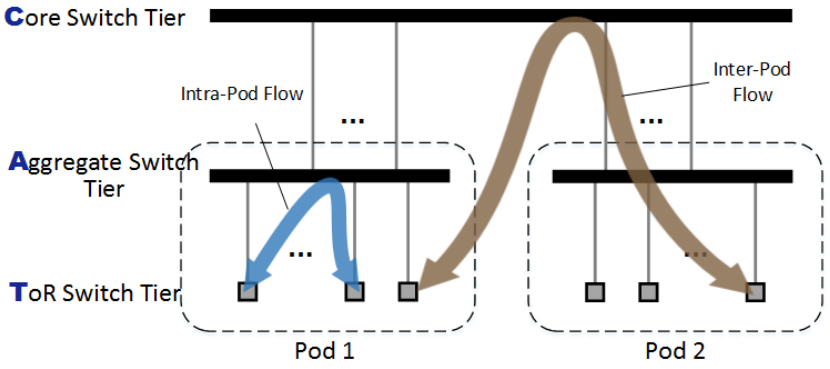


- Fine-grained **WS** is able to **reduce end-to-end latency** compared with **pure EPS**
- **OTSS** is able to achieve **lower latency** due to its finer granularity

Simulation and experimental results

- **Experimental setup**
  - Prototype DCN with Fat-Tree topo
  - **Intra-pod** and **inter-pod** switching
  - OTSS frame length: **100μs**
  - Guard interval: **100ns**
  - Time slice lengths of four data flow: **1μs, 1μs, 10μs and 25μs**

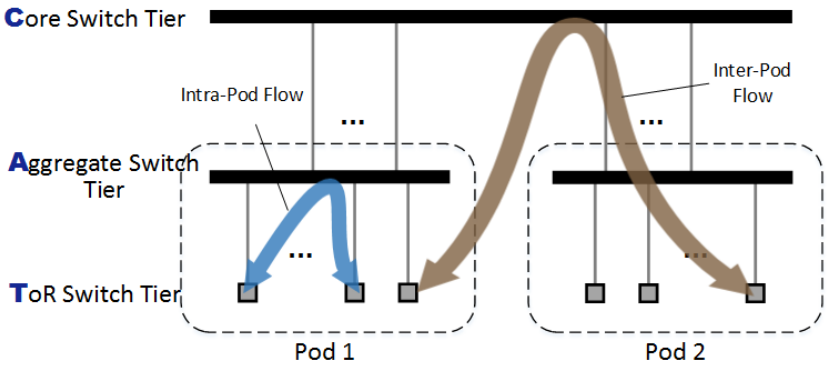


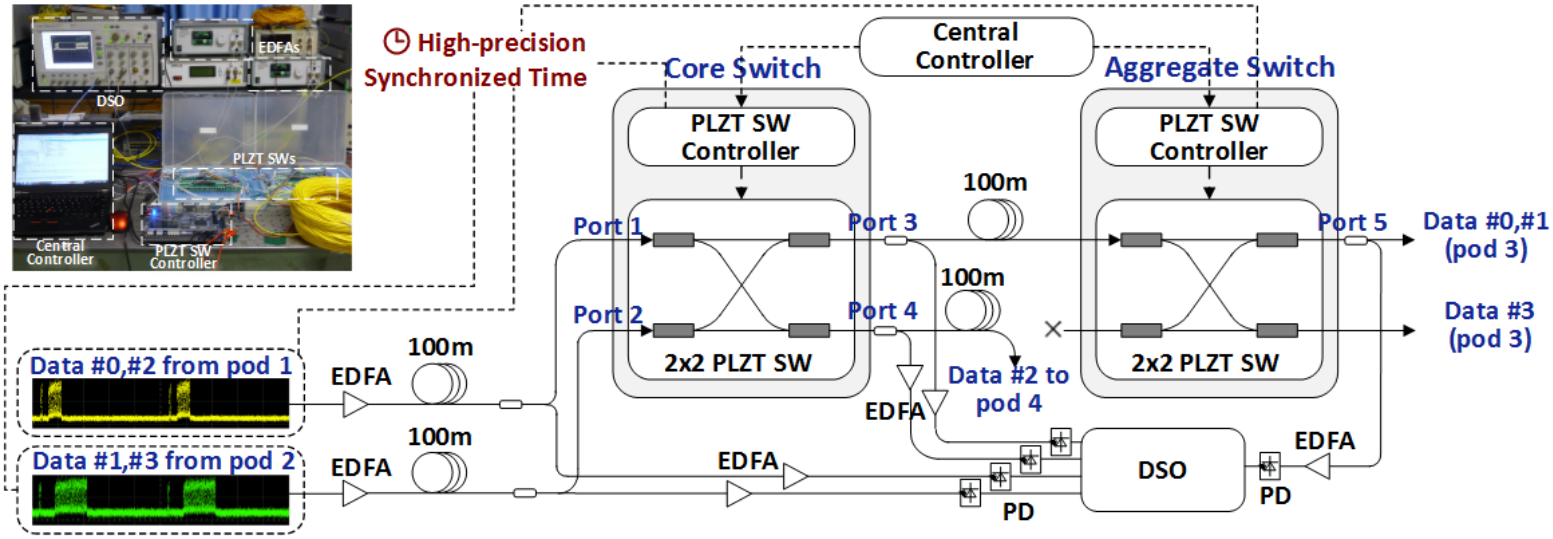


Simulation and experimental results

Experimental setup

- **Intra-pod** and **inter-pod** switching
- OTSS frame length: **100μs**
- Guard interval: **100ns**
- Time slice lengths of four data flow: **1μs, 1μs, 10μs and 25μs**

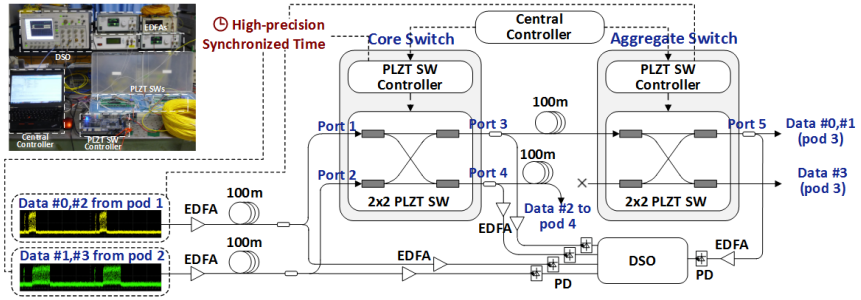


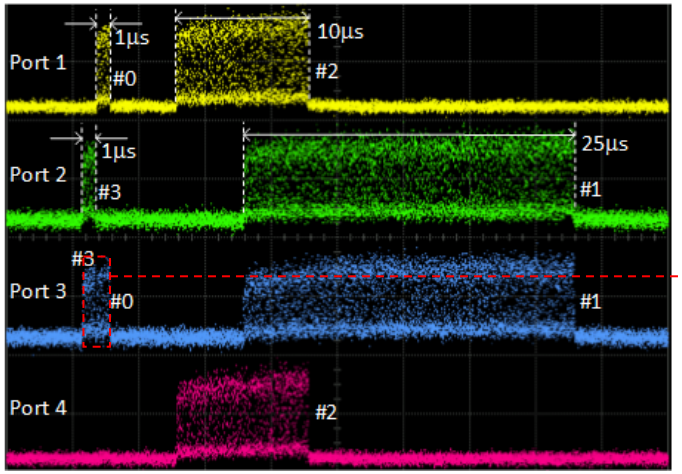


Simulation and experimental results

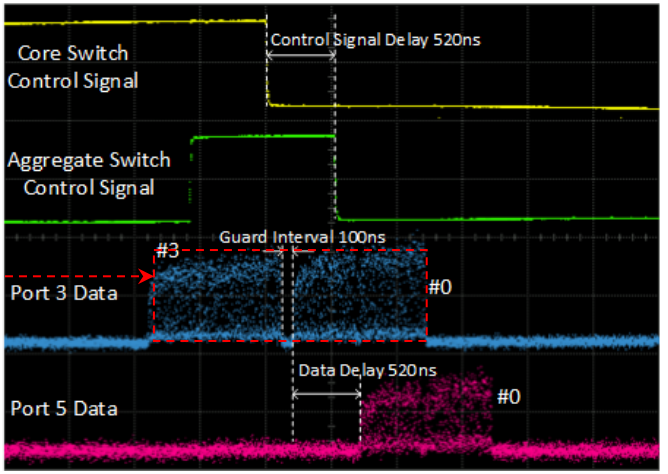
Experimental results

- **Intra-pod** and **inter-pod** switching
- OTSS frame length: **100μs**
- Guard interval: **100ns**
- Time slice lengths of four data flow: **1μs, 1μs, 10μs and 25μs**





Time slice separation and aggregation



Transmission and control signal delay

Outline

- **Motivation**
  challenges for large-scale DC and HPC networks
- **Can optical switching reduce latency?**
  scale, latency, throughput and granularity
- **Temporal-domain optical switching**
  optical time slice switching (OTSS)
- **Simulation and experimental results**
- **Conclusions**

Conclusions

- As the scale of the DC and HPC networks increases, **end-to-end latency** is becoming a fatal problem due to the **electronic processing and queuing delays** at intermediate switching nodes
- **Introducing optical switching** can provide a potential solution to the latency problem; however, its effect is limited for large-scale networks due to the **restricted switching granularity**
- We introduce our proposed fine-grained optical time slice switching (**OTSS**) into DC and HPC networks, which is able to offer **over one thousand channels** on a single wavelength
- Simulation and experimental results demonstrate that **OTSS is able to achieve lower latency** compared with conventional EPS and spectrum-flexible WS **due to its finer granularity**

Thank You

for Your Attention!