Optics for the Cloud

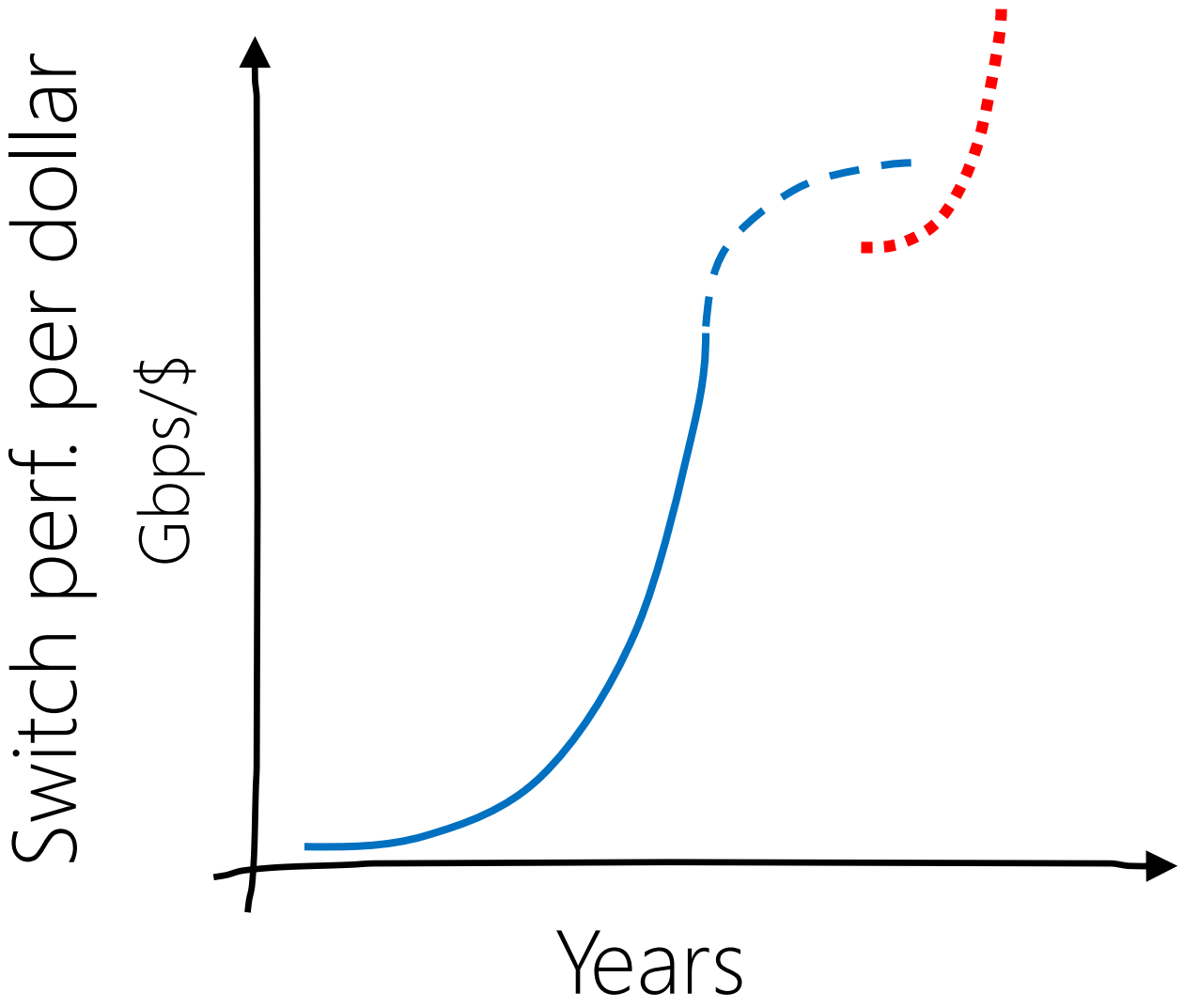
Opportunities and Challenges

Hitesh Ballani

Daniel Cletheroe, Paolo Costa, Istvan Haller, Krzysztof Jozwik, Fotini Karinou, Sophie Lange, Kai Shi, Benn Thomsen

Microsoft Research

Will free scaling of the network continue?

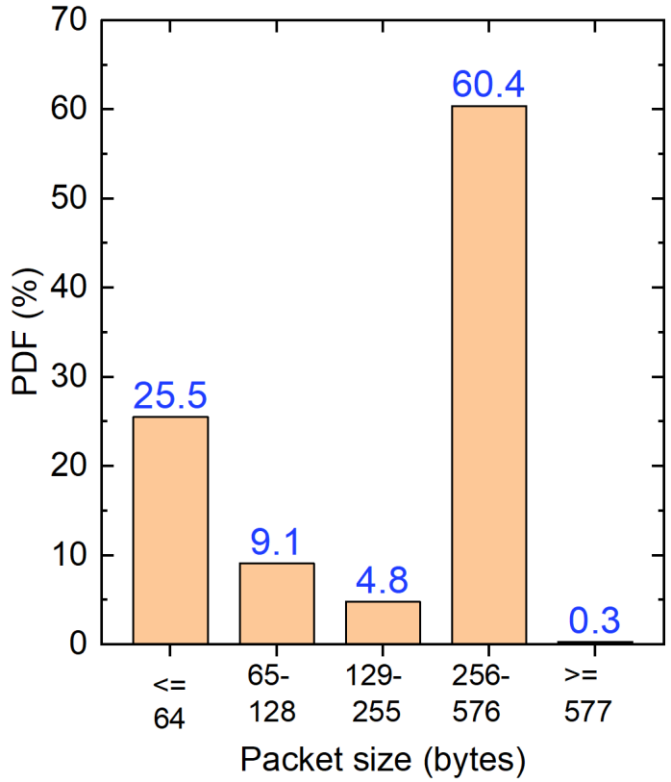


Requirements for data center switching

1. Ultra-fast
   - nanosecond switching

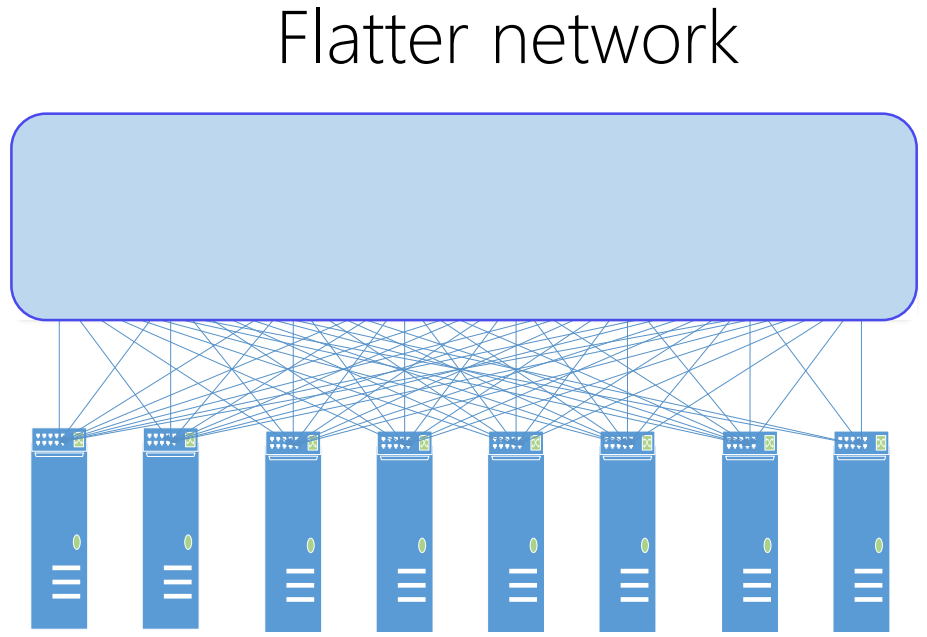
Bursty cloud applications

- 90% packets less than 576 bytes

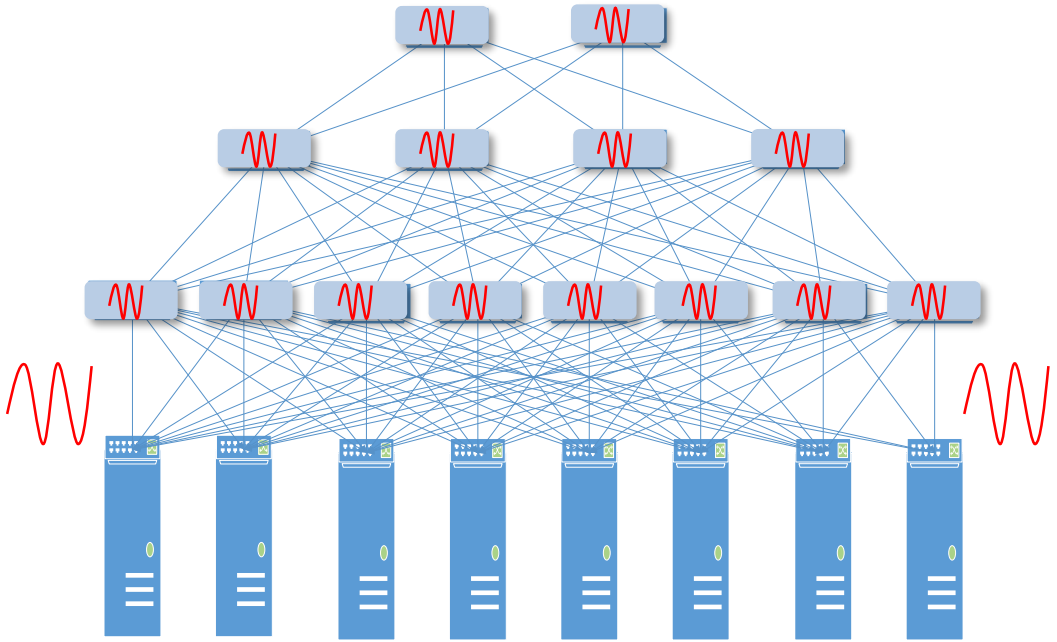


Requirements for data center switching

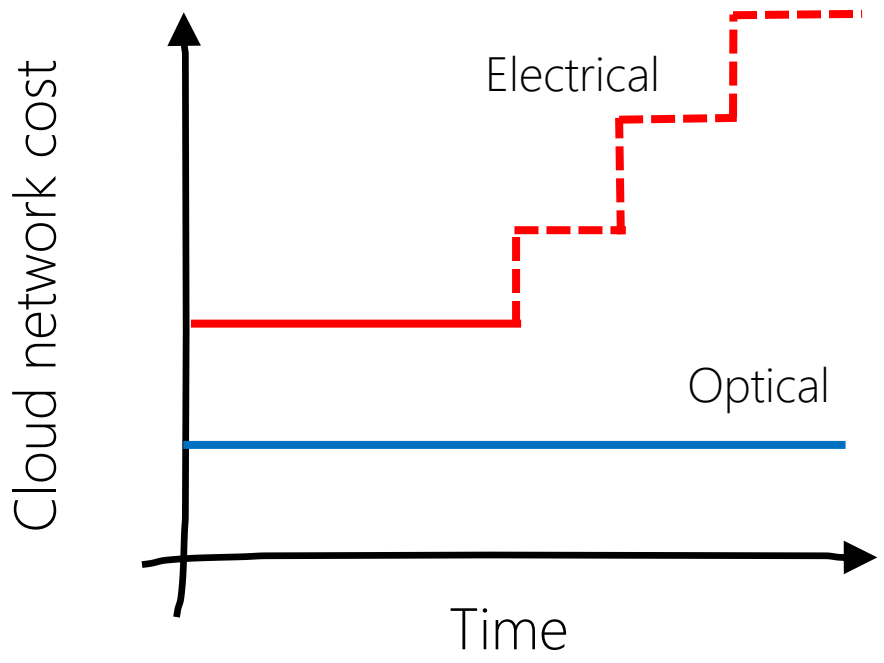
1. Ultra-fast
   - nanosecond switching
2. Scalable (ideally, high-radix)
   - flat network
3. Reliable and easy to manage



Could photonics offer a new growth curve?

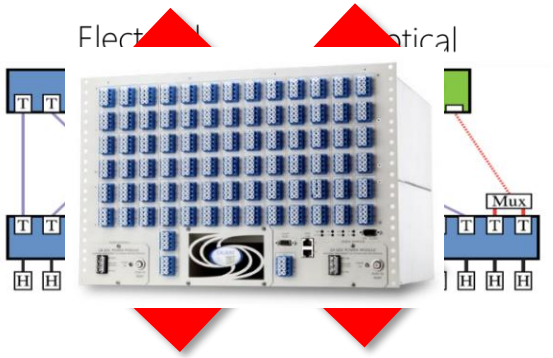


Potential benefits



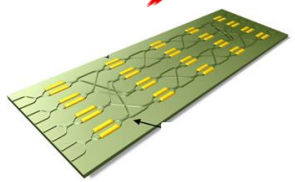
Low and predictable latency

Why the hold-up?



two separate networks
complexity

What is the last mile so hard?



EAMs
MZIs
SOAs
Tunable lasers and AWGRs
…

Data center scale
(1,000s-10,000s)

Cluster scale
(10s-100s)

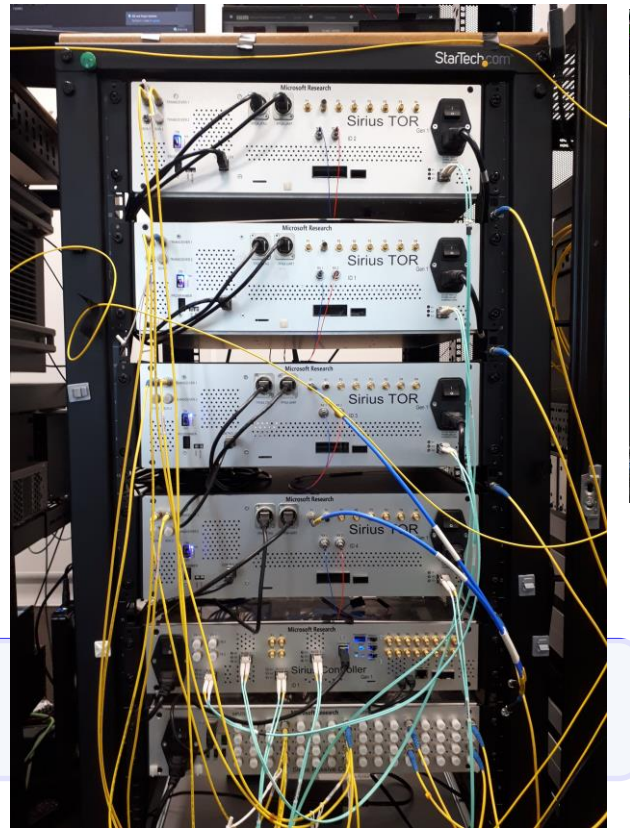
Scale of solution

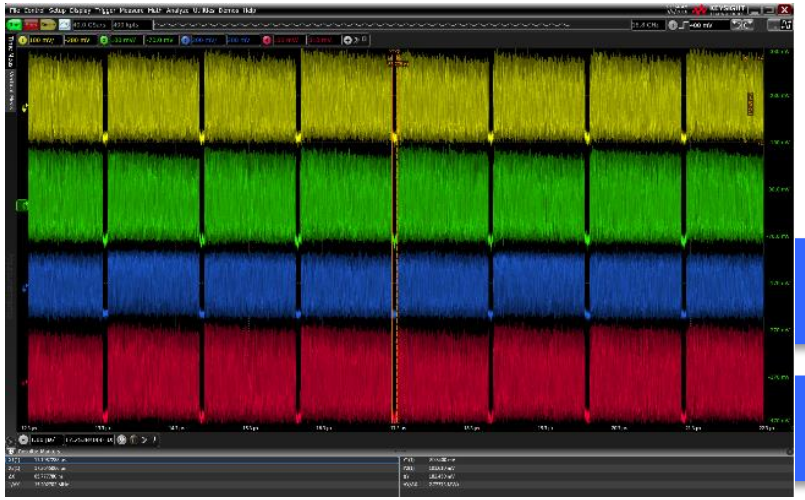
Slow (ms)

Fast (ns)

Switching speed

Bridging the last mile





Apps

Network Stack

NIC

PHY

NIC

PHY

A fundamentally different abstraction

From *asynchronous* packet switches to *synchronous* circuit switches

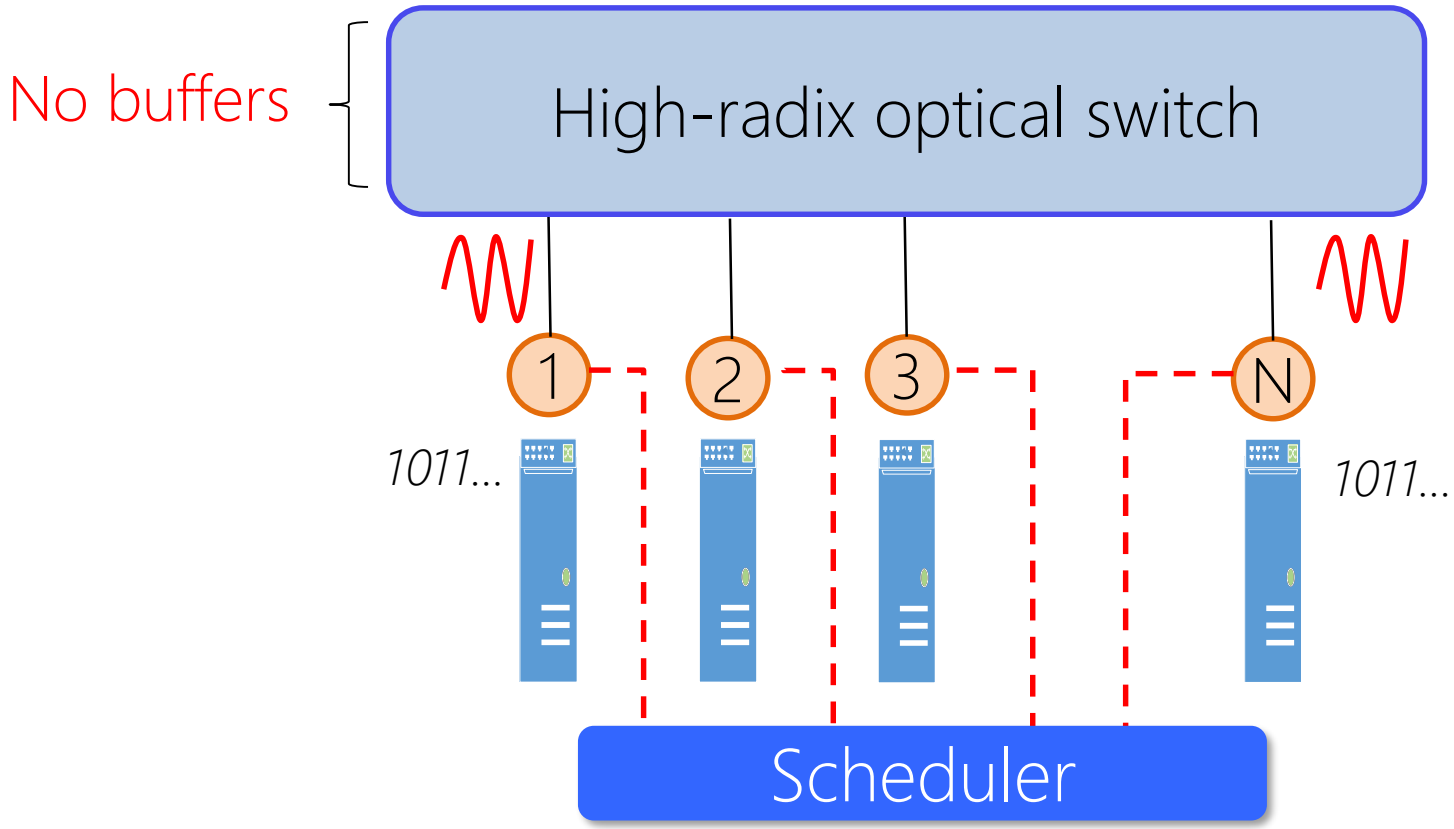
Lots of very hard problems to solve to make optical switching practical

Need for cloud-centric, cross-layer solutions

Two case studies

Problem	Traditional solution
Lack of buffering	Centralized Scheduler
Sub-nanosecond CDR	-

Case study #1: Scheduling the network



Building a data center wide scheduler is hard

Scale to 100K servers, Infer demand, Communicate demand

Scheduler-less network

[Cheng et al., 2000]

A permutation of connections

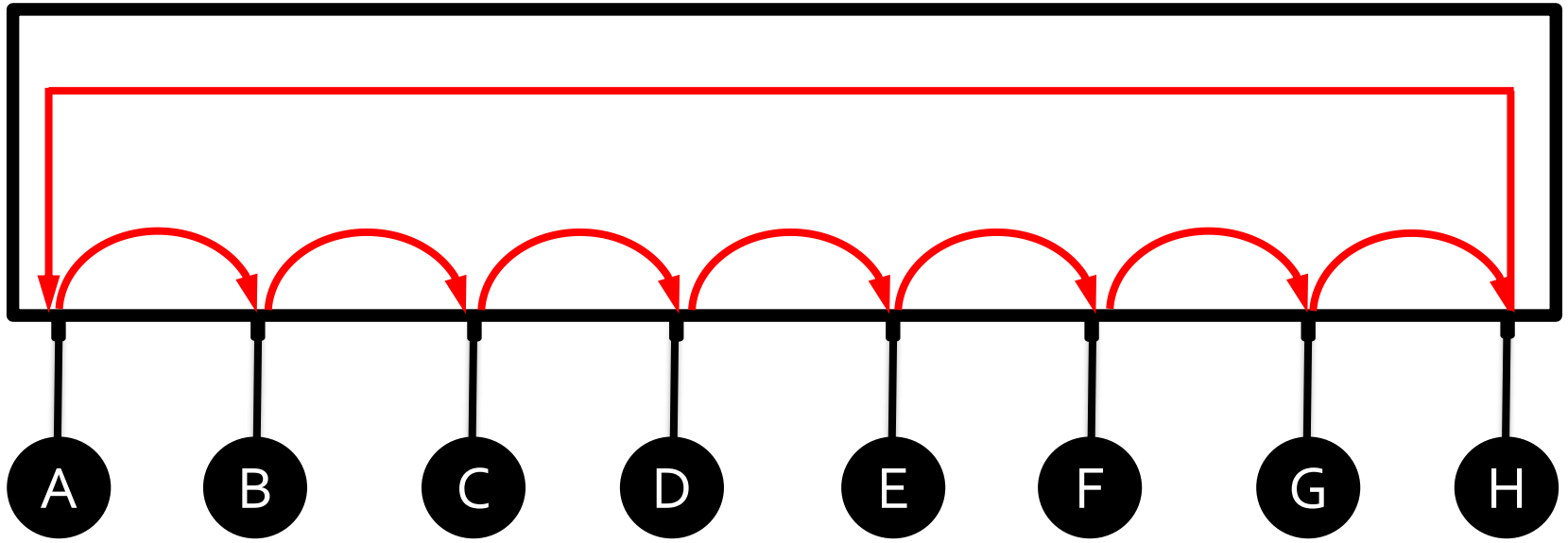
N-1 time slots

Time slot

	1	2	3	4	5	6	7
A	B	C	D	E	F	G	H
B	C	D	E	F	G	H	A
C	D	E	F	G	H	A	B
D	E	F	G	H	A	B	C
E	F	G	H	A	B	C	D
F	G	H	A	B	C	D	E
G	H	A	B	C	D	E	F
H	A	B	C	D	E	F	G

Static pre-defined schedule

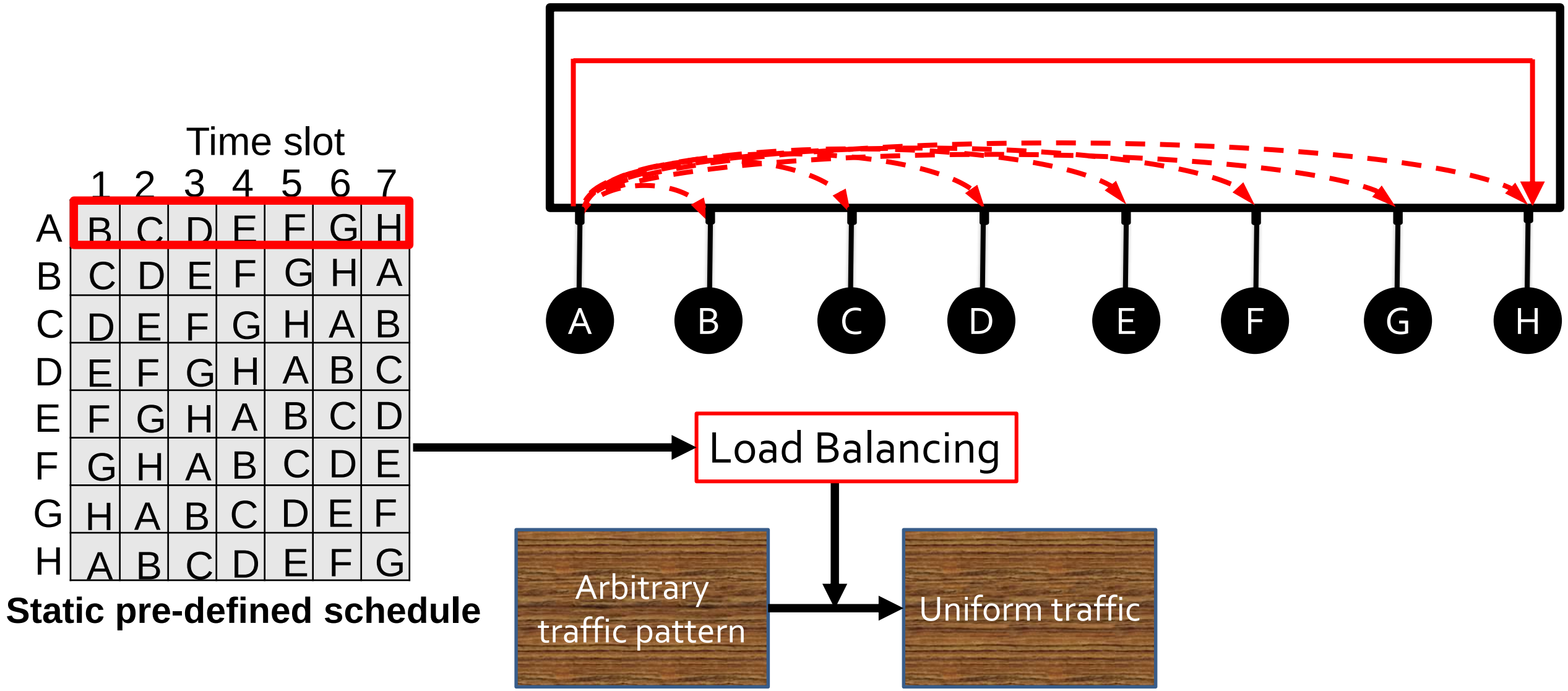
(a cyclic permutation)



Uniform traffic → 100% throughput

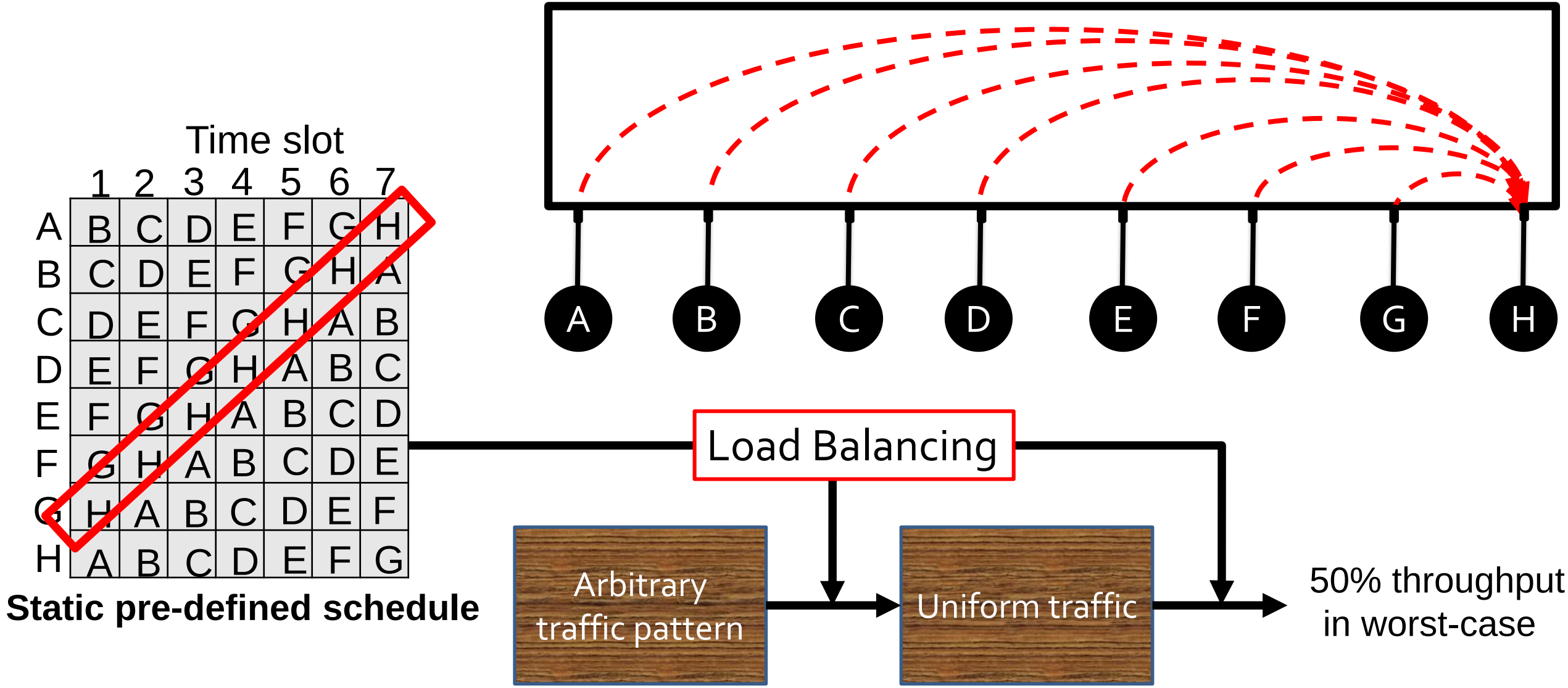
Scheduler-less network

[Cheng et al., 2000]



Scheduler-less network

[Cheng et al., 2000]



Good trade-off

- Through and latency overhead of 2-hop paths

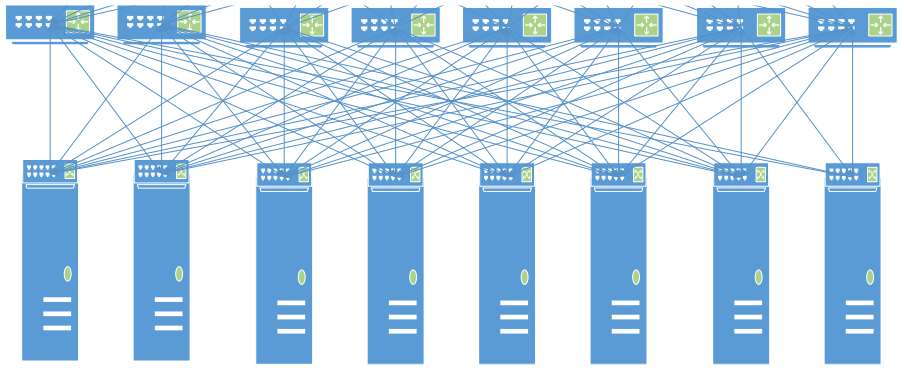
+ *No scheduling!*

"Shoal: A Network Architecture for Disaggregated Racks", Cornell and Microsoft Research, NSDI 2019

"RotorNet: A Scalable, Low-complexity, Optical Datacenter Network", UCSD, SIGCOMM 2017

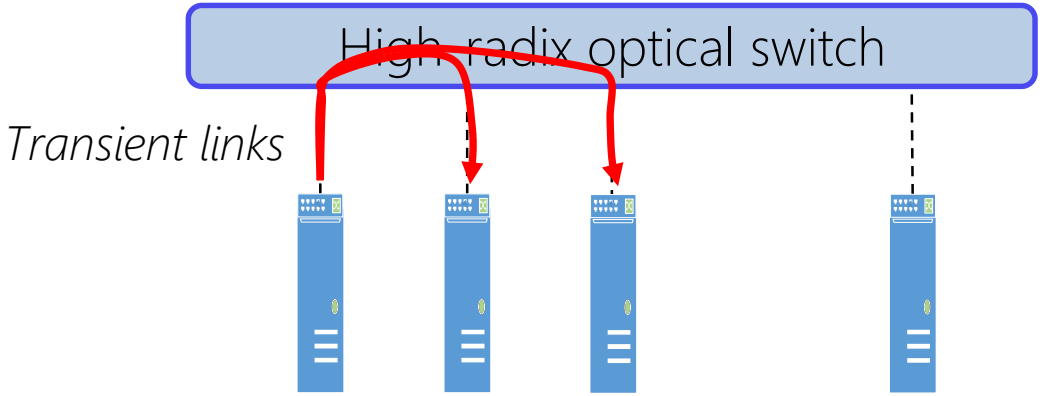
Case study#2: Clock and Data Recovery

Today's network



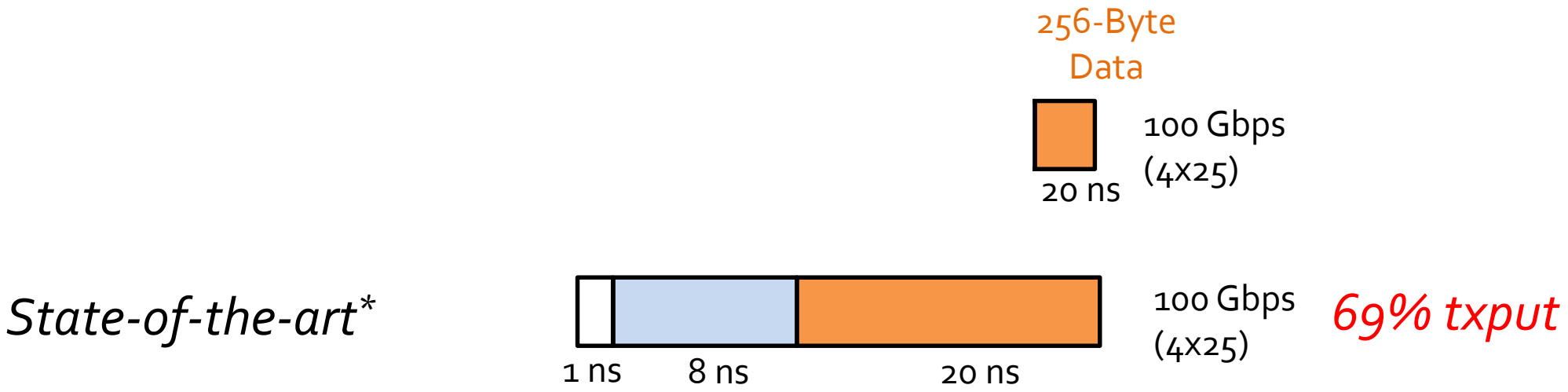
- Links are point to point and always on
  - CDR Locking time does not matter

Optical network



- Links can be established every few ns
  - *Locking time impacts overall throughput*

Status quo on CDR



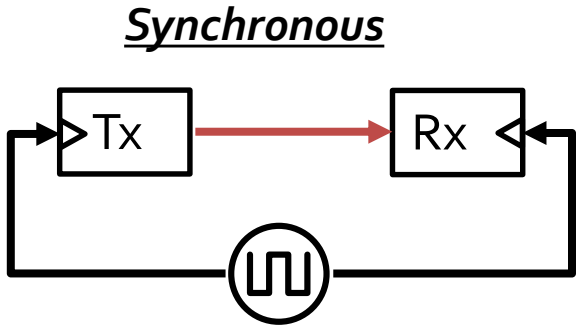
High locking time impacts network throughput

* A Cevrero et al., IBM Research and EPFL, OFC 2018

Need for sub-nanosecond CDR

Phase caching: sub-nanosecond CDR

But Optical Switches *require* ***synchronisation*** to avoid packet collisions*!*

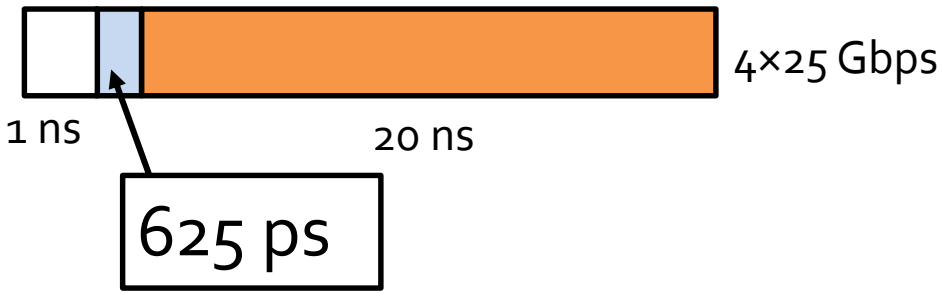


CDR problem reduced to phase discovery

But it can still take >40ns to recover phase

Phase does not change often, cache it!

Fast optical switching starts looking viable



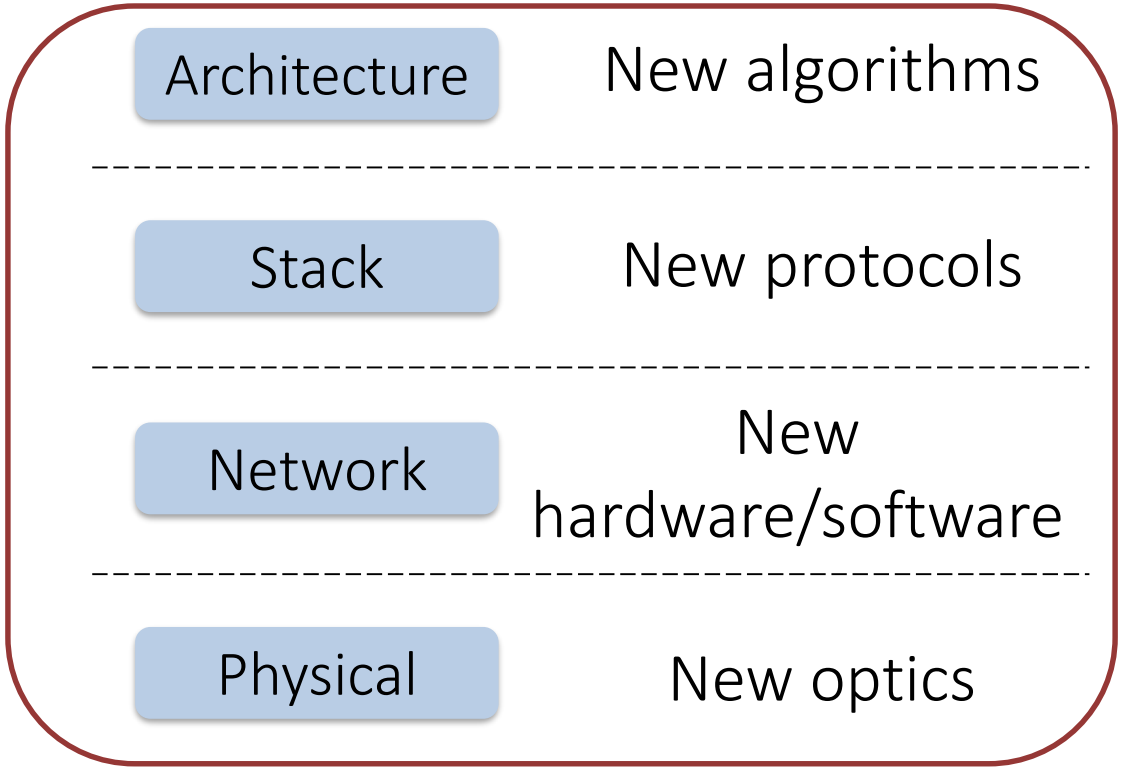
Cross-layer approach that leverages DC peculiarities to achieve sub-ns CDR

Good trade-off in this setting

"Sub-Nanosecond Clock and Data Recovery in an Optically-Switched Data Centre Network", UCL and Microsoft Research, ECOC PDP, 2018

Innovation across the cloud stack needed

Need an end-to-end, cloud-centric approach to make optical switching viable



http://aka.ms/OpticsForTheCloud