Reshaping Transport and Traffic Engineering in Reconfigurable Data Center Networks

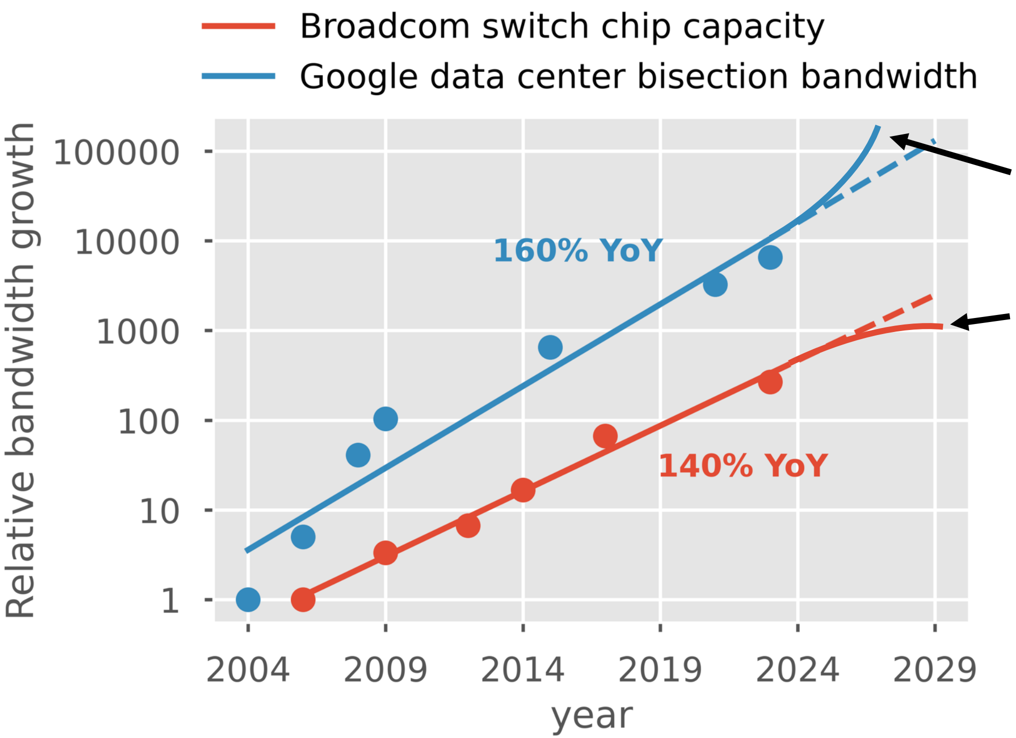
Shawn Shuoshuo Chen

Workshop on Reconfigurable Networks
June 2, 2025

Carnegie Mellon University

Funding to attend this conference was provided by the CMU GSA/Provost Conference Funding.

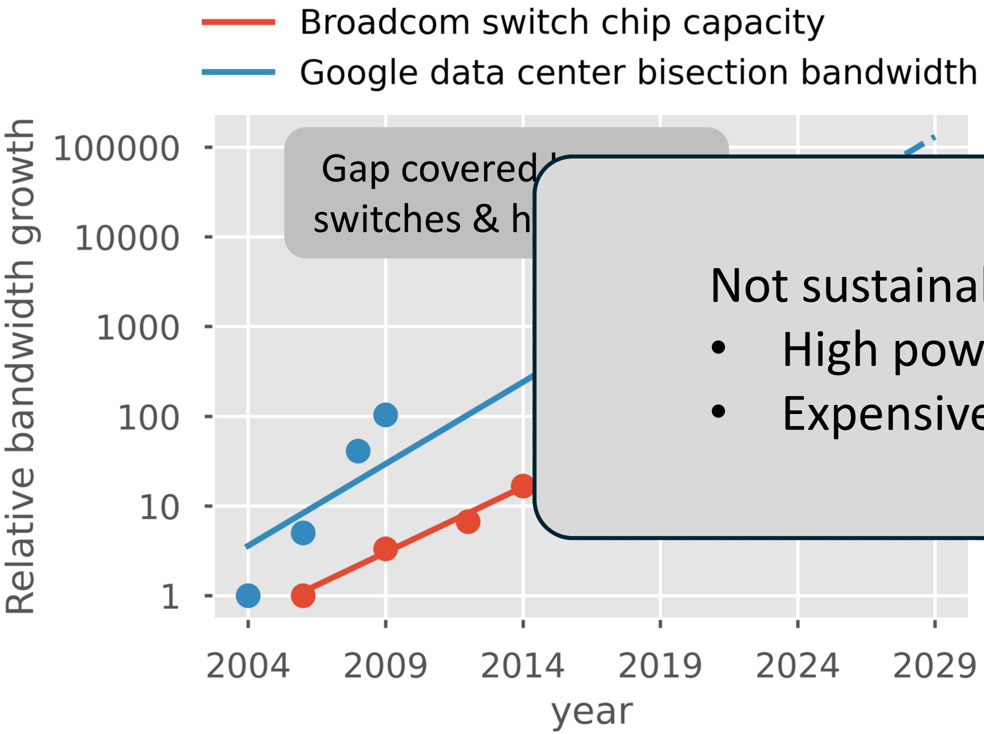
The **scaling crisis** of data center networks



Driven by AI, AR/VR

Dying Moore’s Law

The **scaling crisis** of data center networks



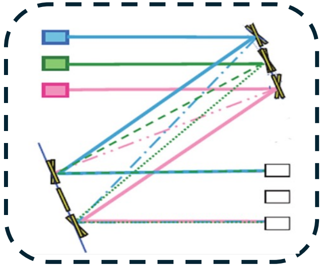
🟧 Switch

⬜ Host

Not sustainable

- High power consumption
- Expensive

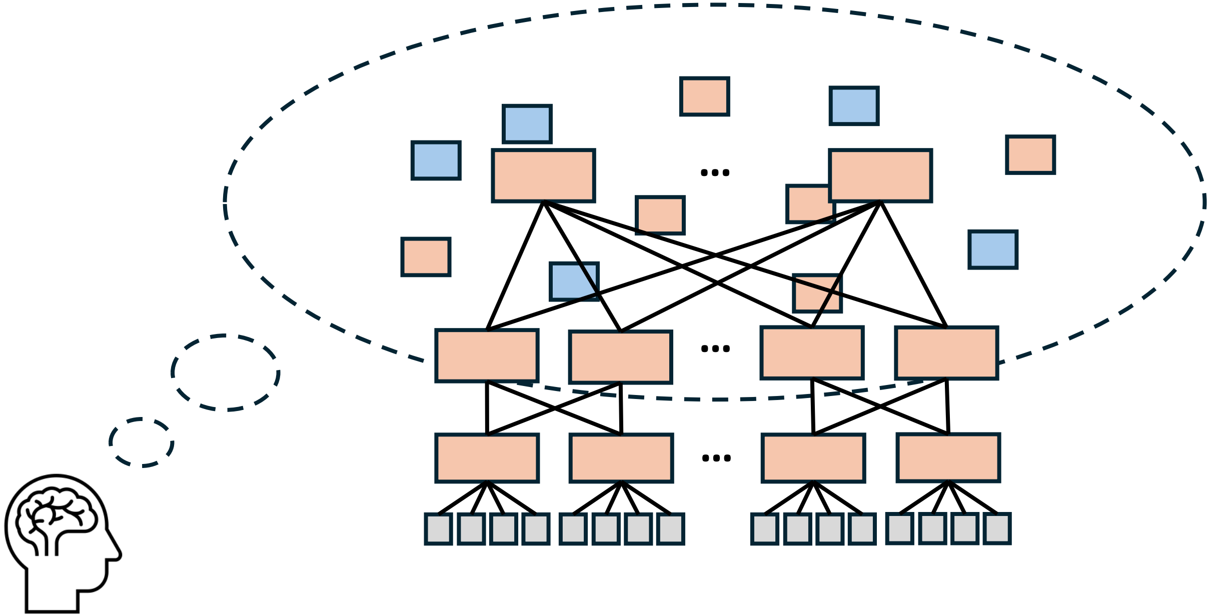
What is optical circuit switch (OCS)?



Optical circuit switch
1:1 in/out mapping
Down during **reconfiguration** (µs-ms)
Much higher b/w, lower latency
Data rate agnostic

Electrical packet switch
Packet-level multiplexing
No down period
Lower b/w, higher latency
Fixed rate per generation

Reconfigurable Data Center Networks (RDCNs)



🟦 Optical circuit switch (OCS)

🟧 Electrical packet switch (EPS)

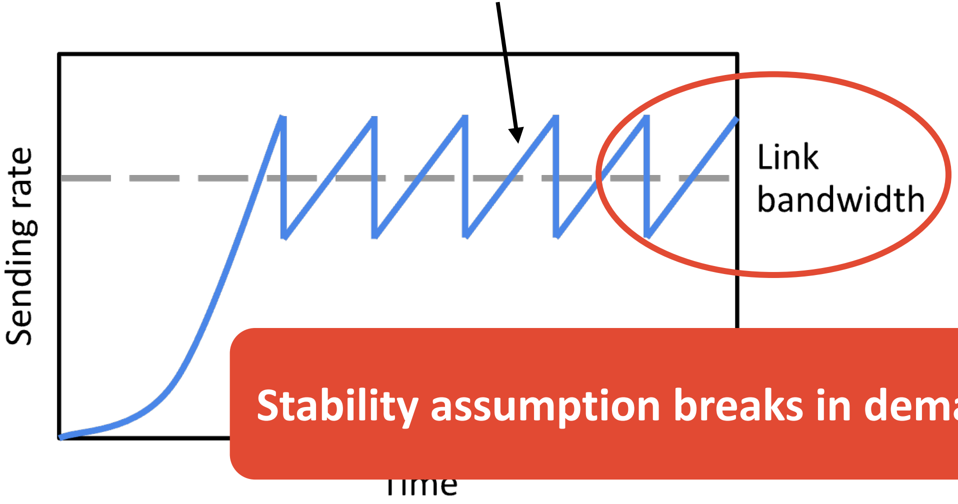
⬜ Host

Today's talk

- **Transport**
  - Time-division TCP for demand-oblivious RDCNs *[SIGCOMM'22]*
- **Traffic engineering**
  - Precise traffic engineering for demand-aware RDCNs *[NSDI'24]*

Existing TCP’s assumption: stable network path

TCP’s goal: match sending rate to available bandwidth.

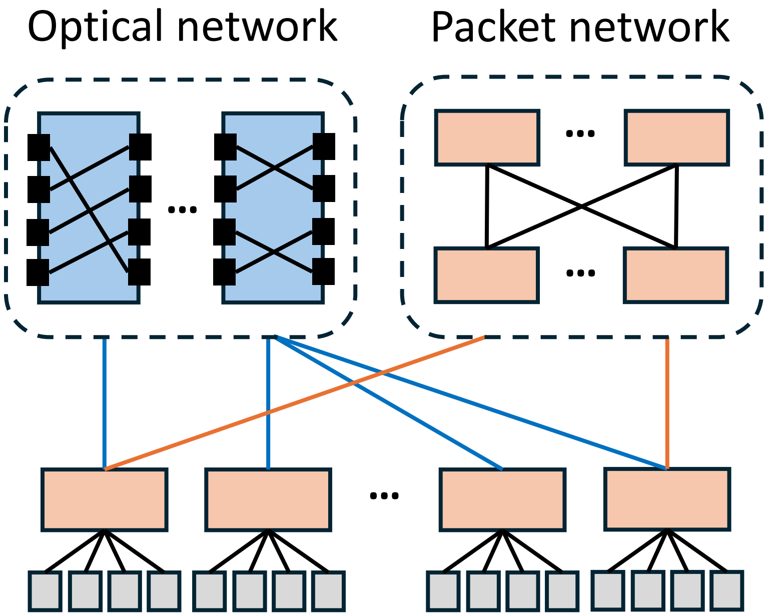


TCP’s mechanism

- Probe & converge
  - On round-trip time (RTT) scale
- Model path characteristics
  - *cwnd*: sending rate

Stability assumption breaks in demand-oblivious RDCN.

Demand-oblivious RDCN

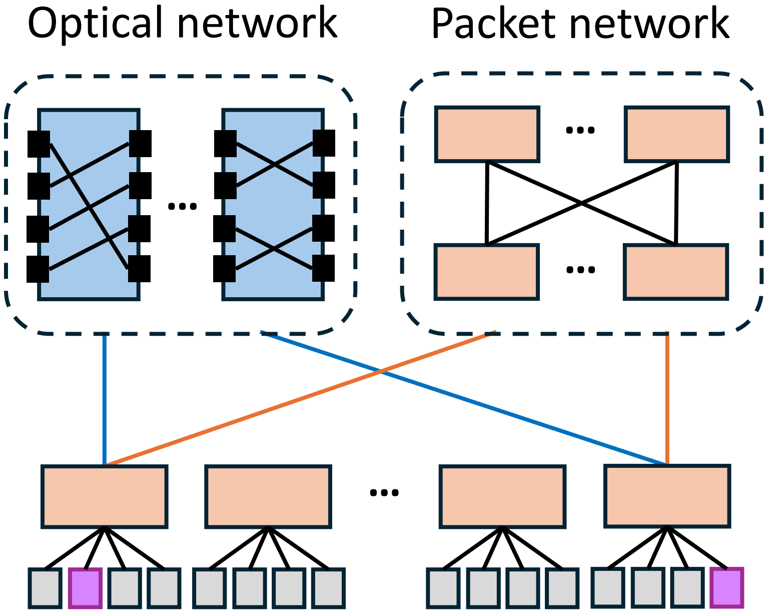


🟦 Optical circuit switch (OCS)

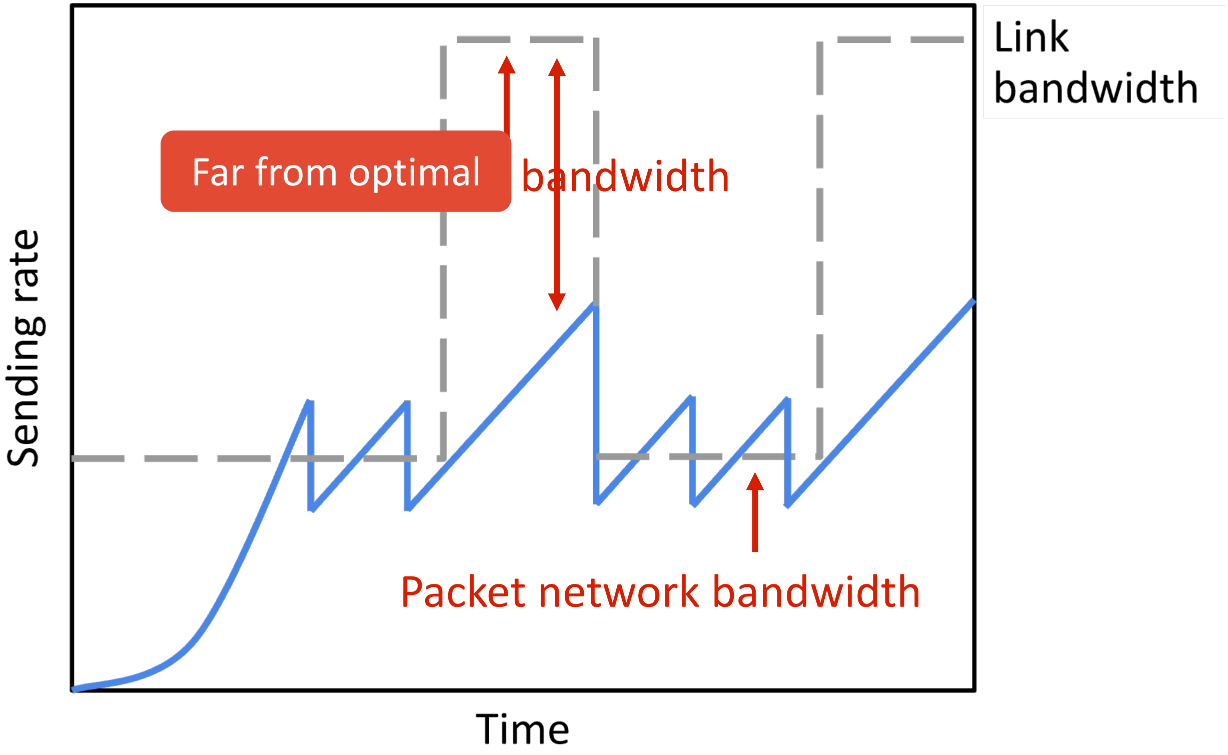
🟧 Electrical packet switch (EPS)

⬜ Host

Demand-oblivious RDCN

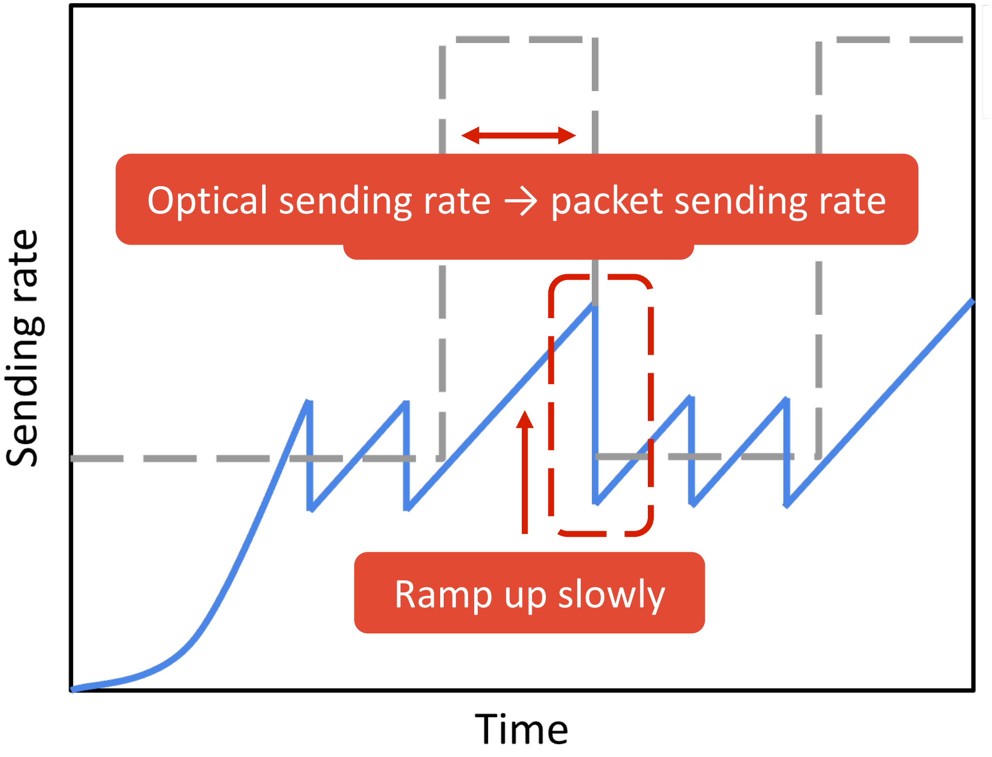


TCP performs poorly under invalid assumption



Far from optimal

TCP performs poorly under invalid assumption



What happened?

1) Reactive probing

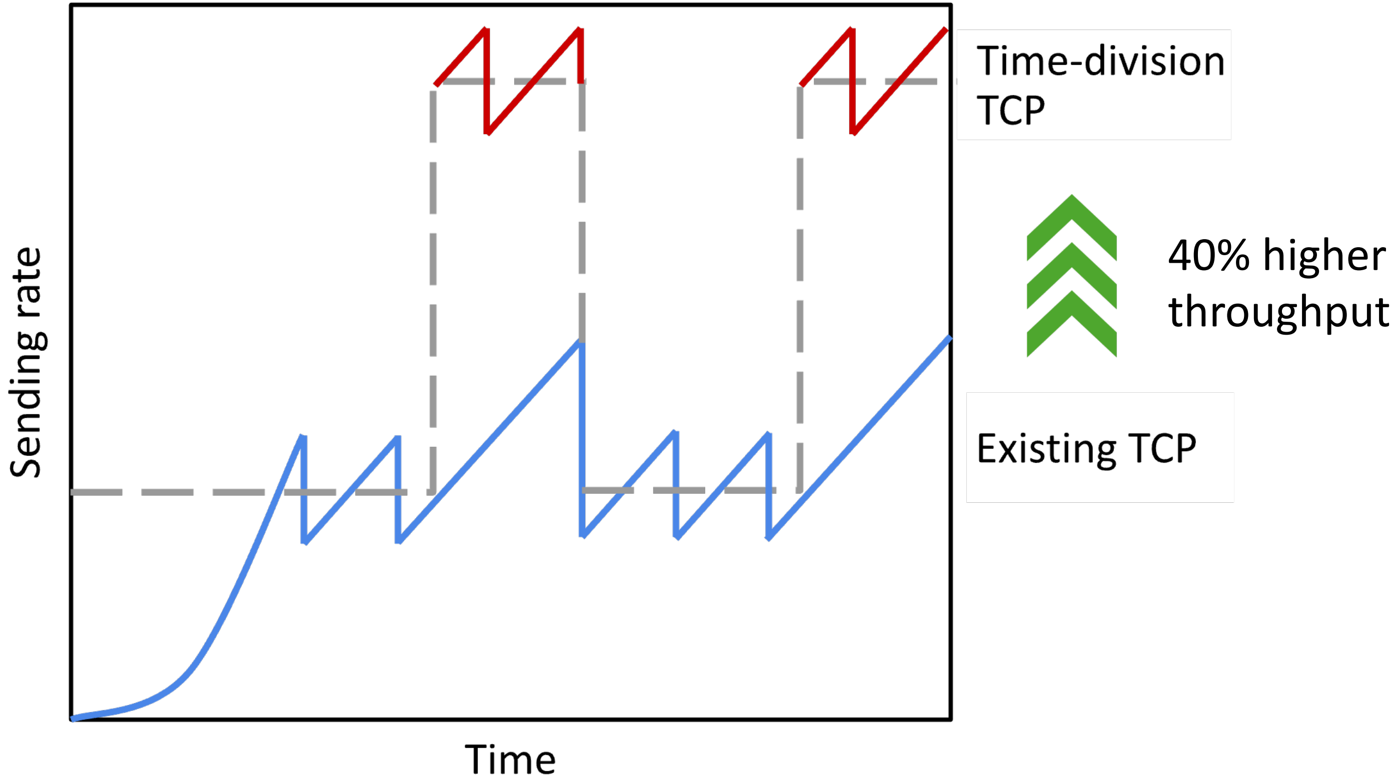
2) Insufficient time to converge

3) Overwritten states

Our proposal: Time-division TCP

	Existing TCP	Time-division TCP
Change discovery	Reactive - in-band, probing	Proactive - out-of-band, switch notification
Path modeling	One state: - *cwnd, srtt*	2 (N) states: - *cwnd[], srtt[]*

Time-division TCP outperforms existing TCP

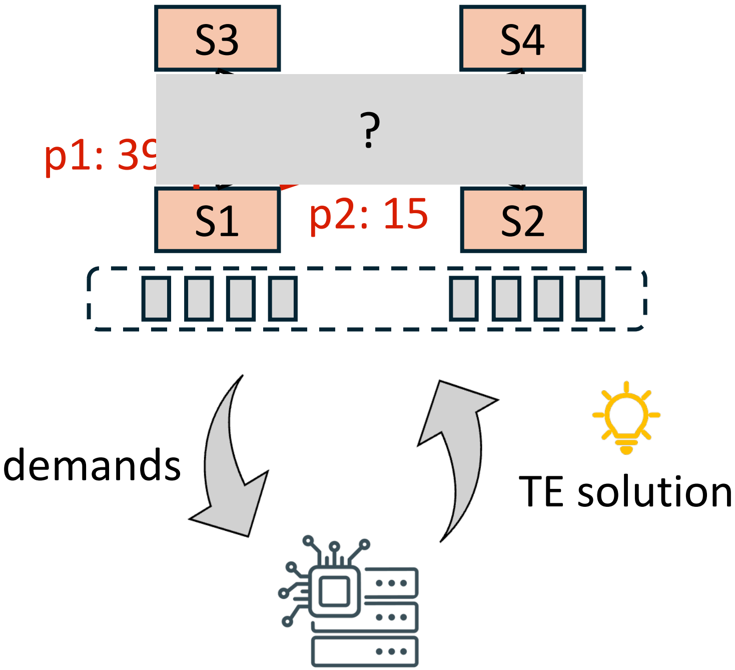


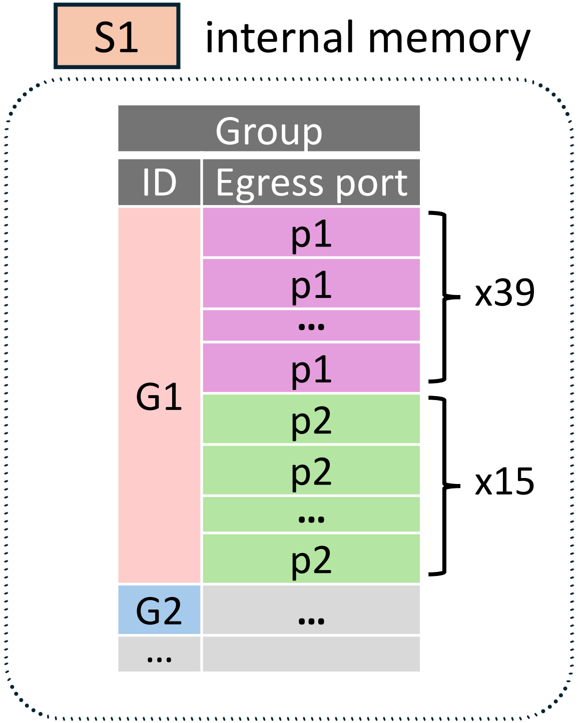
Today's talk

- **Transport**
  - Time-division TCP for demand-oblivious RDCNs *[SIGCOMM'22]*

- **Traffic engineering**
  - Precise traffic engineering for demand-aware RDCNs *[NSDI'24]*

How does a traffic engineering (TE) system work?



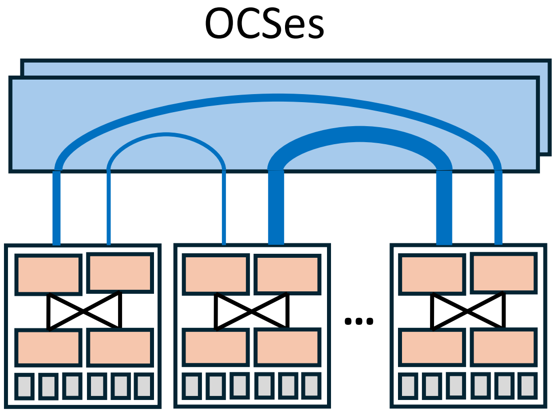


Group	
ID	Egress port
G1	p1
	p1
	…
	p1
	p2
	p2
	…
	p2
G2	…
…	

Demand-aware RDCN

Demand-awareness introduces heterogeneity

- Skewed traffic distribution (weight ratios)
- Ratios have different actual impact on traffic

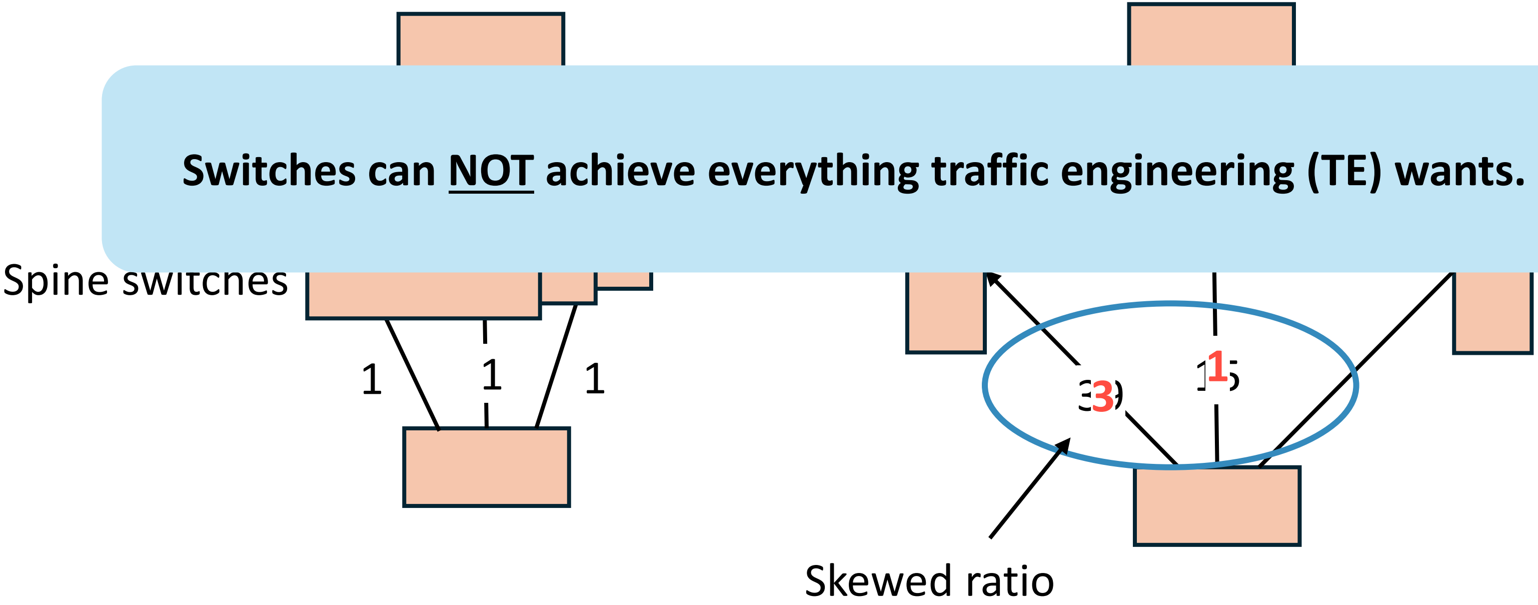


🟦 Optical circuit switch (OCS)

🟧 Electrical packet switch (EPS)

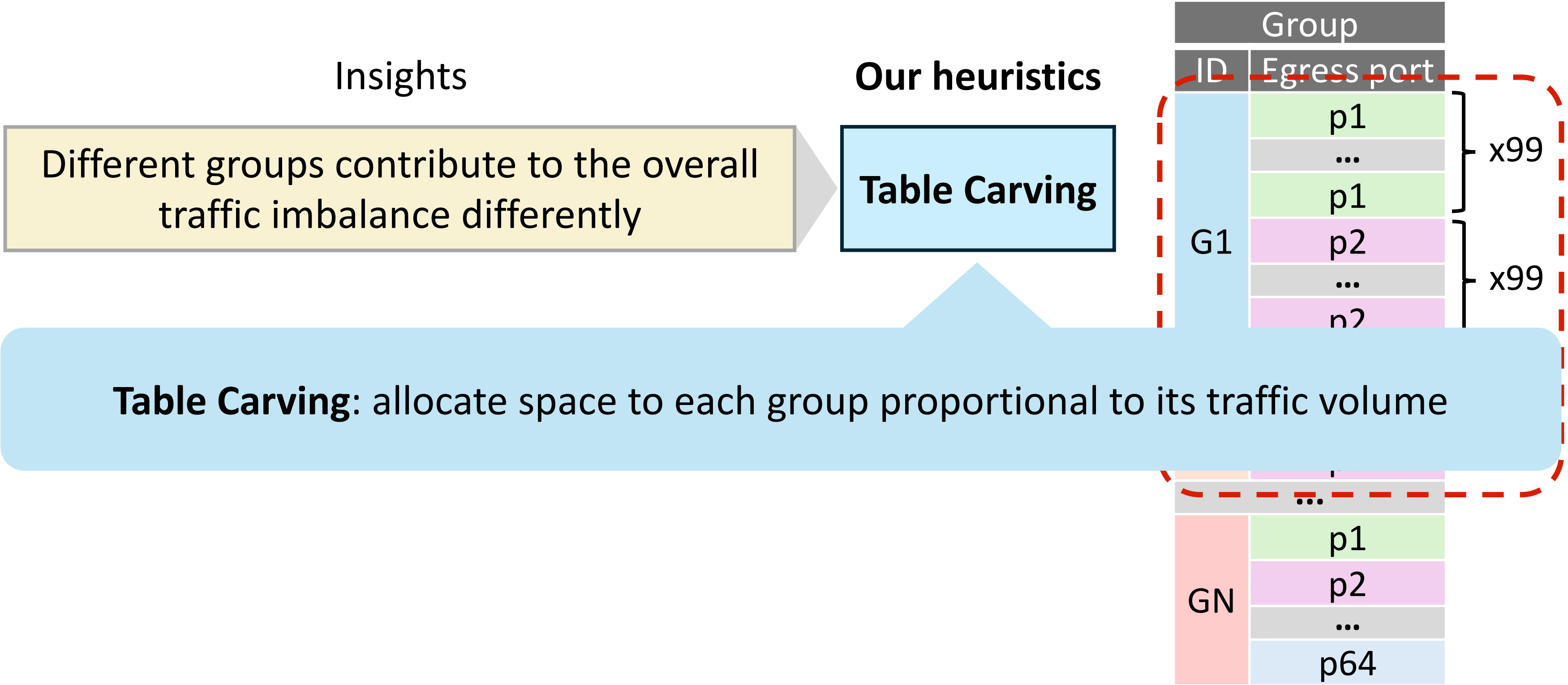
⬜ Host

Traffic engineering’s assumption: omnipotent switches



Switches can <u>NOT</u> achieve everything traffic engineering (TE) wants.

Heuristics to reduce group space usage



Insights

Different groups contribute to the overall traffic imbalance differently

Our heuristics

Table Carving

Table Carving: allocate space to each group proportional to its traffic volume

Heuristics to reduce group space usage

Insights

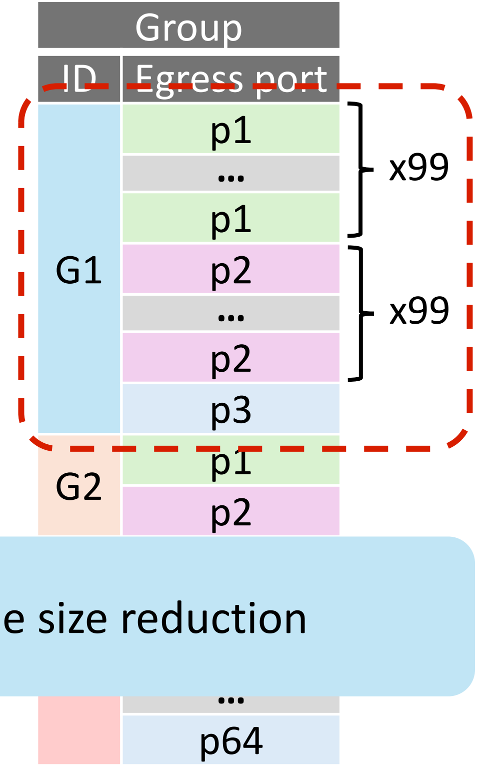
Different groups contribute to the overall traffic imbalance differently

Not all ports need to be preserved

Our heuristics

Table Carving

Group Pruning

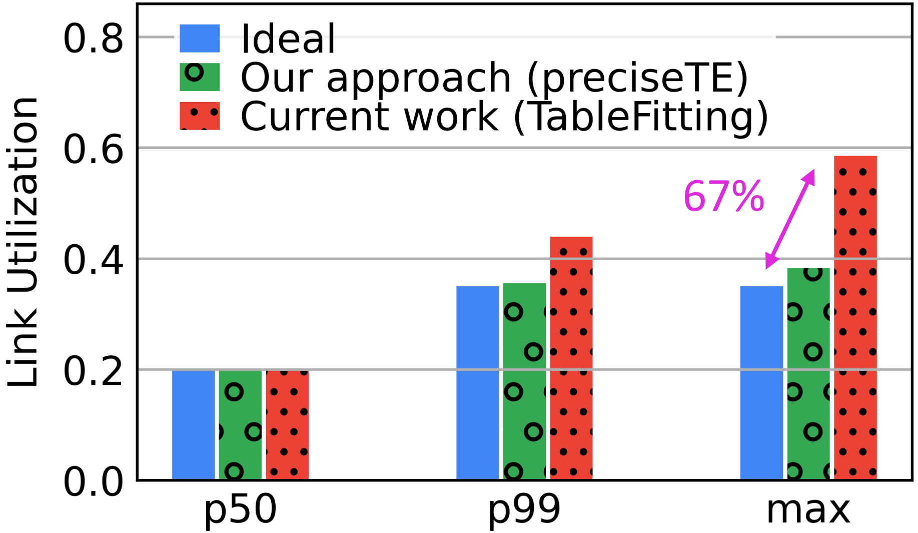


Group	
ID	Egress port
G1	p1
	…
	p1
	p2
	…
	p2
	p3
G2	p1
	p2
	…
	p64

Group Pruning: prune select ports from a group to enable size reduction

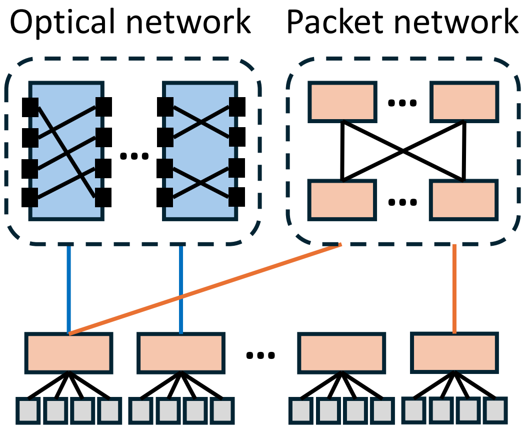
Our approach is more precise than current work.

- preciseTE **7% error** vs. TableFitting **67% error**
- Being deployed at Google



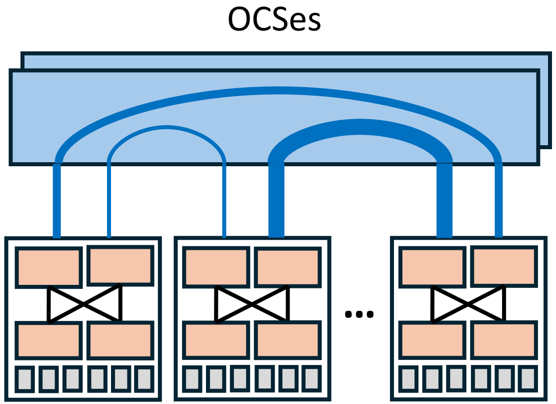
Summary

Demand-oblivious RDCN



Transport: coordination

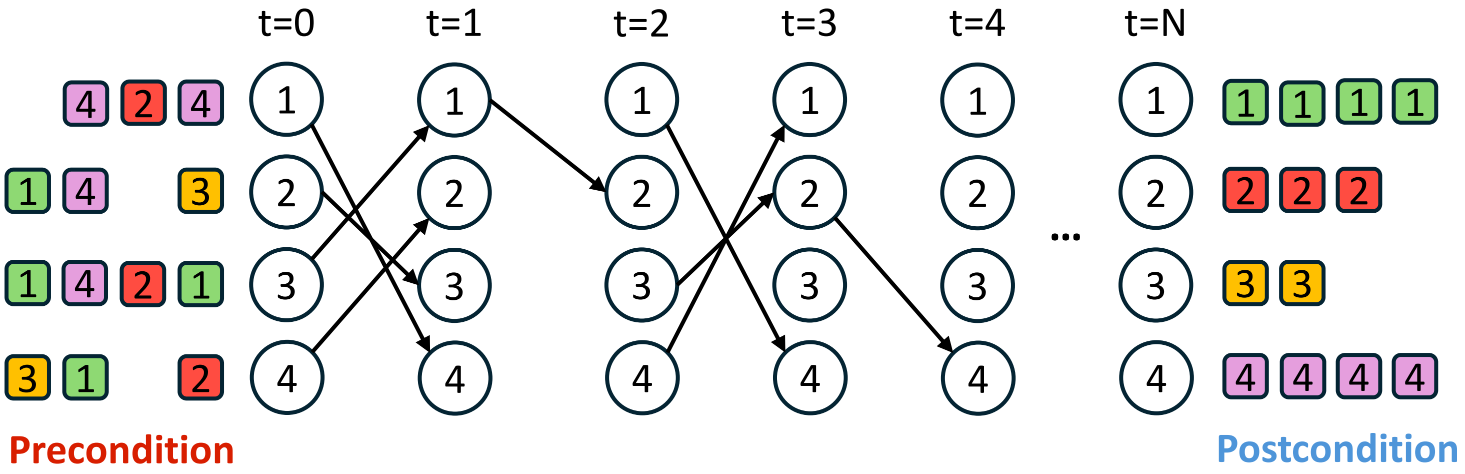
Demand-aware RDCN



TE: managing heterogeneity

Future direction: All-optical RDCN

- Fast OCS (optical packet switching)
- Fully scheduled, source-routed network
- Scheduling challenge
  - “Incast” avoidance


t=0
t=1
t=2
t=3
t=4
t=N
4 2 4
1 4 3
1 4 2 1
3 1 2
1
2
3
4
...
1 1 1 1
2 2 2
3 3
4 4 4 4
Precondition
Postcondition