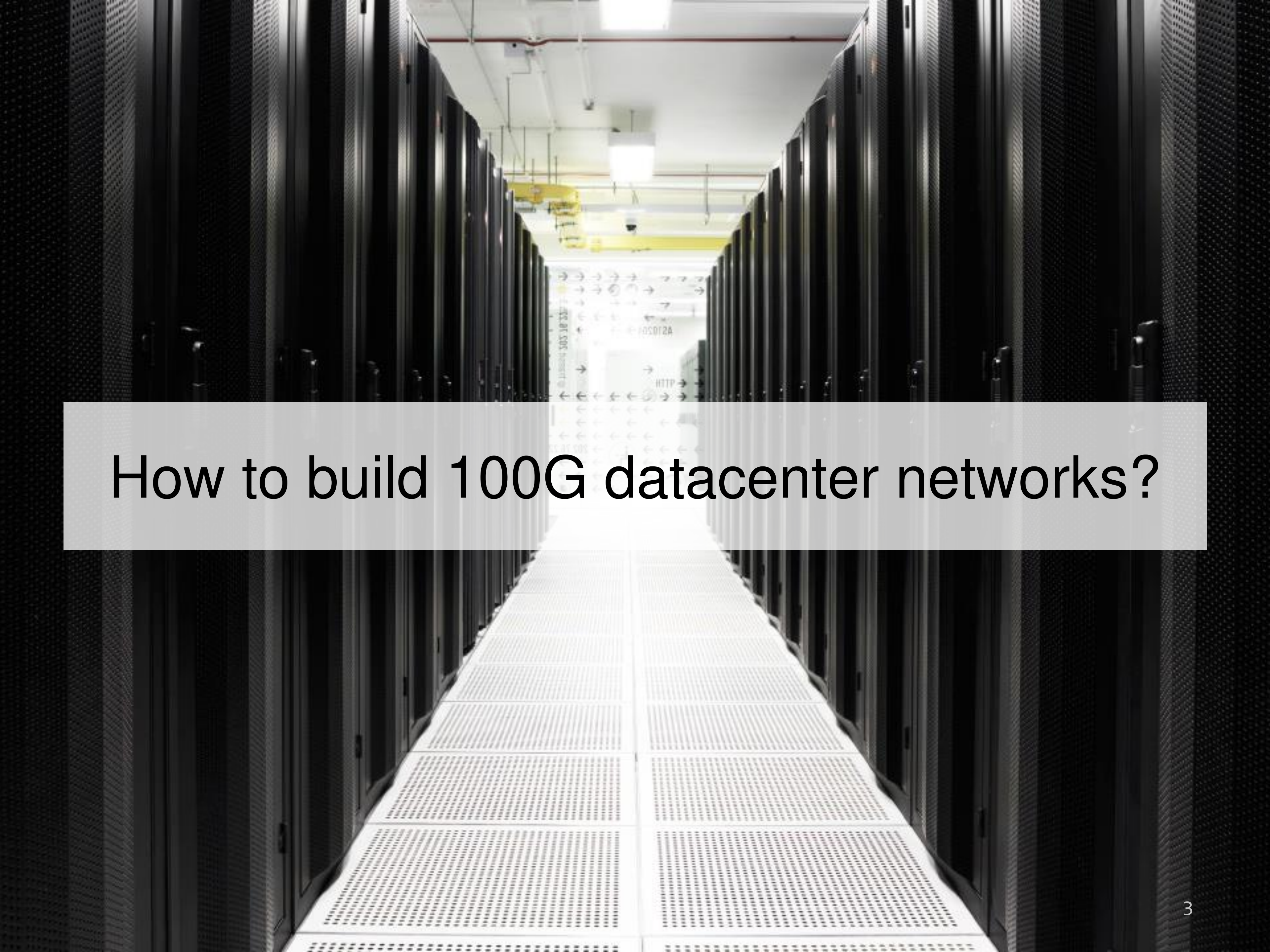
Circuit Switching Under the Radar with REACToR

He Liu, Feng Lu, Alex Forencich, Rishi Kapoor
Malveeka Tewari, Geoffrey M. Voelker
George Papen, Alex C. Snoeren, George Porter

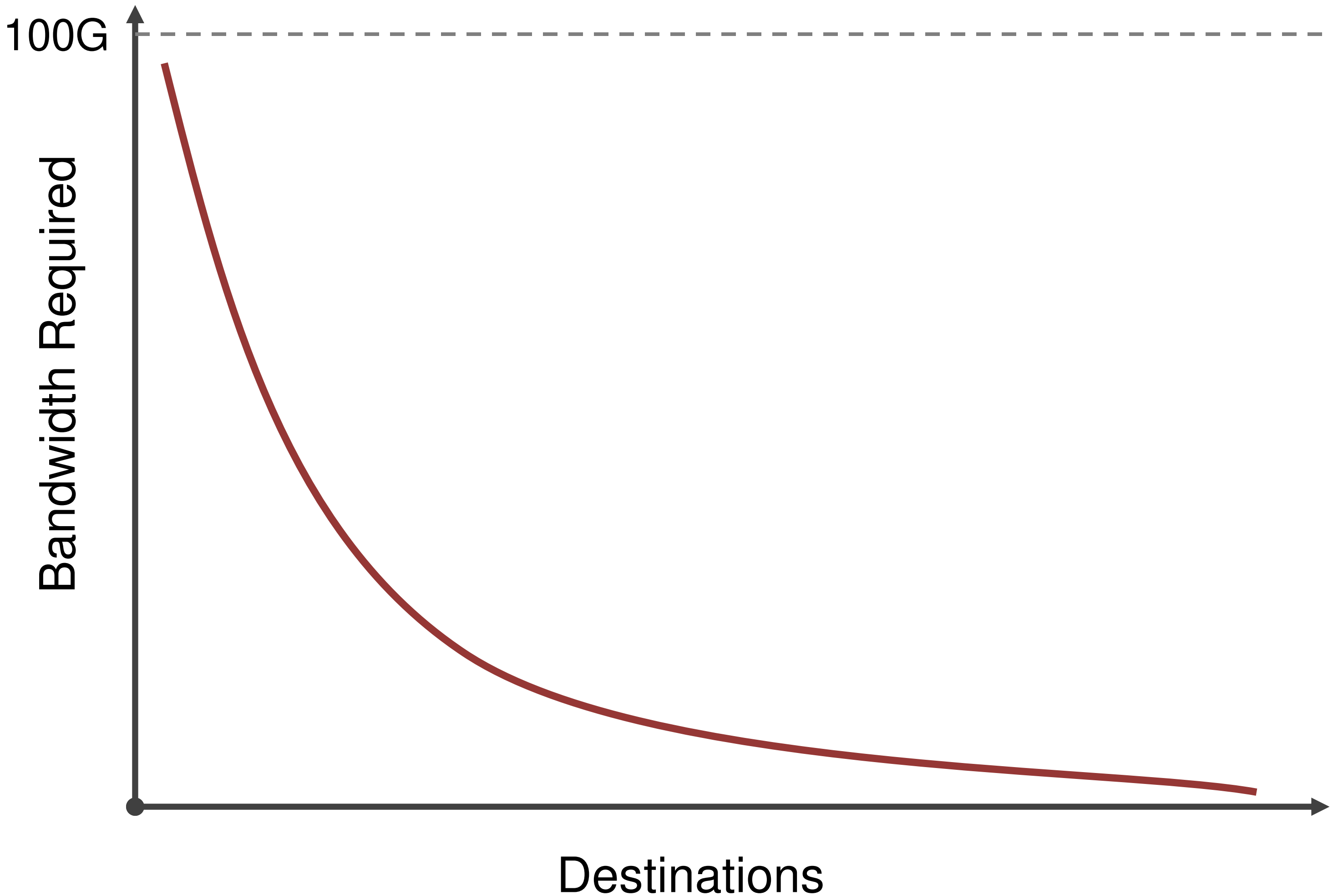
UCSDCSE
Computer Science and Engineering


HTTP

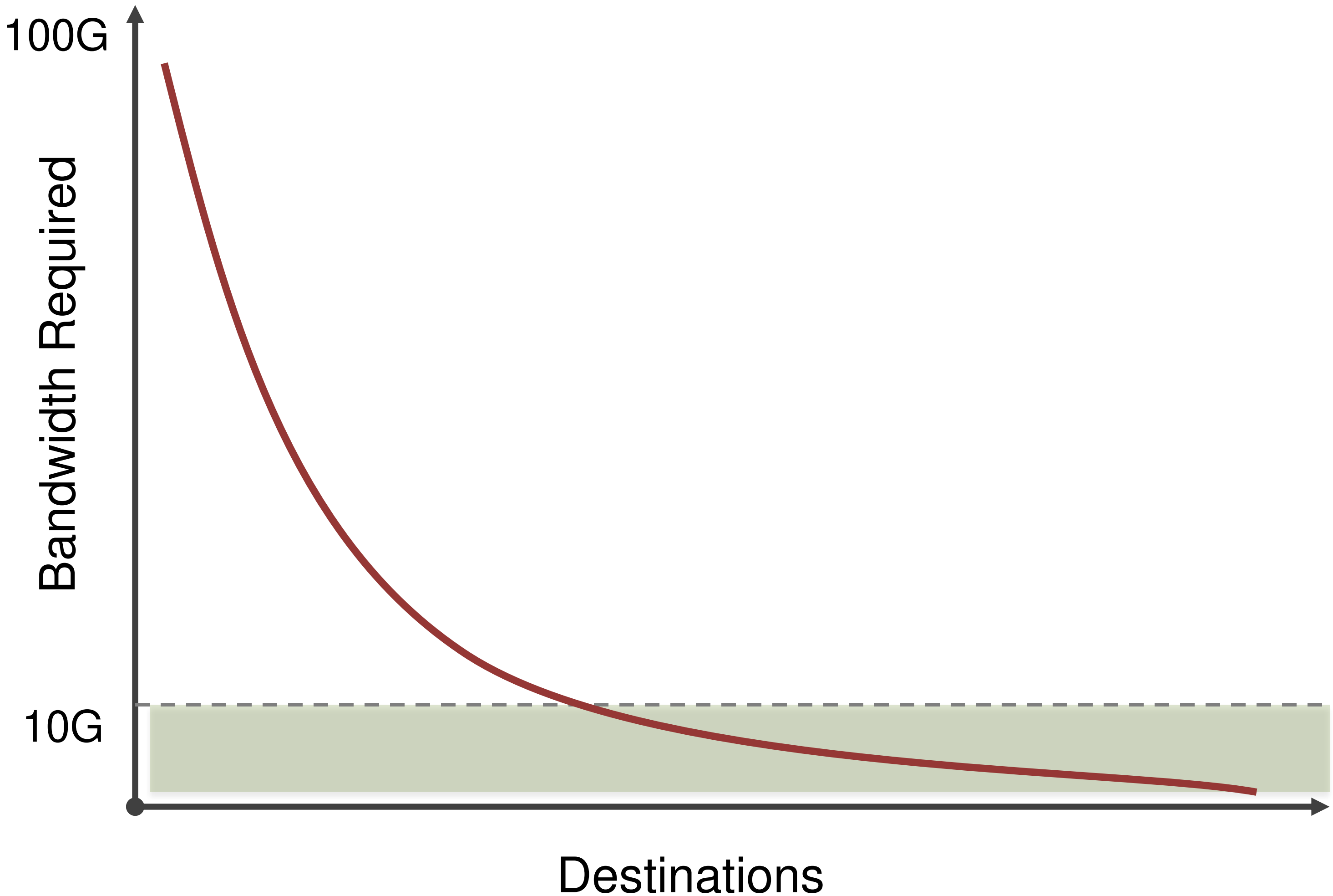
How to build 100G datacenter networks?



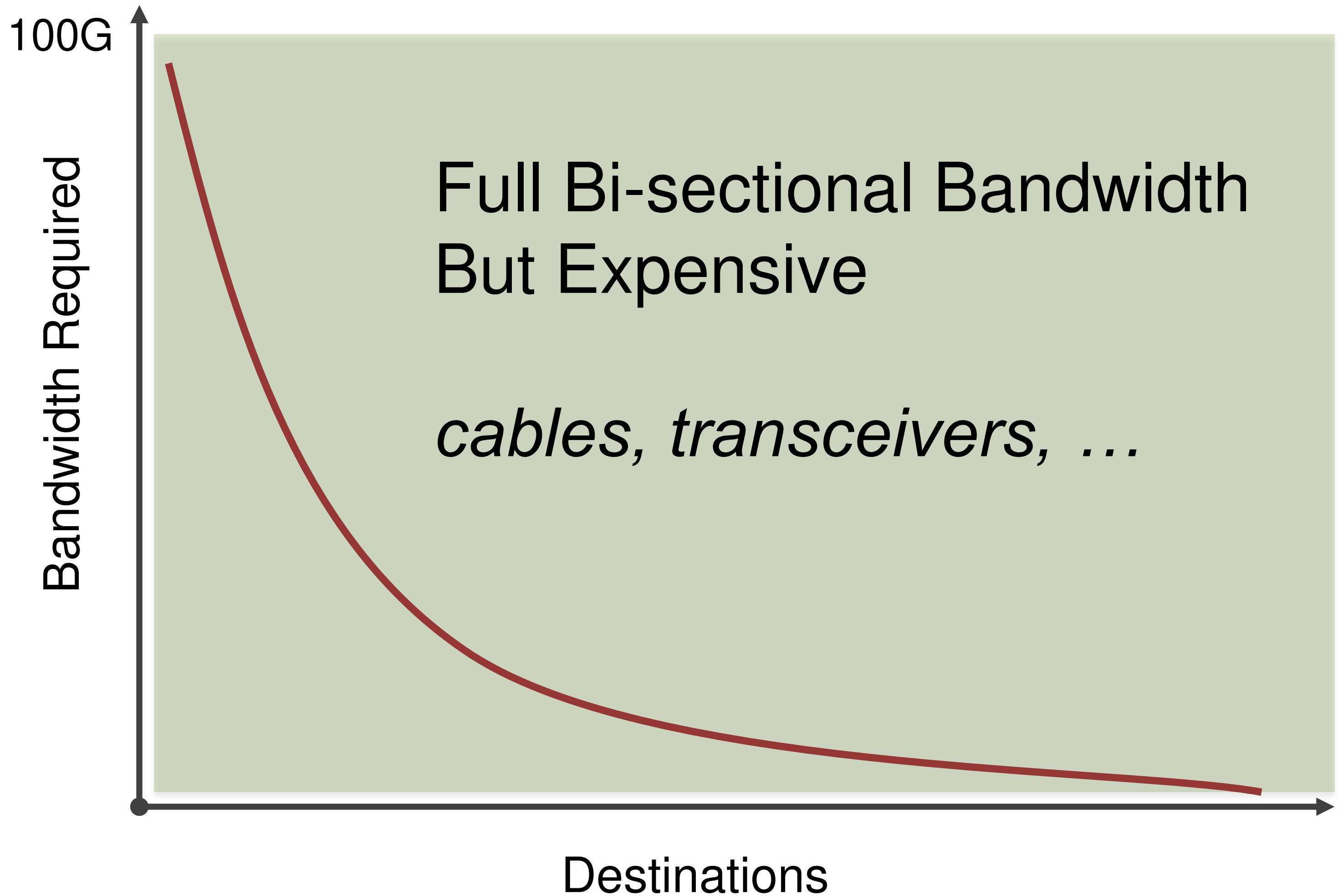
Datacenters Traffic Is Skewed



10G Fat-Tree

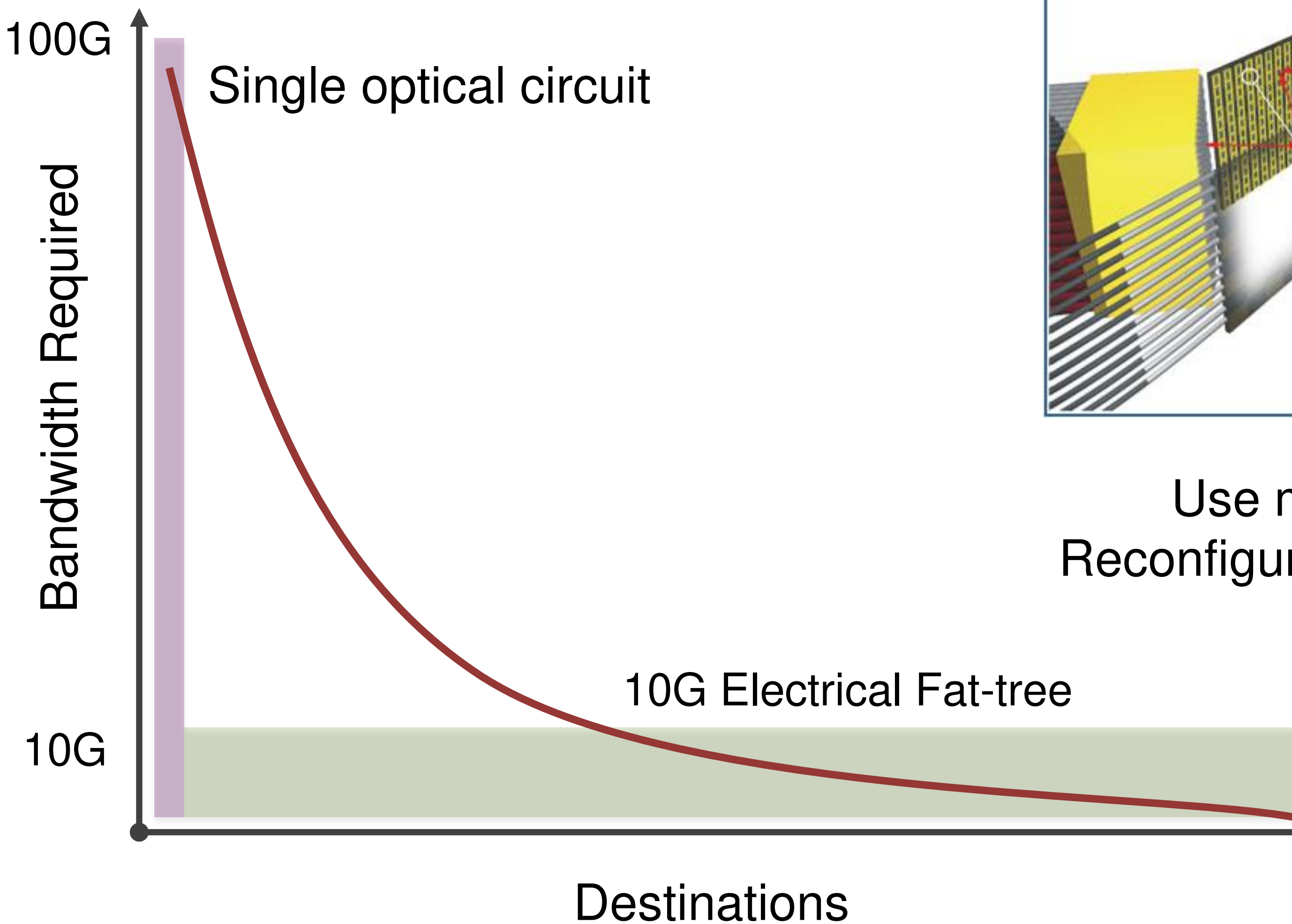


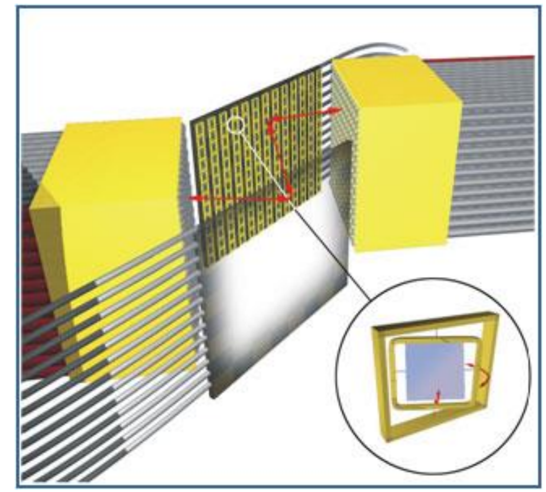
100G Fat-Tree



[SIGCOMM 2010]

Helios, c-Through: Hotspot Circuits

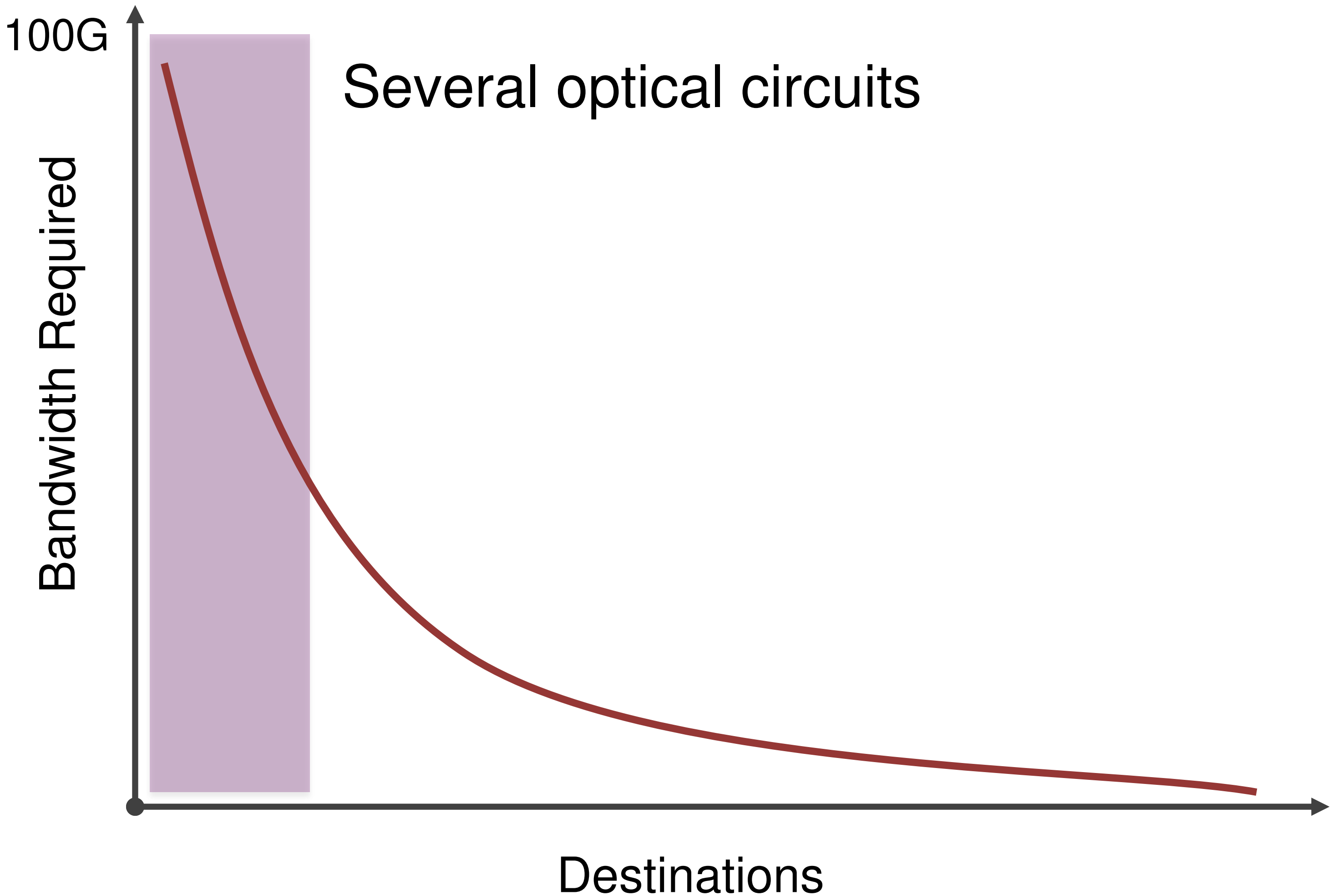




Use mirrors
Reconfigures in 10ms

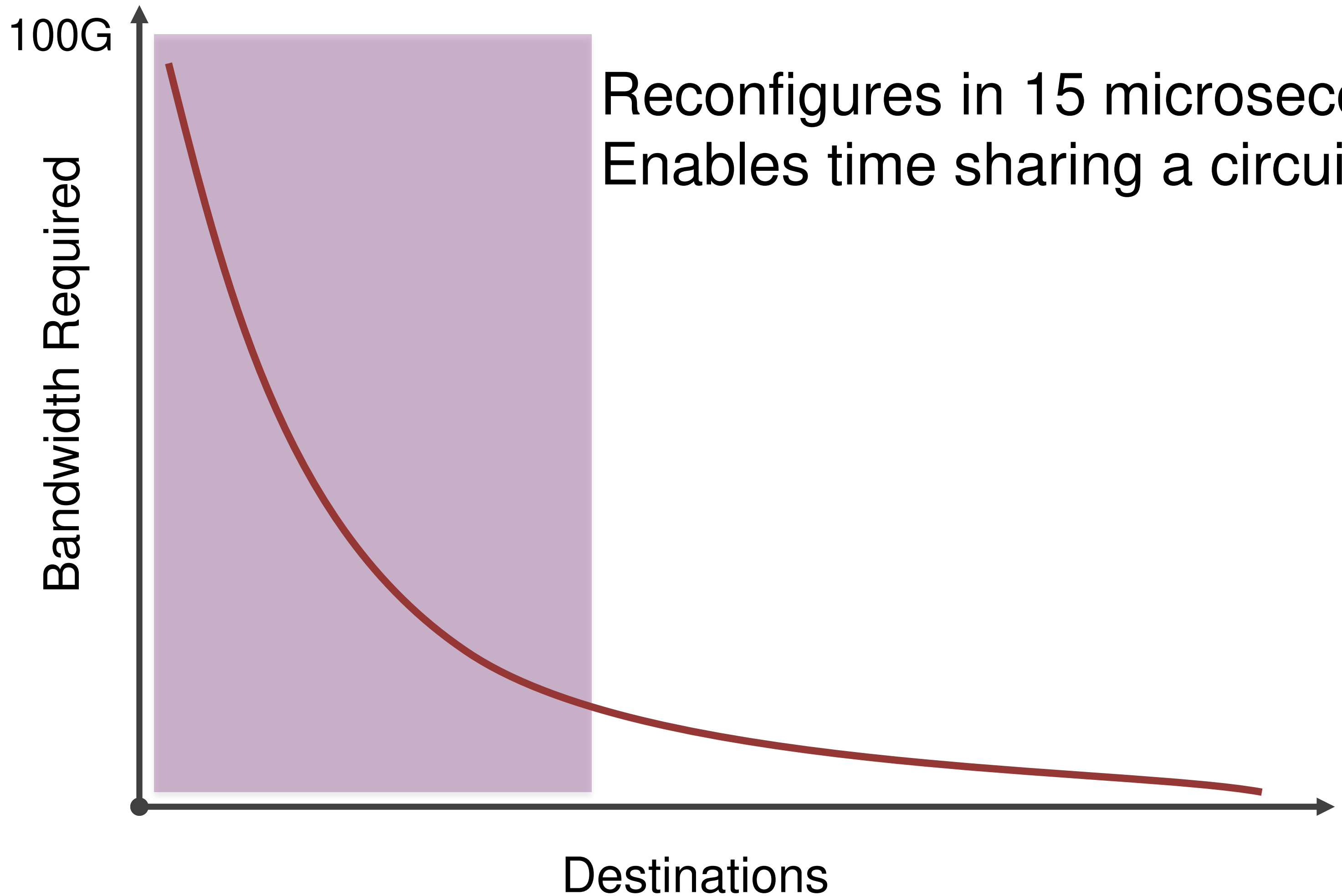
[NSDI 2012]

OSA: More Circuits



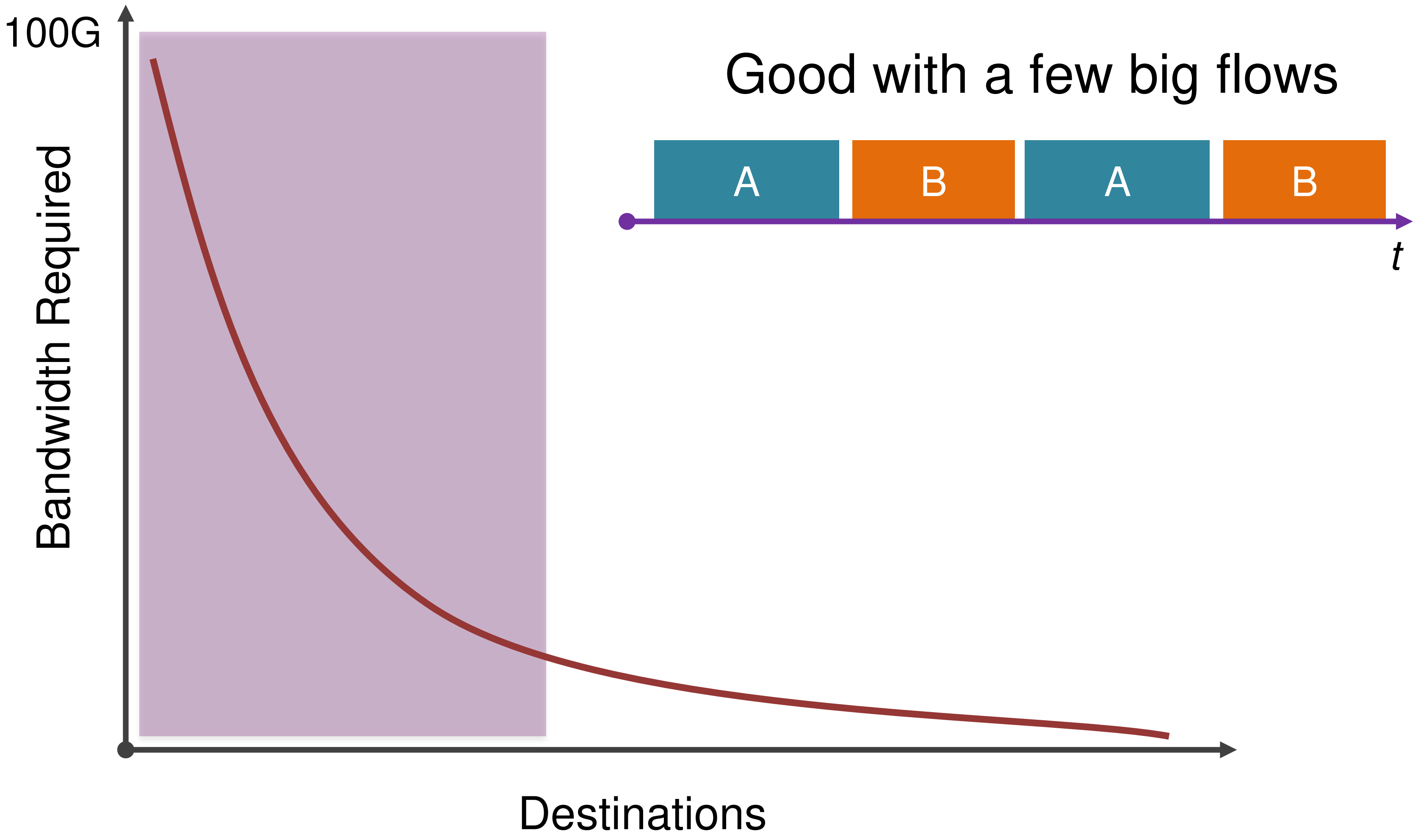
[SIGCOMM 2013]

Mordia: Fast Circuit Switching

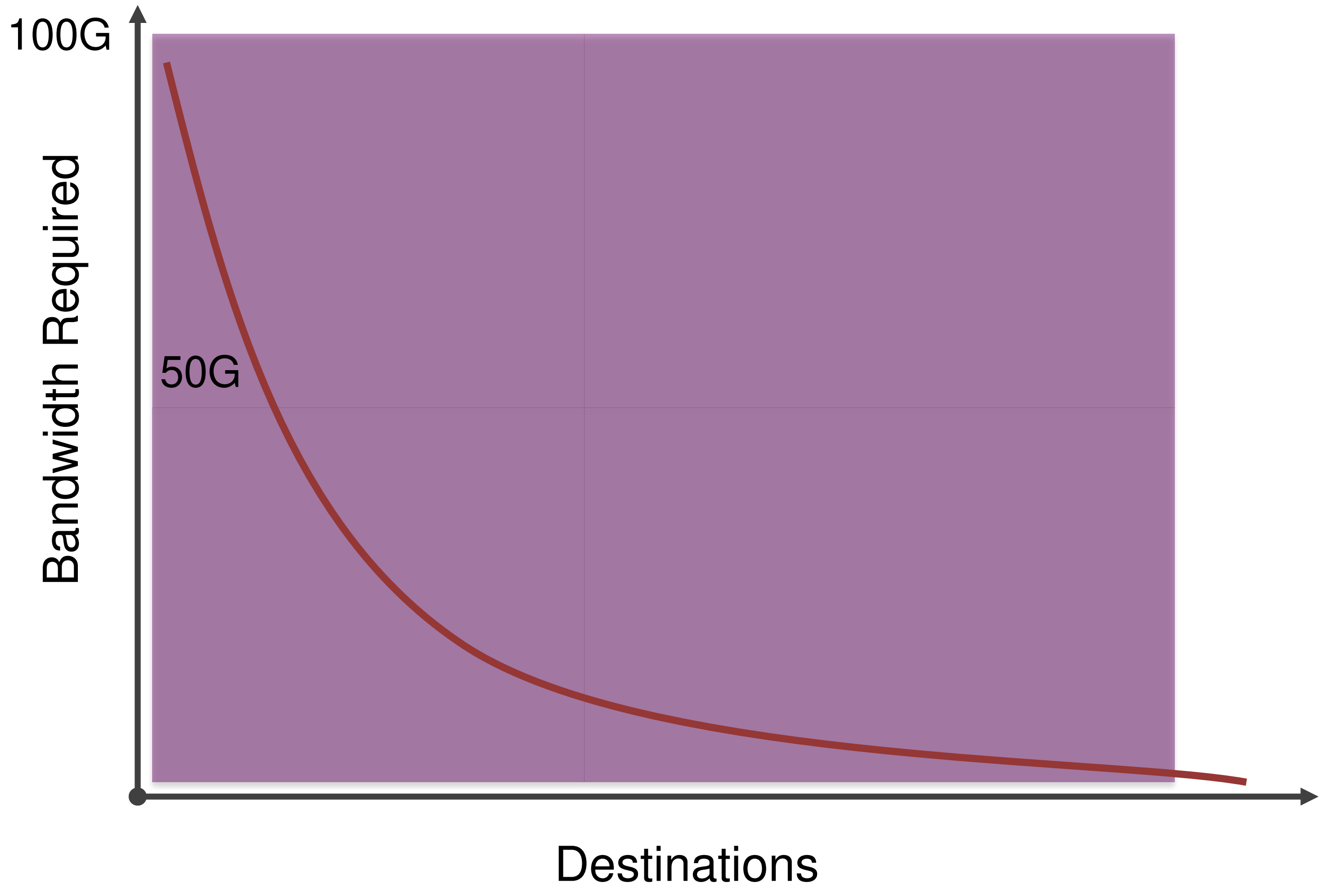


Reconfigures in 15 microseconds.
Enables time sharing a circuit.

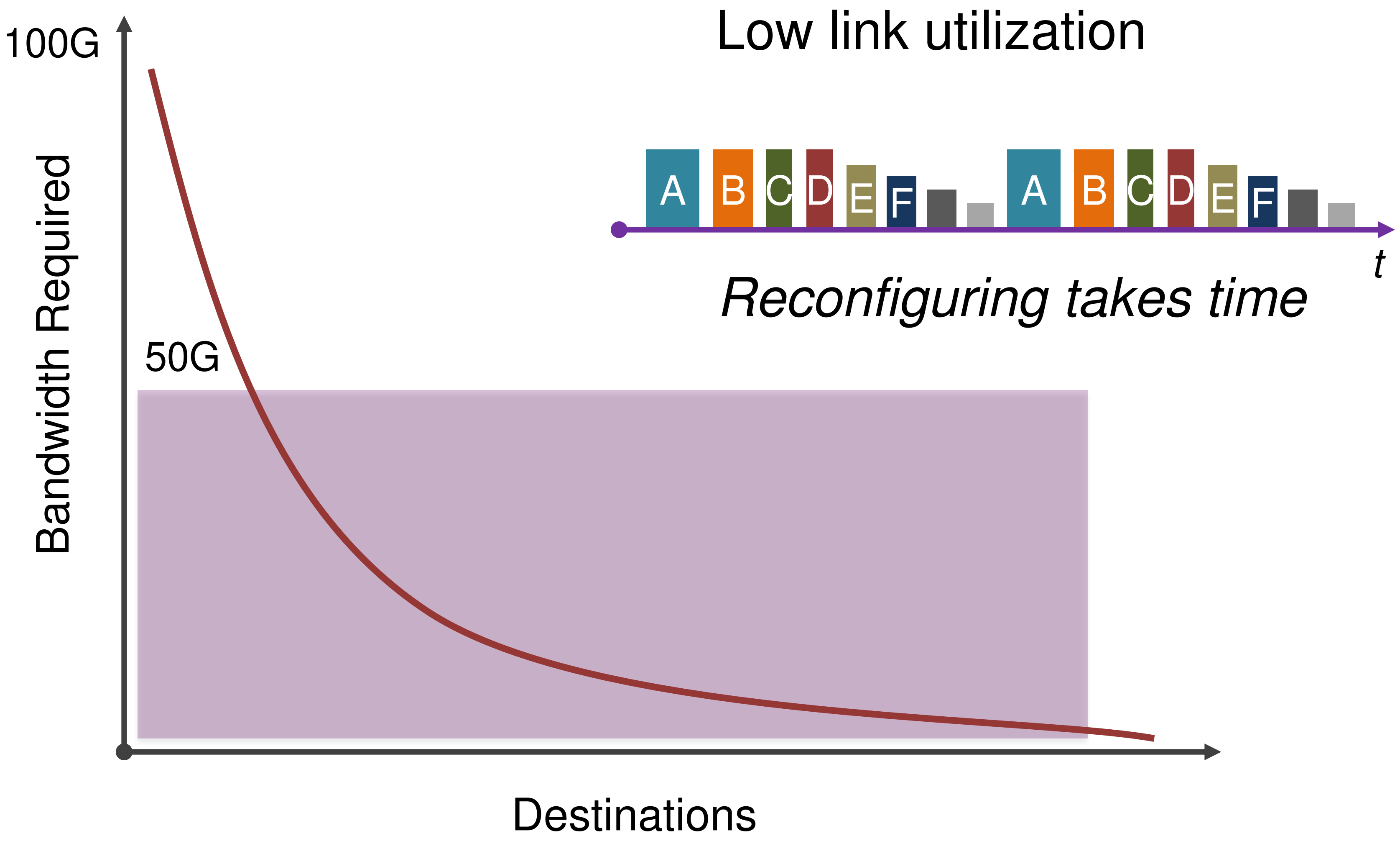
Limitation: Still Circuit Switching



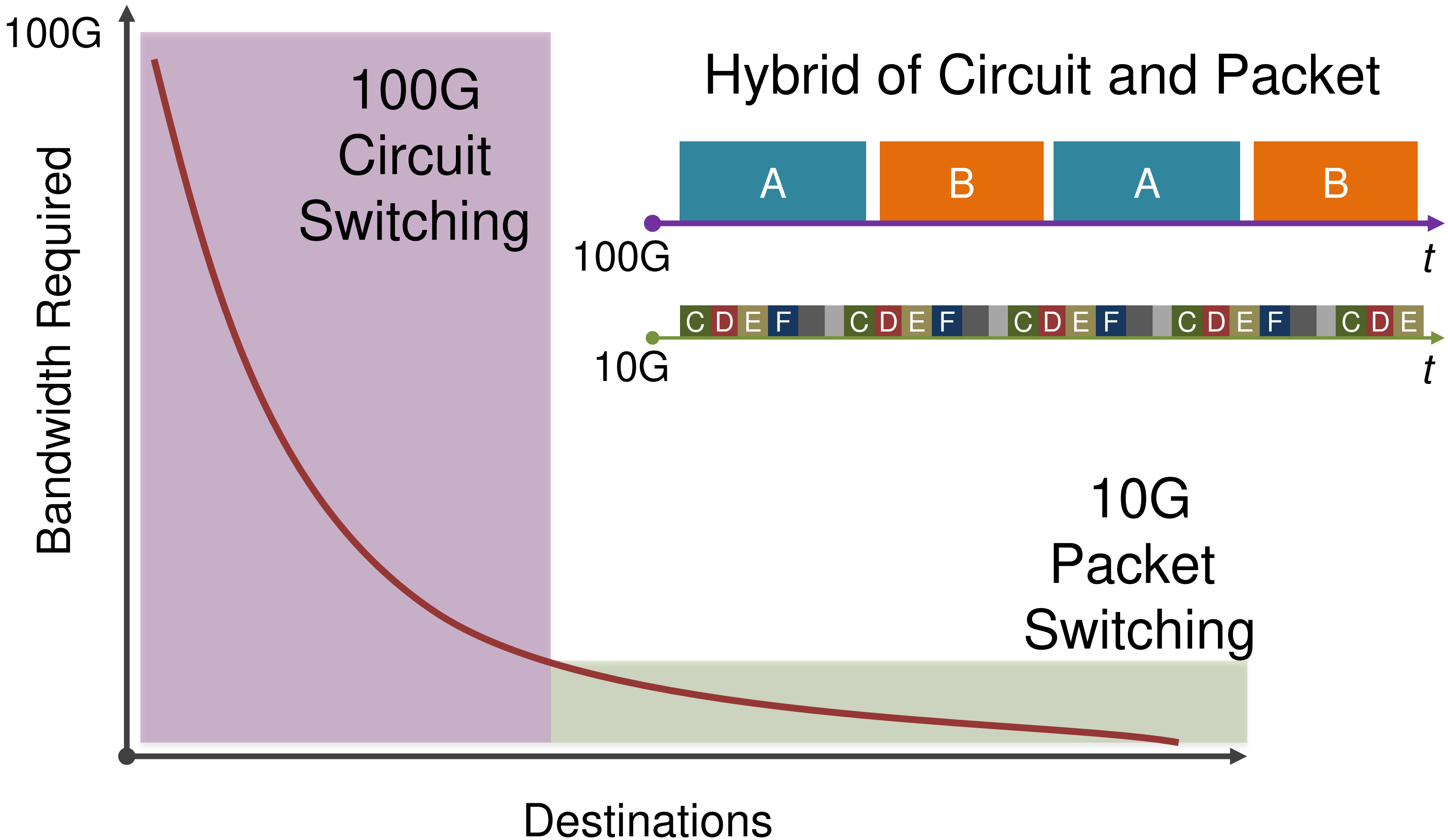
Limitation: Still Circuit Switching



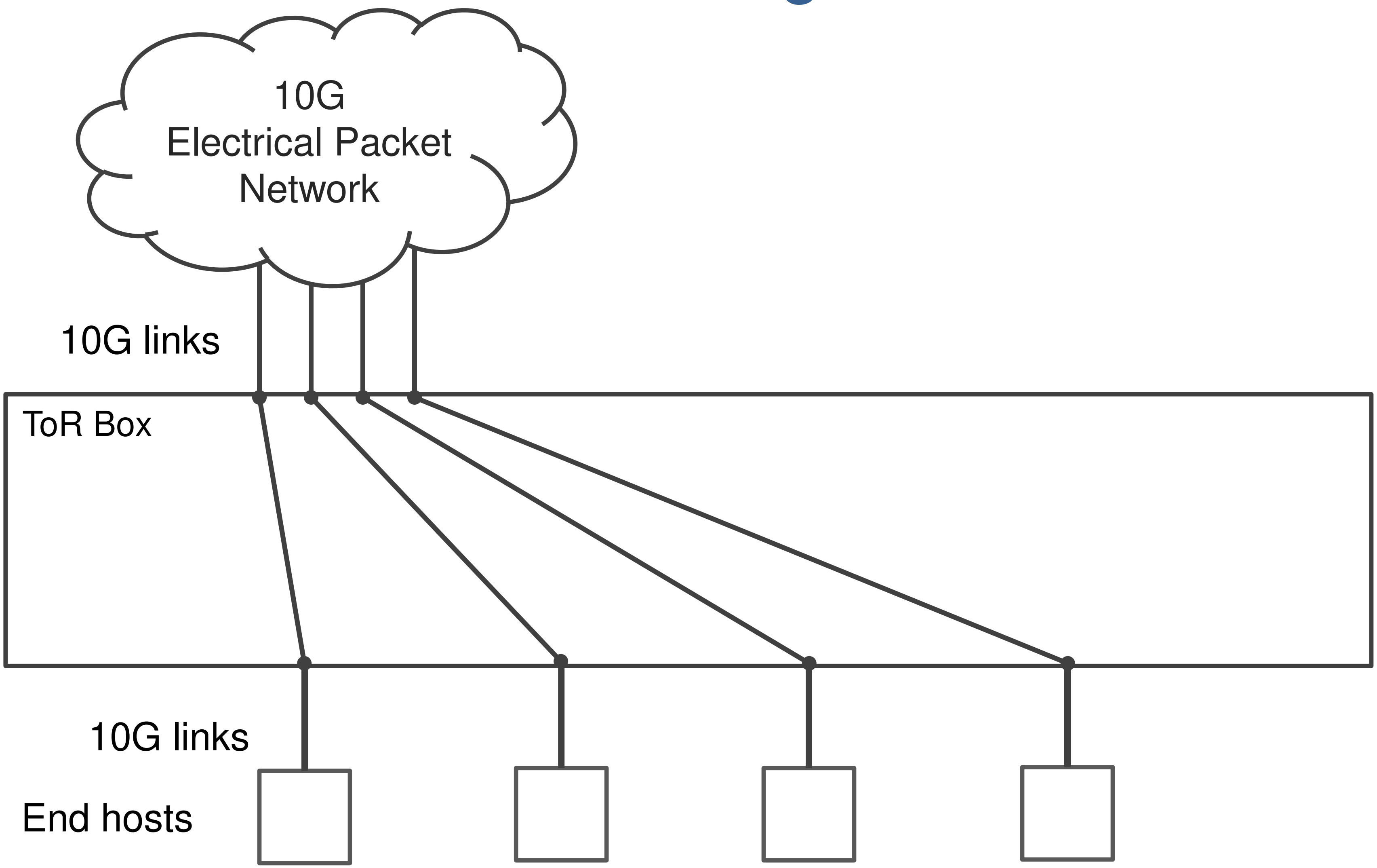
Limitation: Inefficient with Small Flows



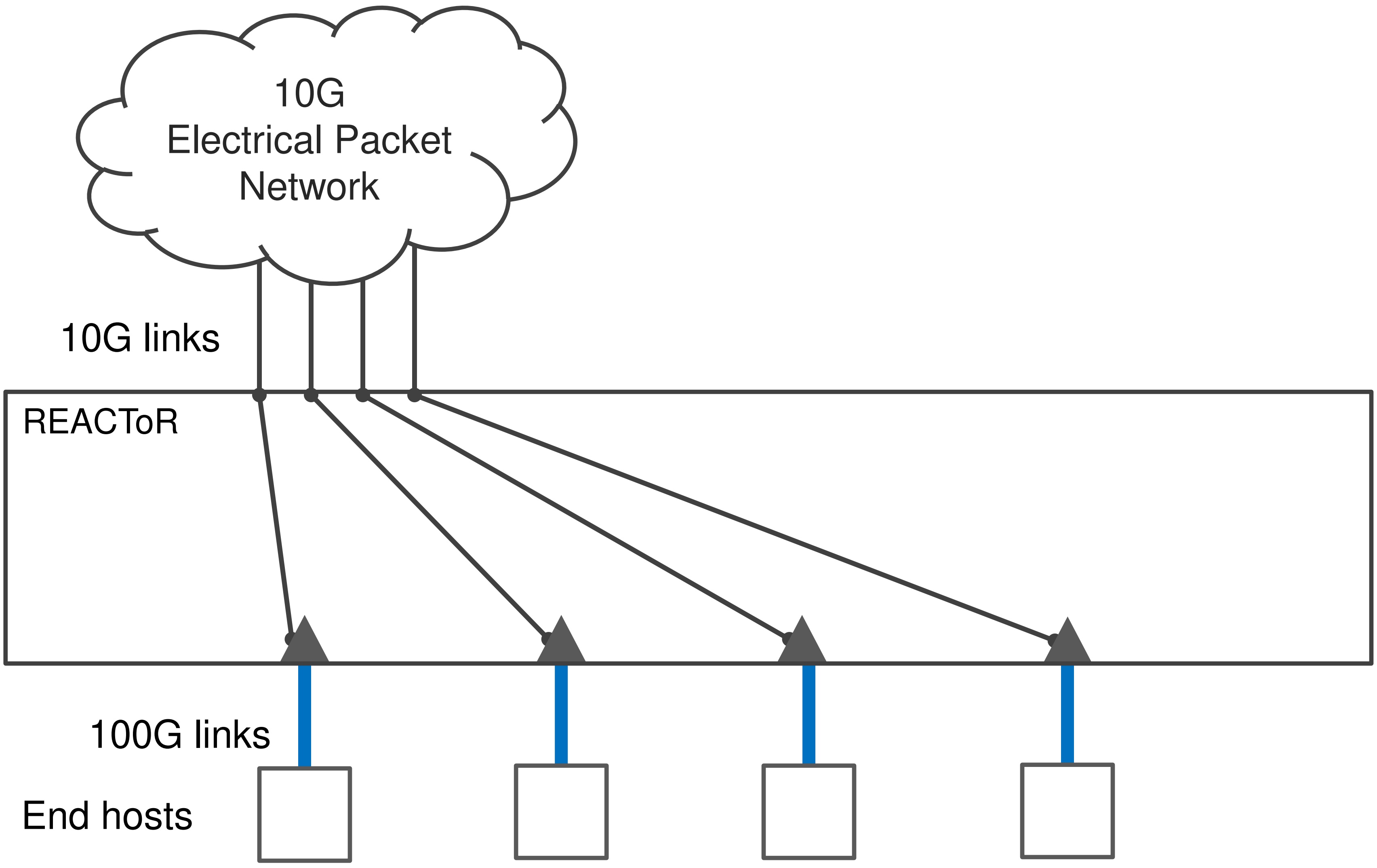
Our Approach: REACToR



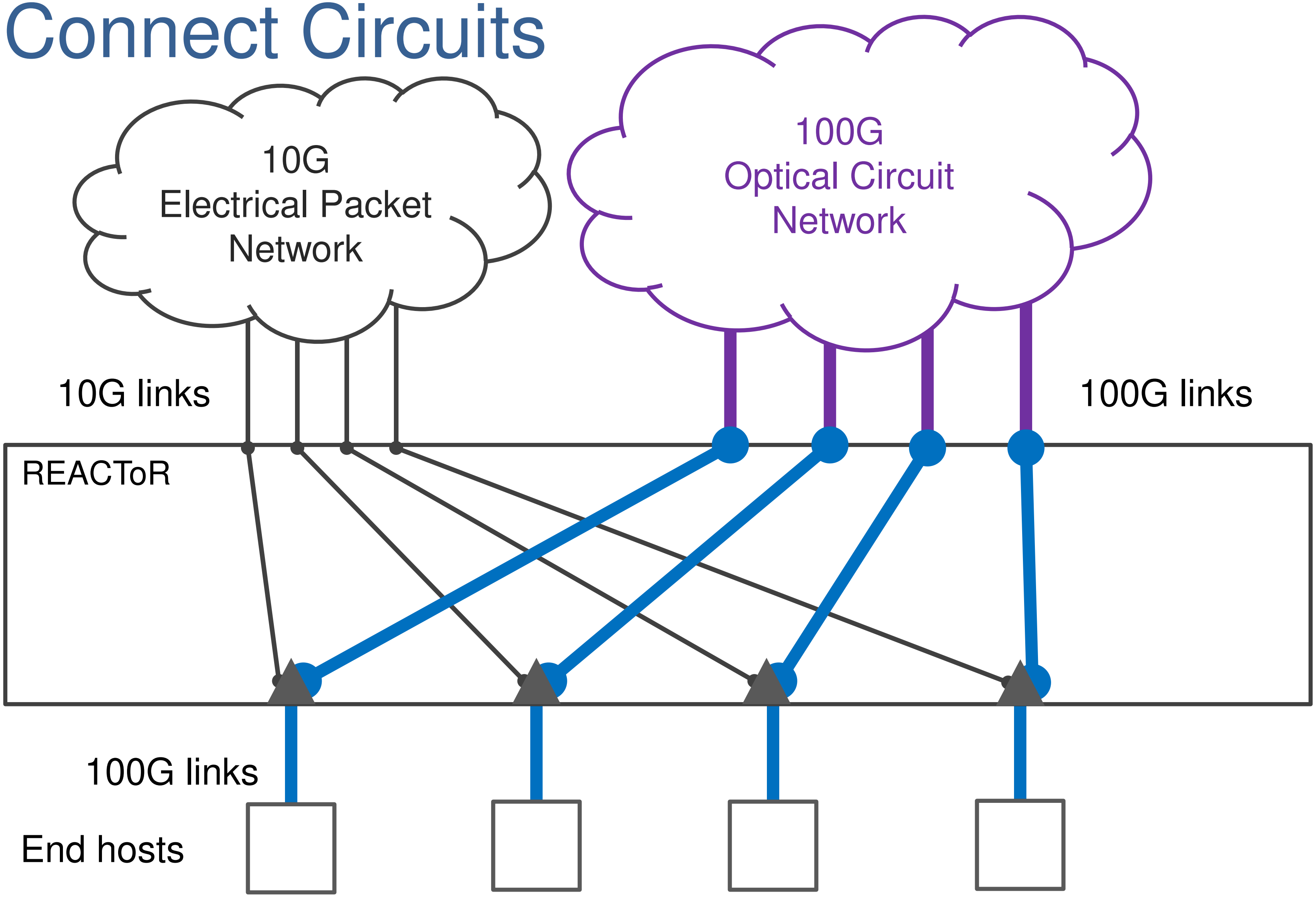
Start with a Pre-existing 10G Network



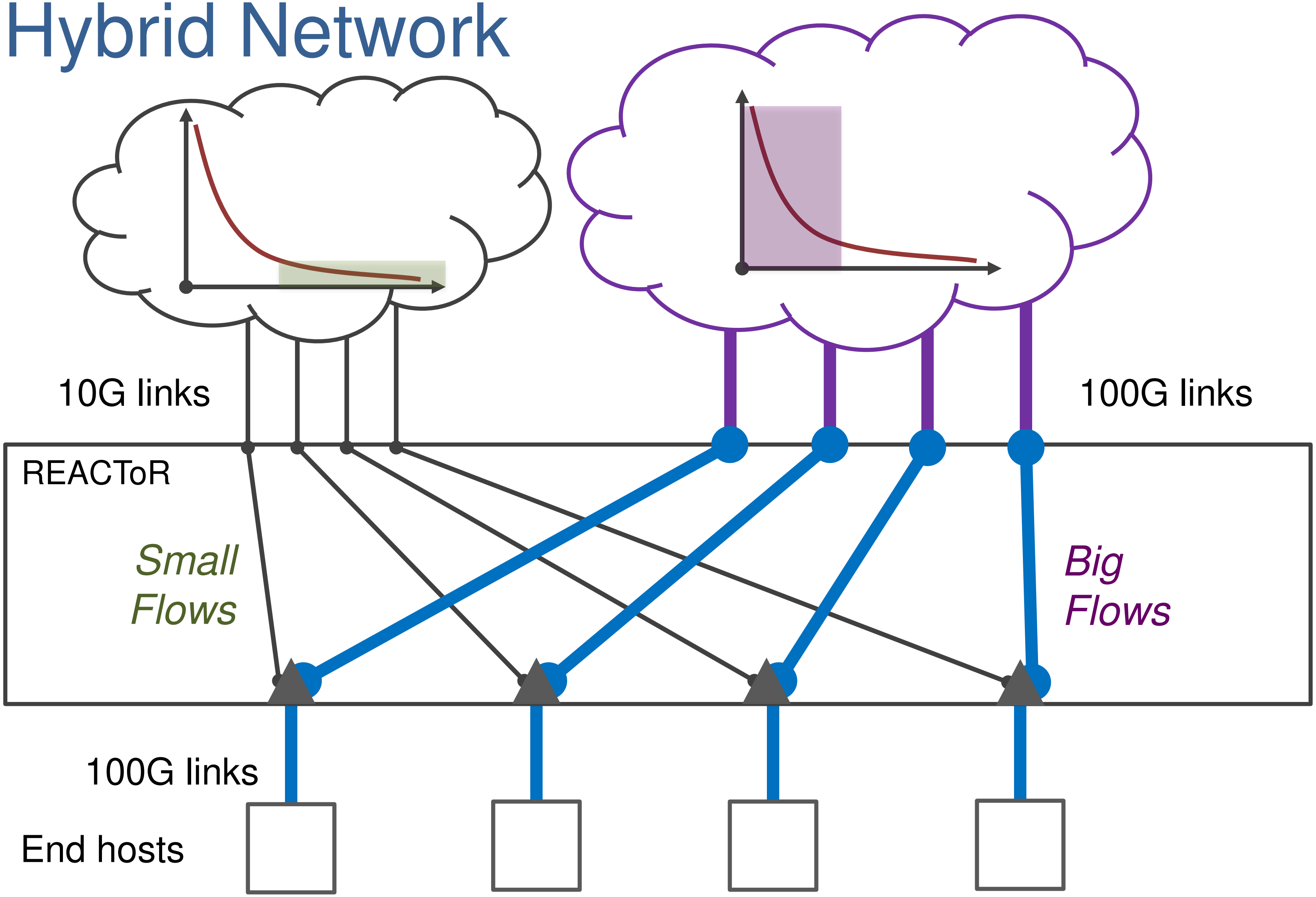
Connect via REACToR



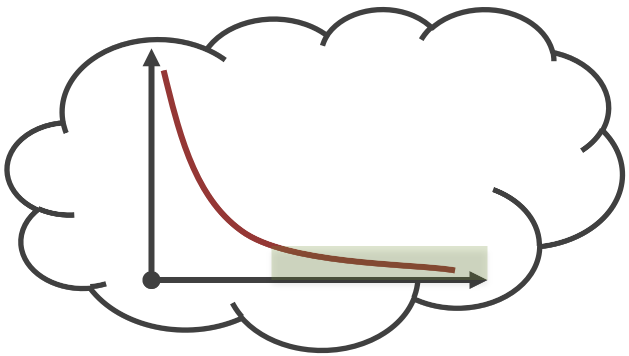
Connect Circuits

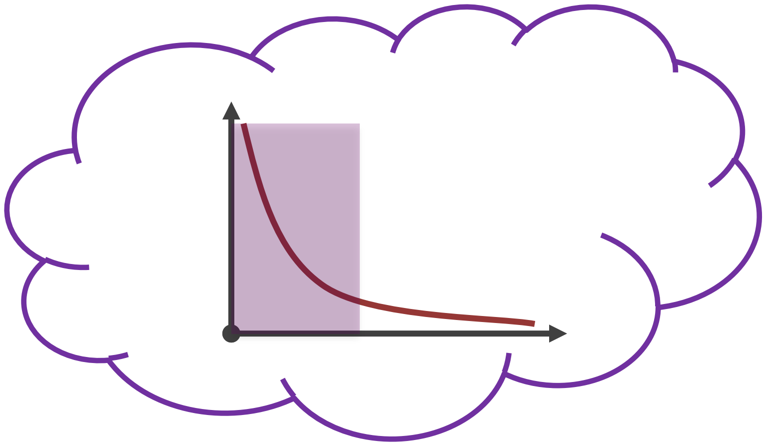


Hybrid Network



Challenge: Two Different Networks





Electrical Packet

- Low bandwidth
- Buffers all the way
- Tx at any time

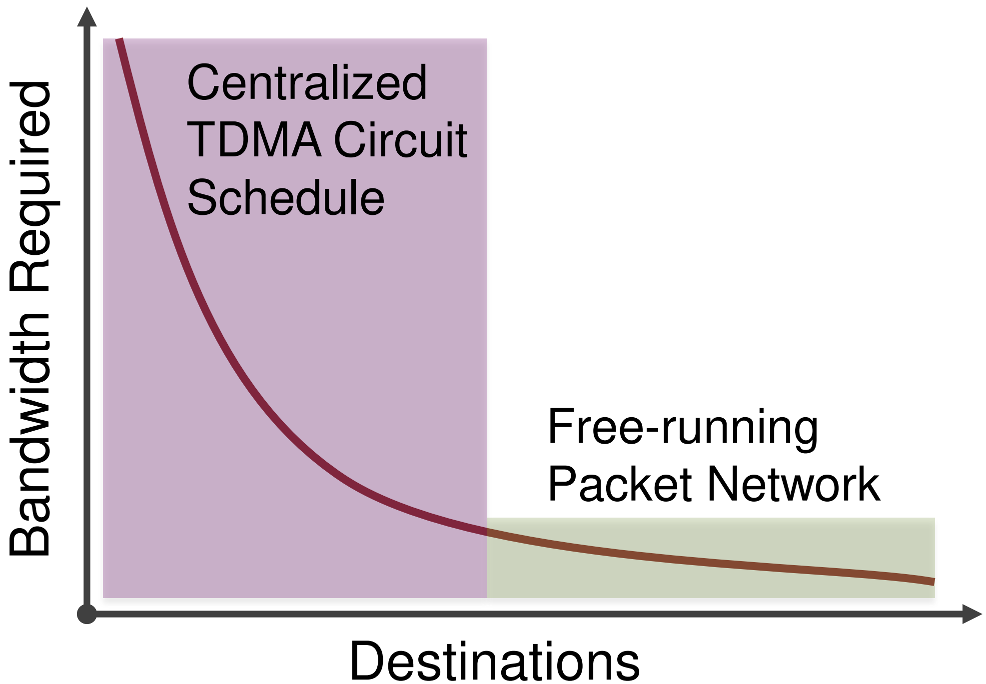
Optical Circuit

- High bandwidth
- Bufferless TDMA
- Tx only when circuit connects

Design Requirements

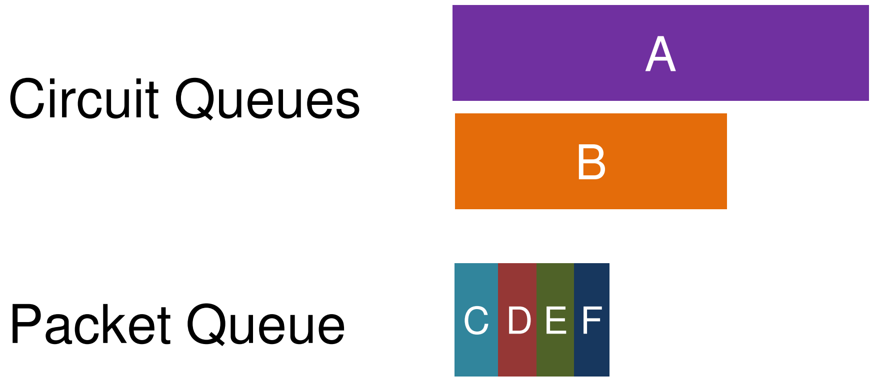
- Hybrid scheduling: classify traffic into circuits or packets
- Buffer packets at source hosts until circuit is available
- Have sources transmit when the circuit is connected
- Rate control to prevent downlink overload

The Hybrid Scheduling Problem



- Collect traffic demand from all hosts
- TDMA schedule the big flows on the circuit path
- Schedule the rest on the packet path
- An oracle predicts the demand and builds the schedules.

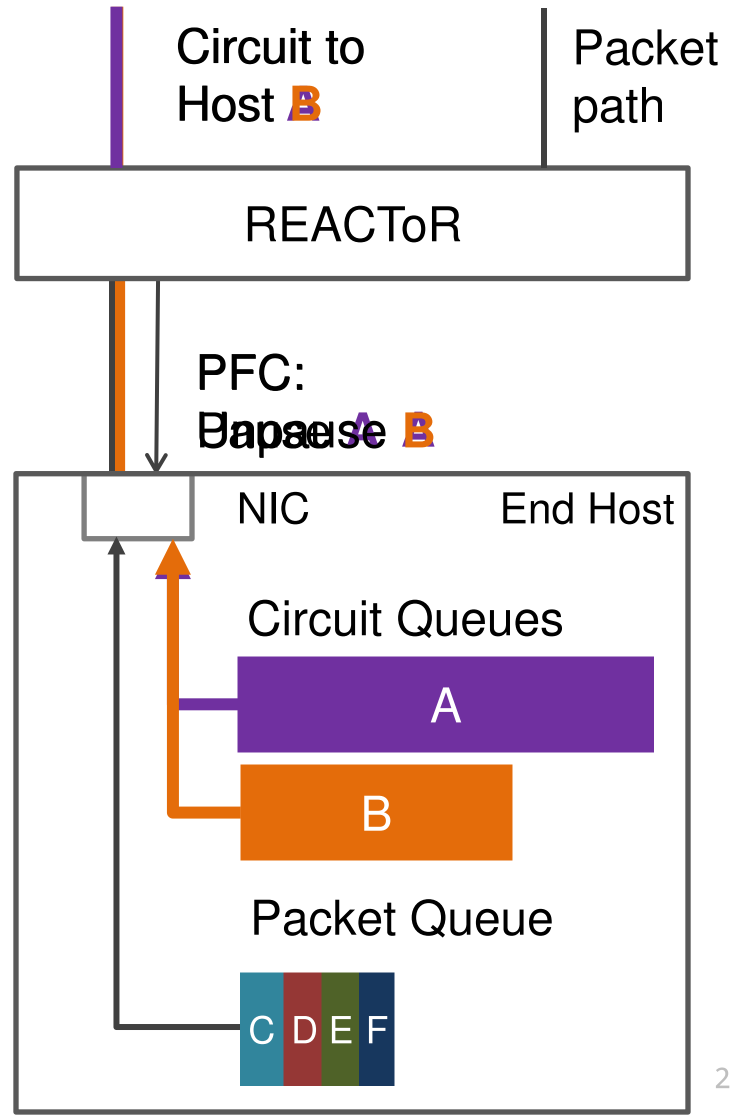
End Host: Classify and Buffer Packets



- Classify packets and map into different hardware queues
  - Based on the schedule
- Packet path: one hardware queue for all destinations
  - Can transmit at any time, but at 10G
- Circuit path: one hardware queue for each destination
  - Can only transmit when the particular circuit is connected
- Buffer the packets in end-host memory

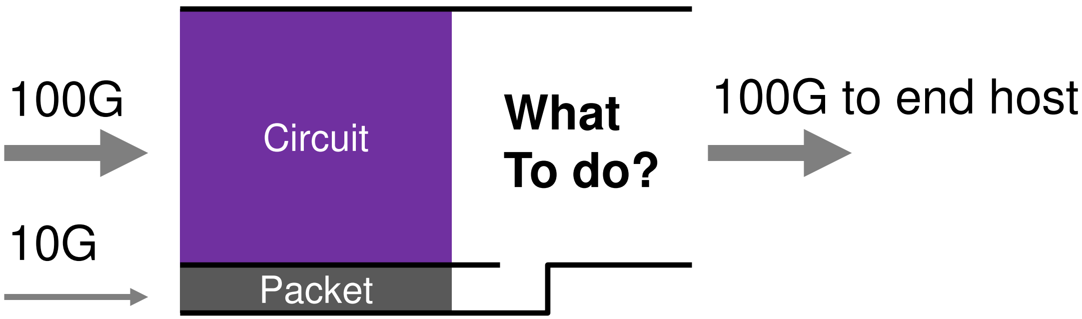
Packet Transmission

- Packet path: Rate limit to 10G
- Circuit path: Transmit only when the circuit is connected
- REACToR pulls packets from the circuit queue in real-time
- Use PFC frames to selectively unpause queues



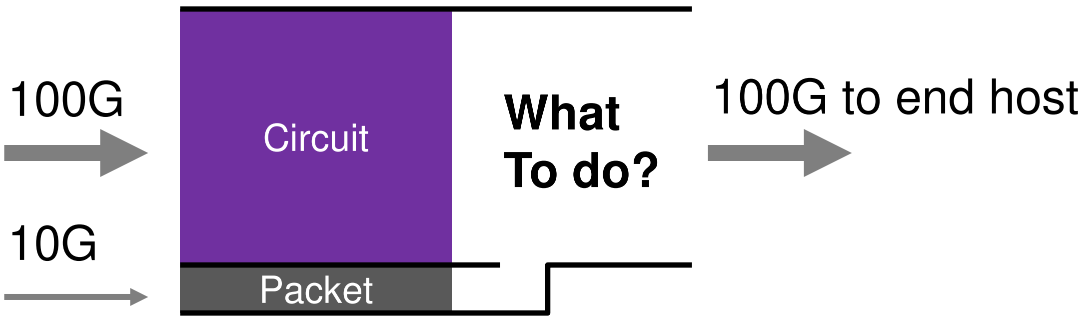
Rate Control

- Problem: downlink merging 100G + 10G to 100G

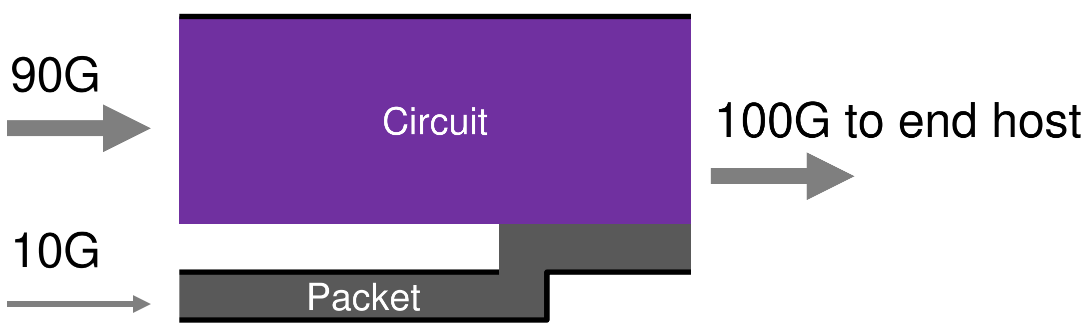


Rate Control

- Problem: downlink merging 100G + 10G to 100G

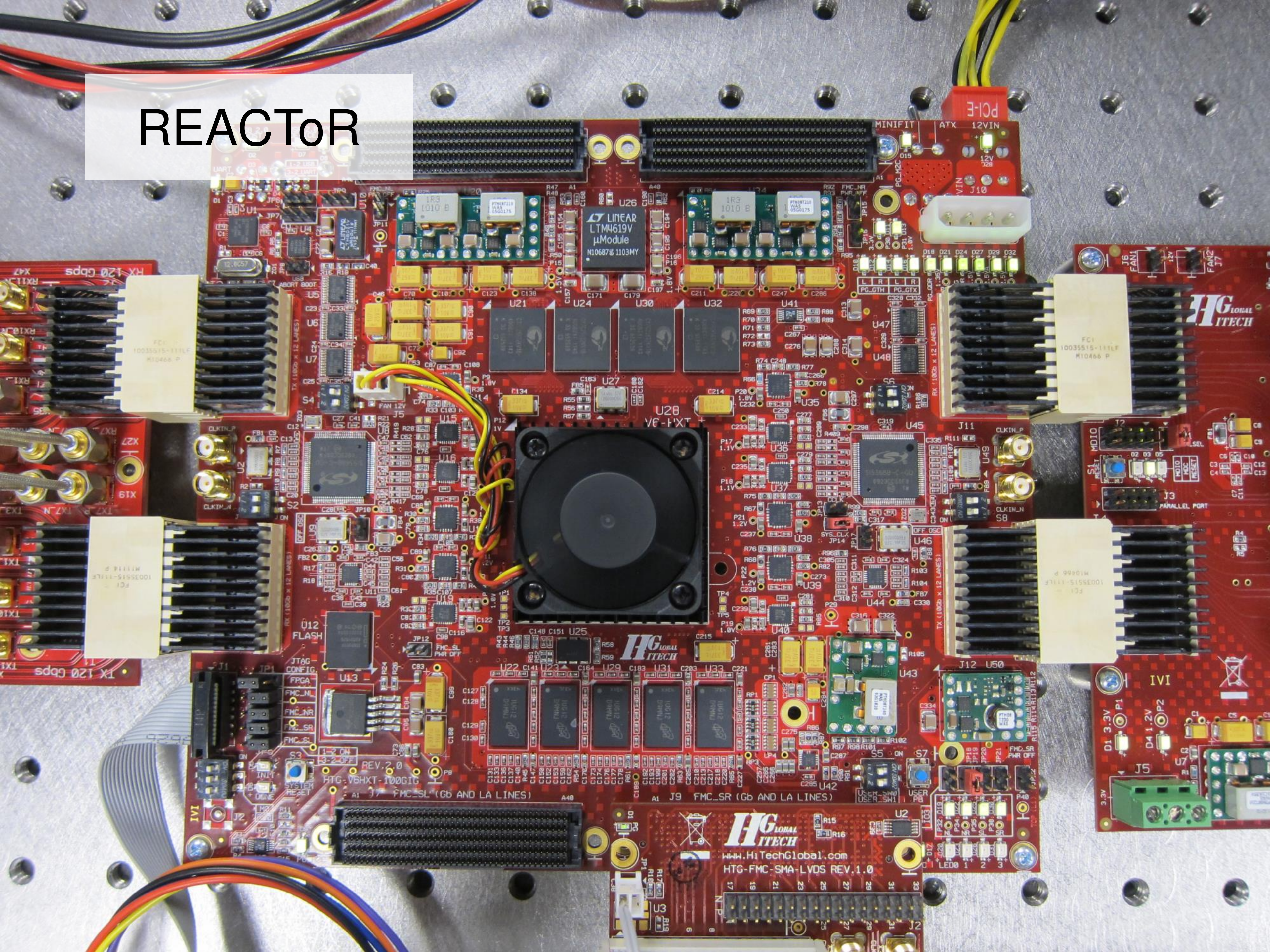


- Our approach: Rate limit the circuit path at the source to avoid overloading

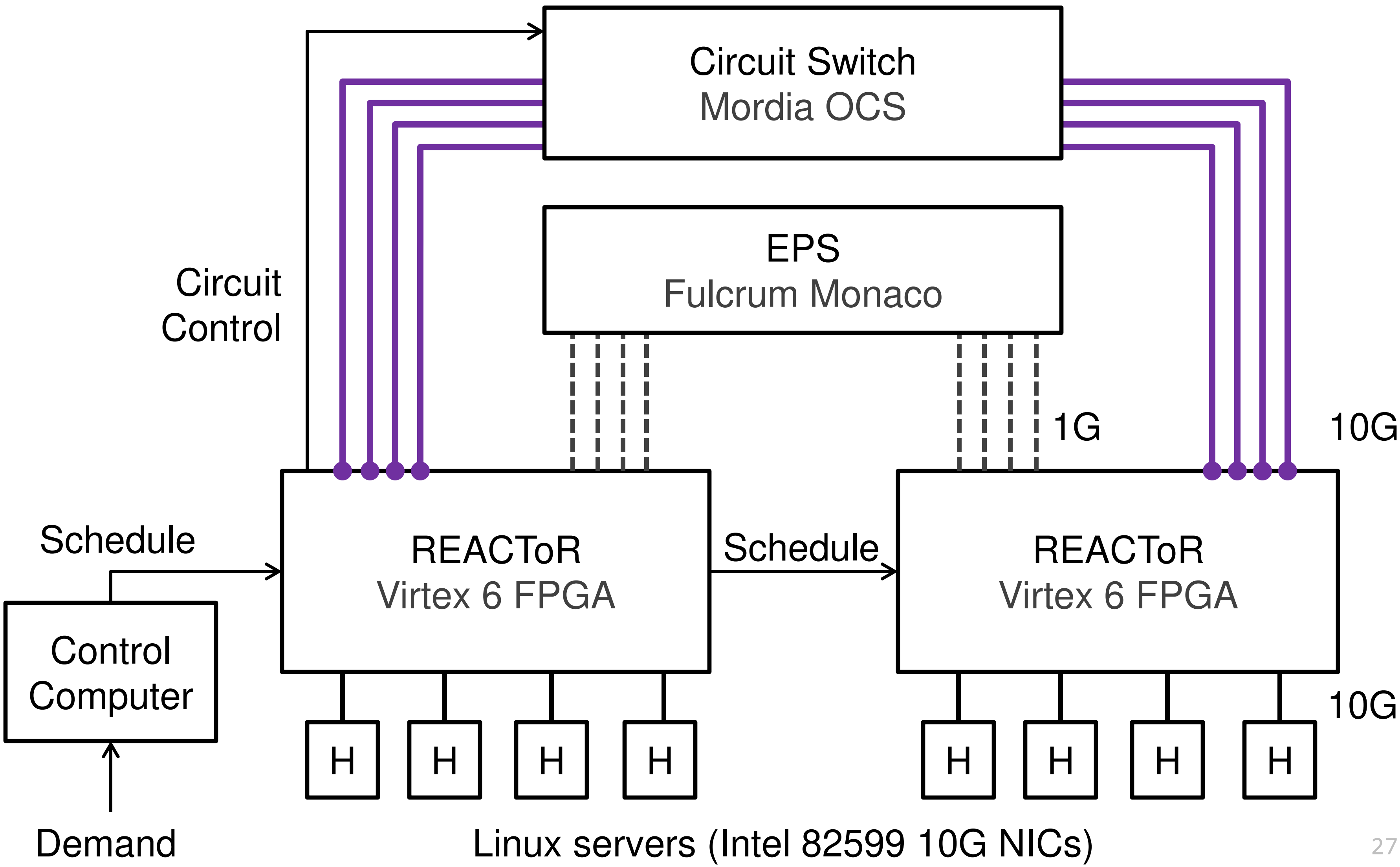


Implementation

REACToR

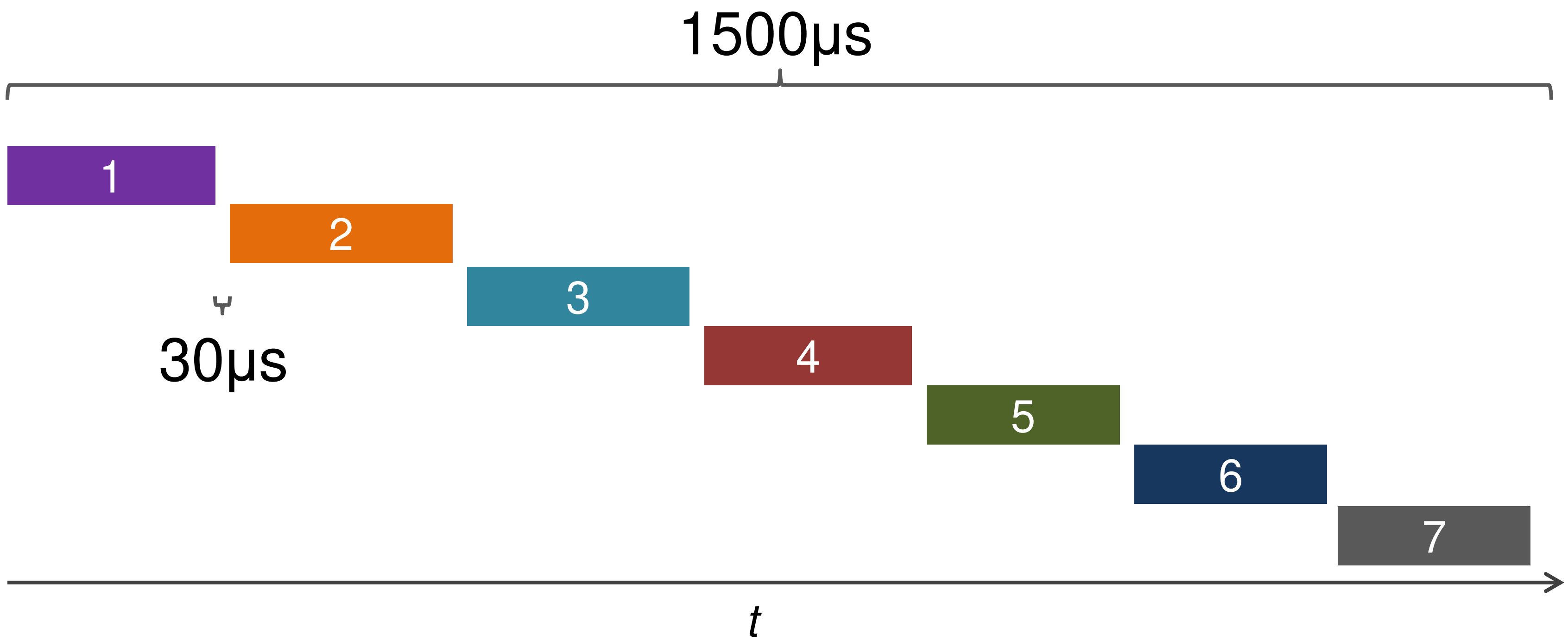

LINEAR
LTM4619V
µModule
HTG-V6HXT-100GIG
www.HiTechGlobal.com
HTG-FMC-SMA-LVDS REV.1.0
FMC_SL (Gb AND LA LINES)
FMC_SR (Gb AND LA LINES)
TX 120 Gbps
RX 120 Gbps

10G/1G Prototype



Timing Parameters

- End-to-end reconfiguration time: 30 μs
- Schedule reconfigures every 1500 us
- Example: 7 flows TDMA, 86% duty cycle



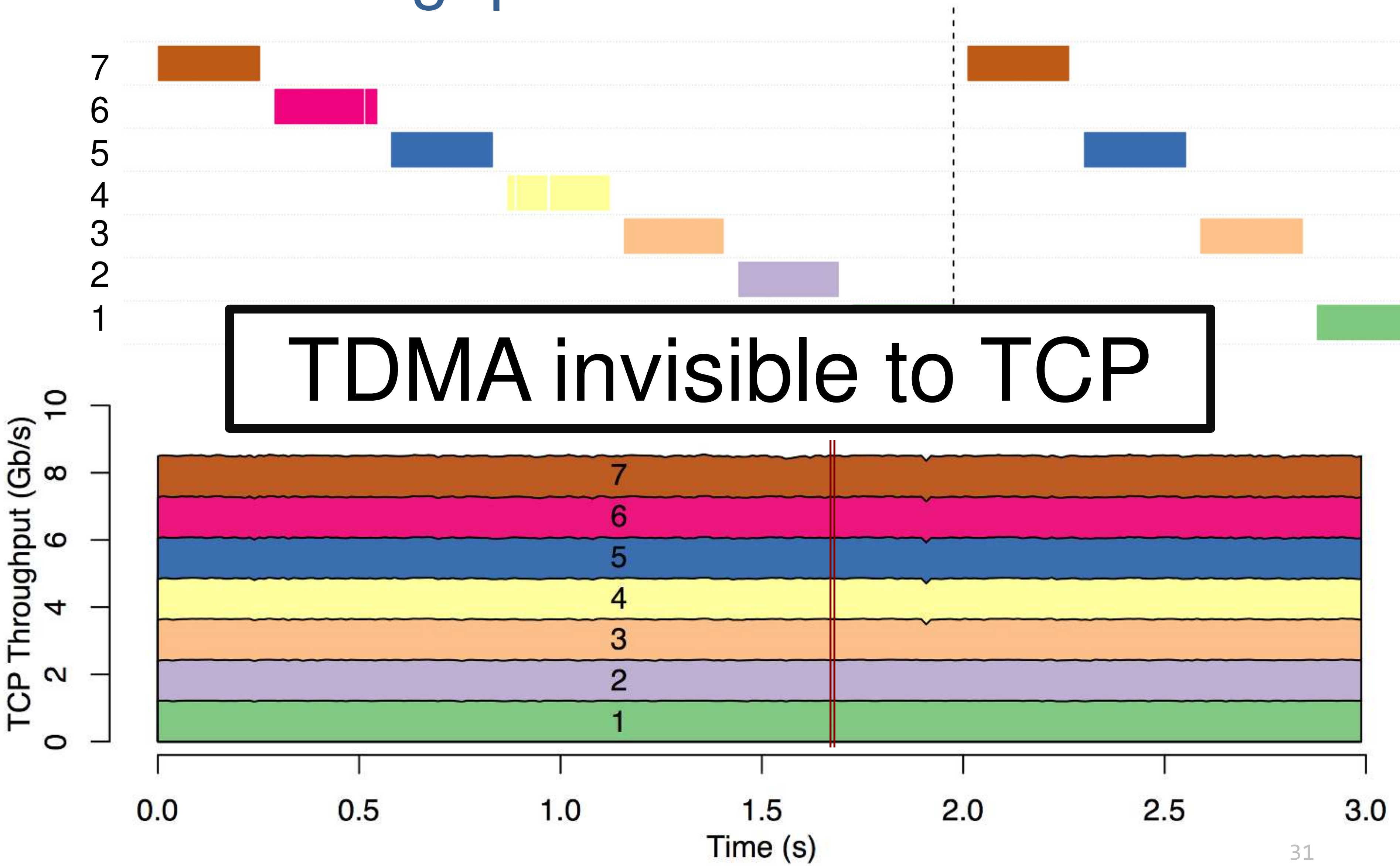
Evaluation

- Experiment 1: Supporting TCP
  - The performance on working with stock network stack
- Experiment 2: React to demand changes
  - The dynamics on handling changes and mispredictions
- Experiment 3: Demonstrate the benefit of using hybrid
  - The performance gain on handling skewed demand

Experiment 1: Supporting TCP

- Each host receives 7 TCP flows from all other hosts
- Hybrid schedule: data packets via OCS, ACKs via EPS
- 7 flows TDMA, fair sharing the link
- Check if TCP works with high throughput

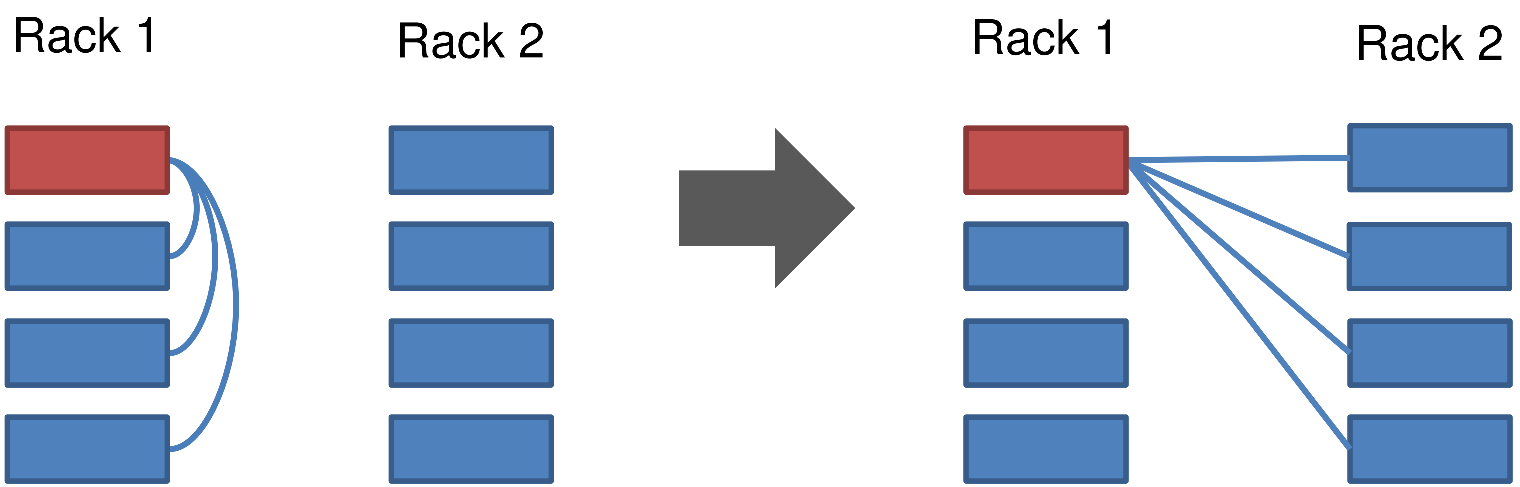
TCP Throughput



Experiment 2: React to Demand Changes

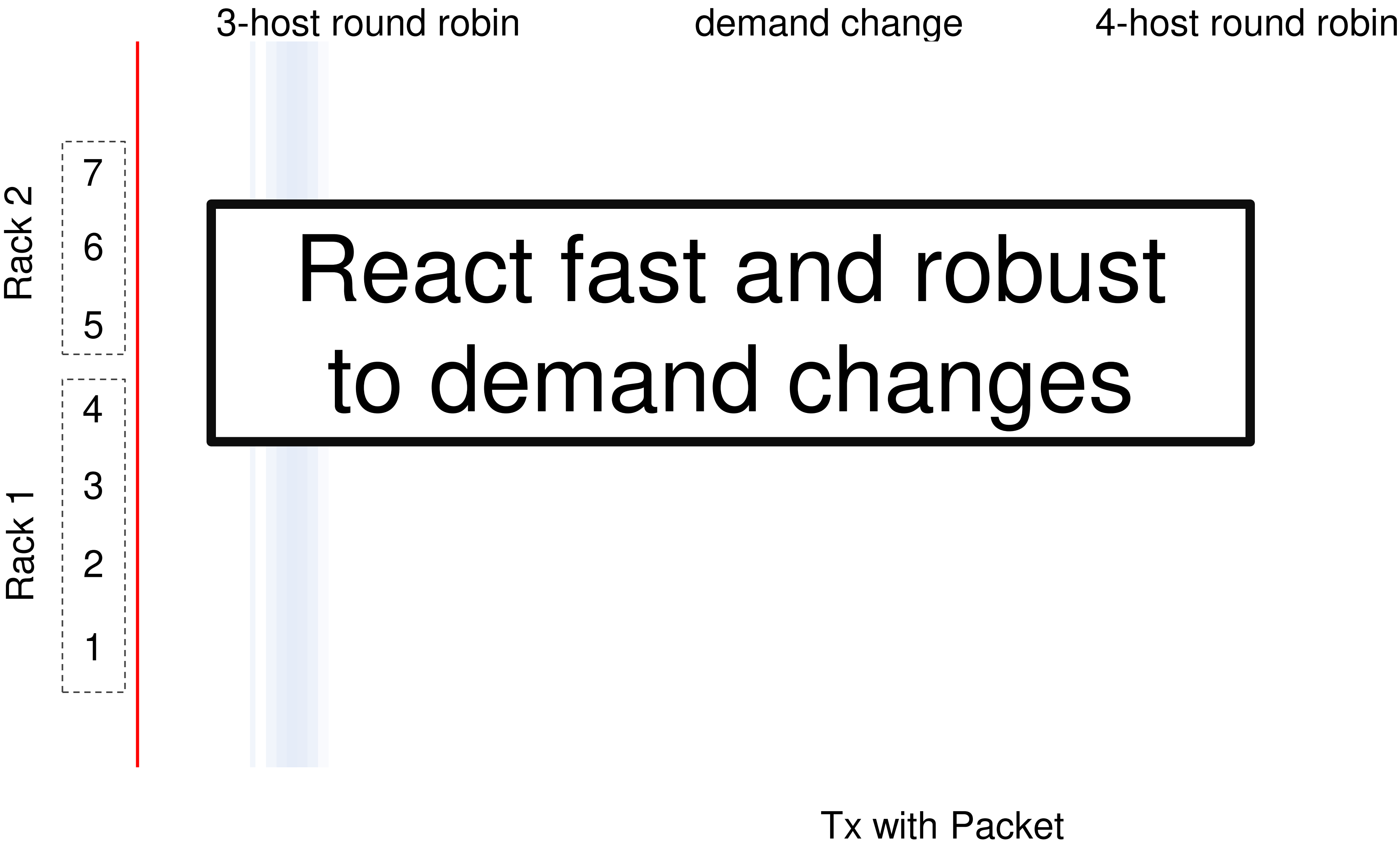
From: Intra-rack Traffic

To: Inter-rack Traffic



Use pktgen to impose precise and sudden traffic pattern change.
See if REACToR can *react* in time.

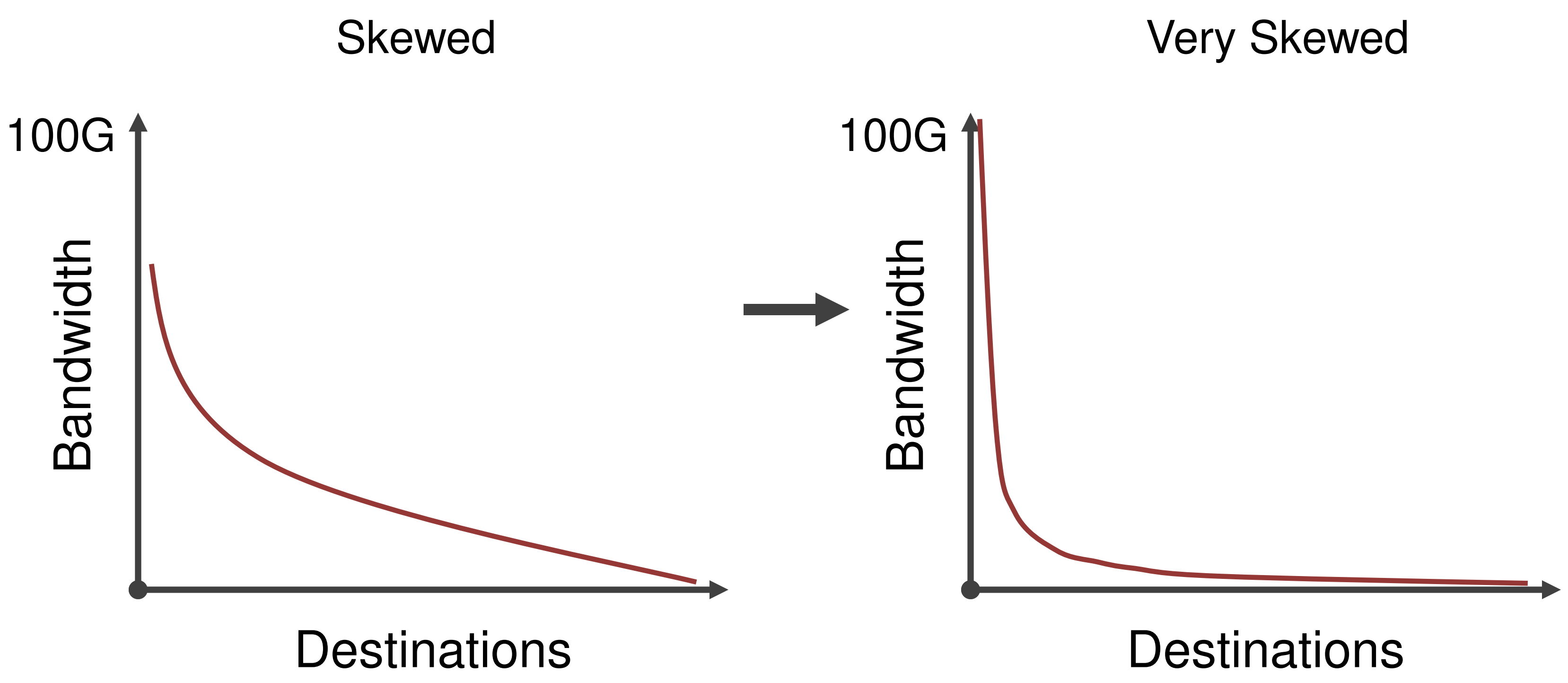
React to Demand Changes



React fast and robust to demand changes

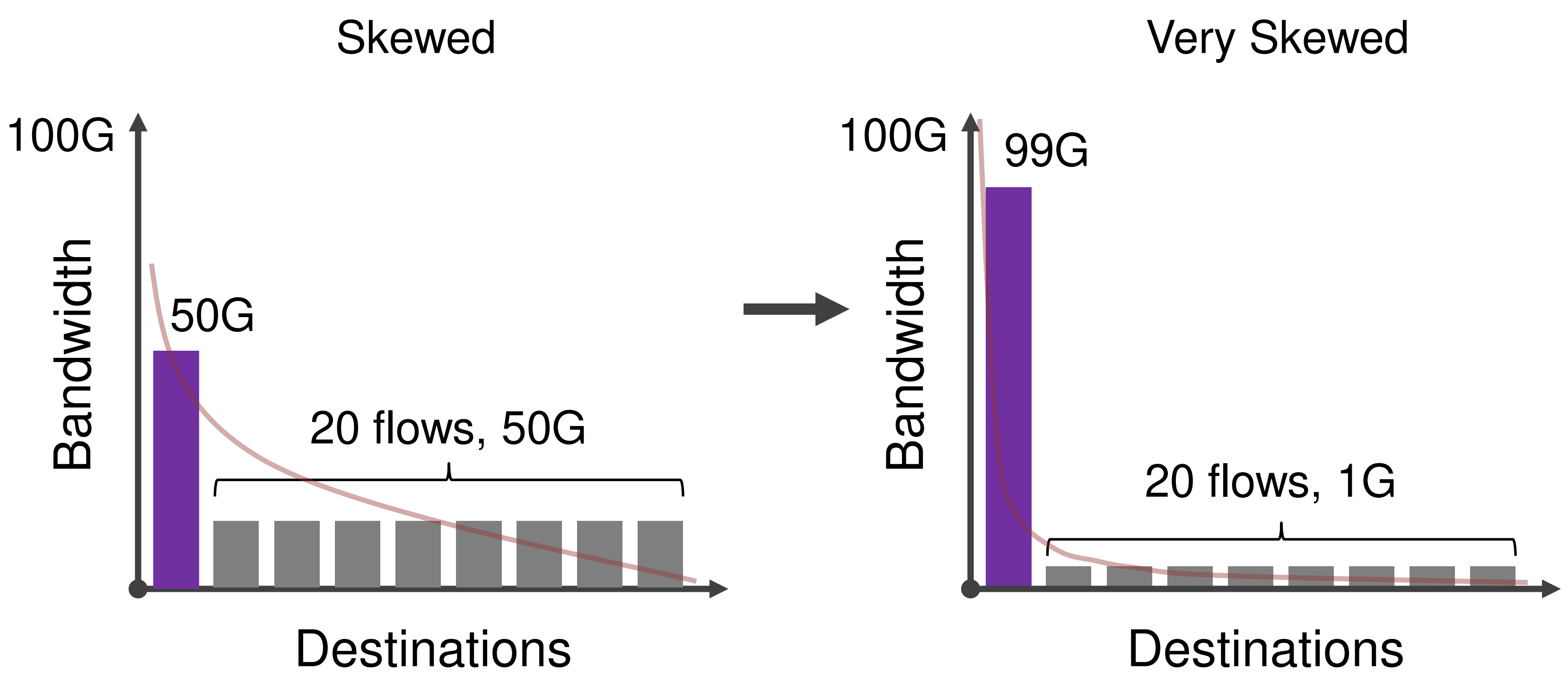
Experiment 3: Demonstrating Hybrid

- Simulated 64 hosts with demand of different skewness
- Big benefit from a small electrical packet switch

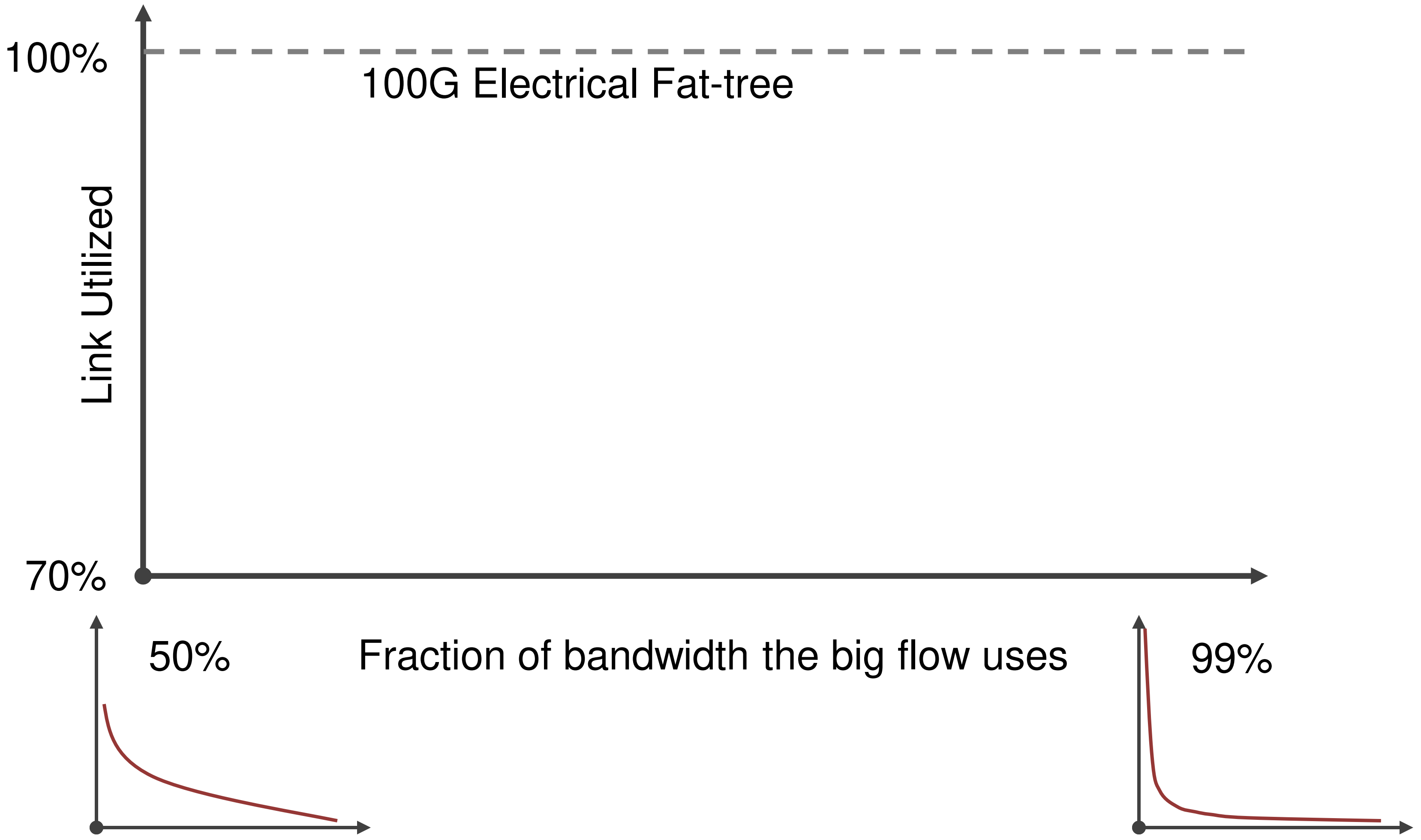


Experiment 3: Demonstrating Hybrid

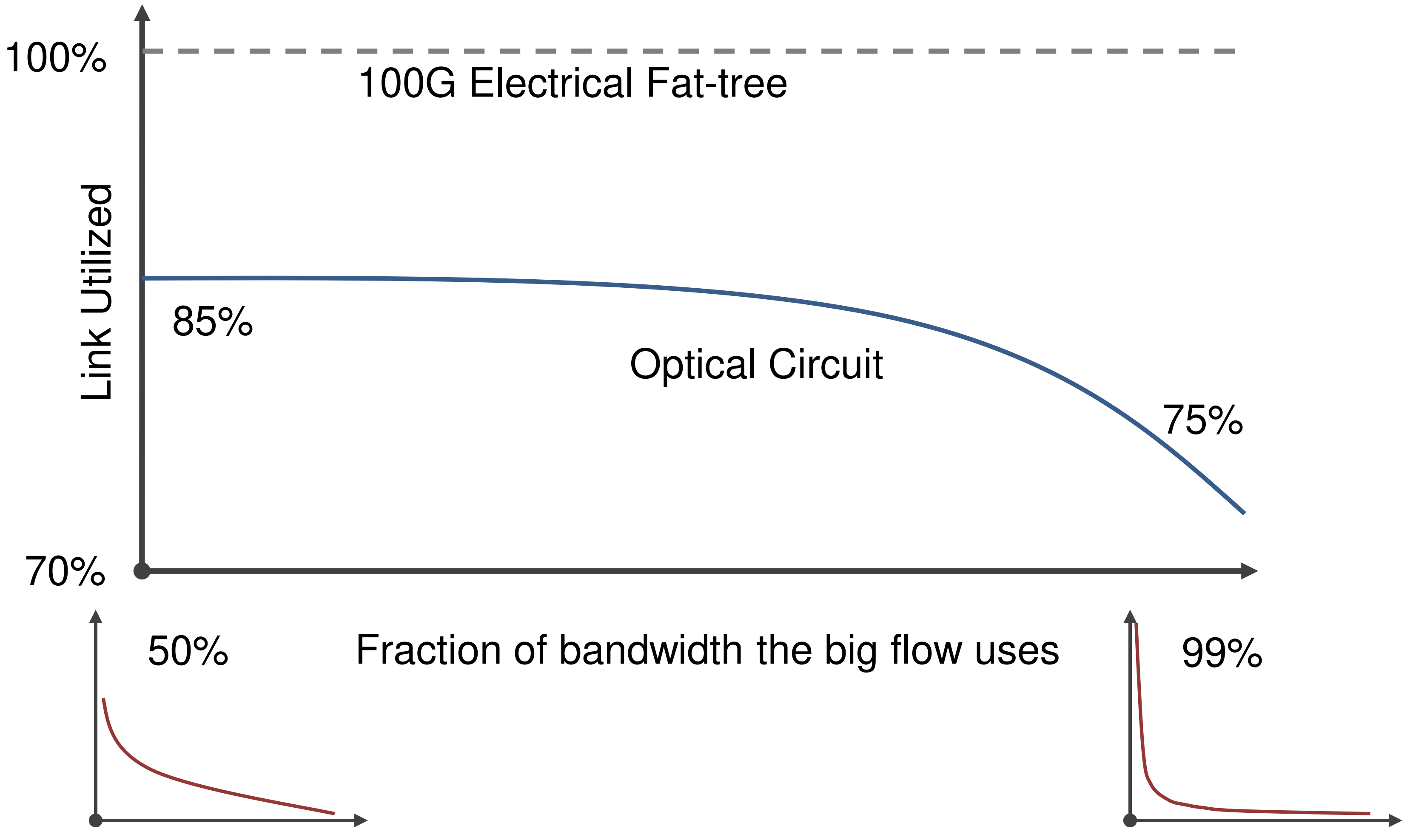
- Simulated 64 hosts with demand of different skewness
- Big benefit from a small electrical packet switch



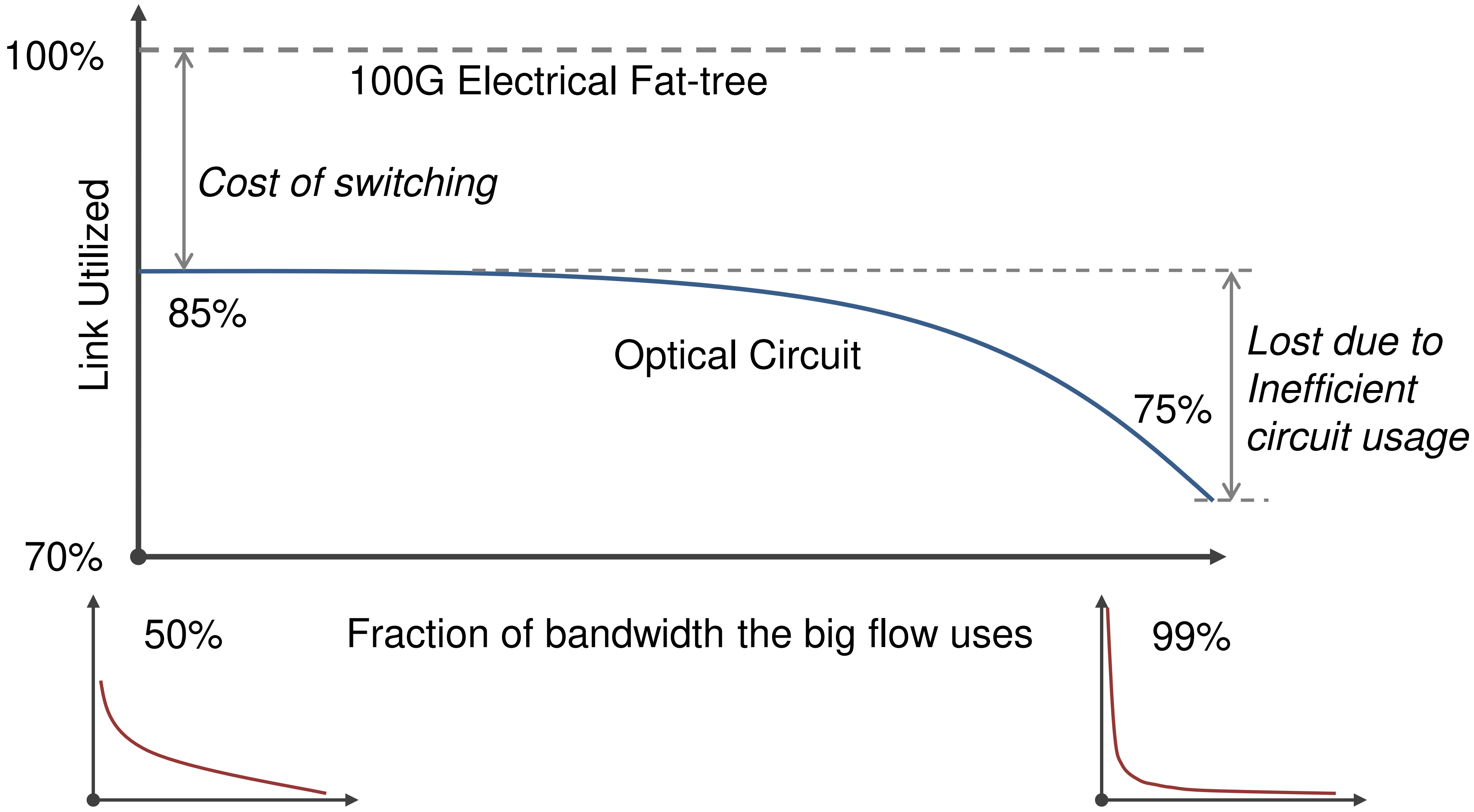
Optical Circuit Switching Not Enough



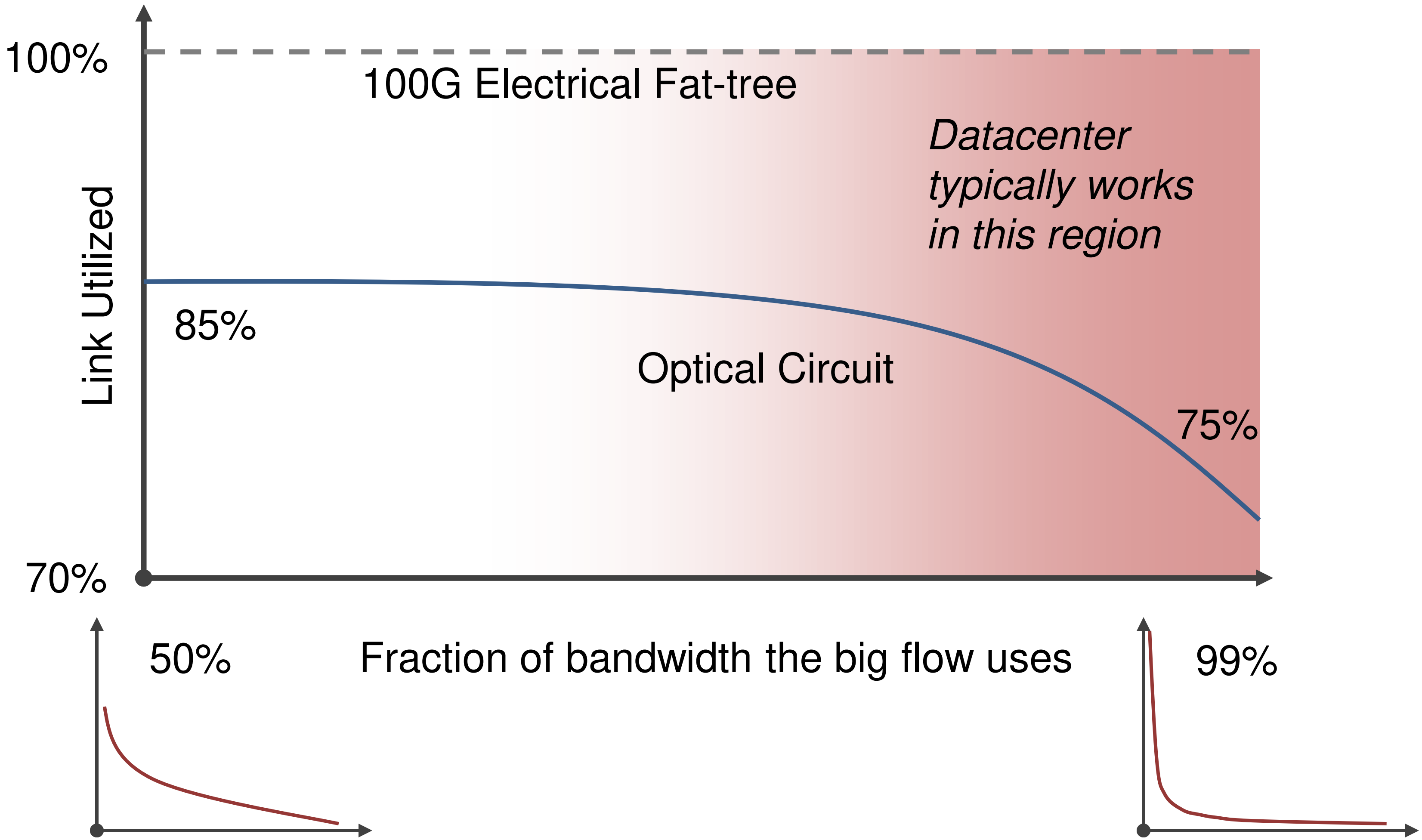
Optical Circuit Switching Not Enough



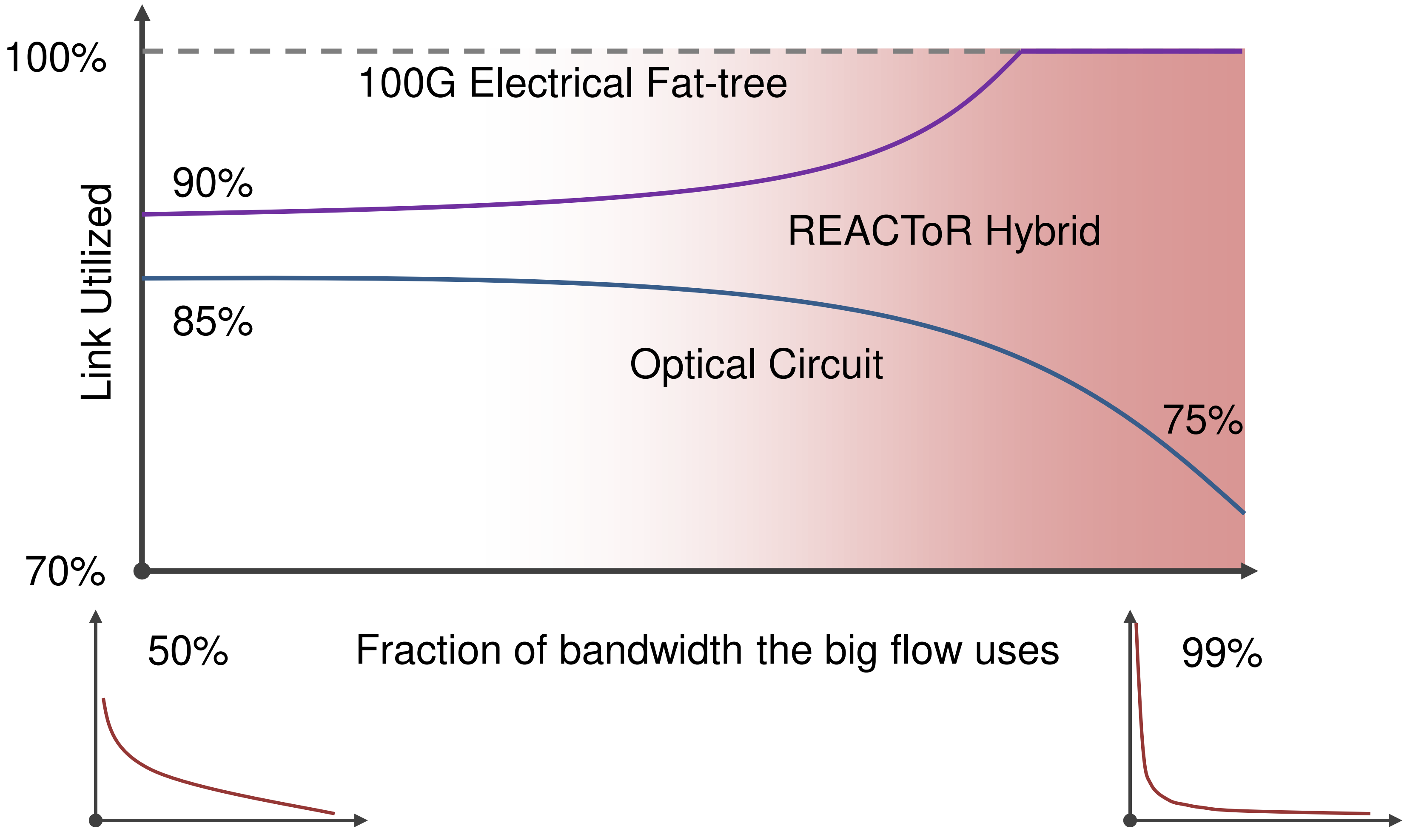
Optical Circuit Switching Not Enough



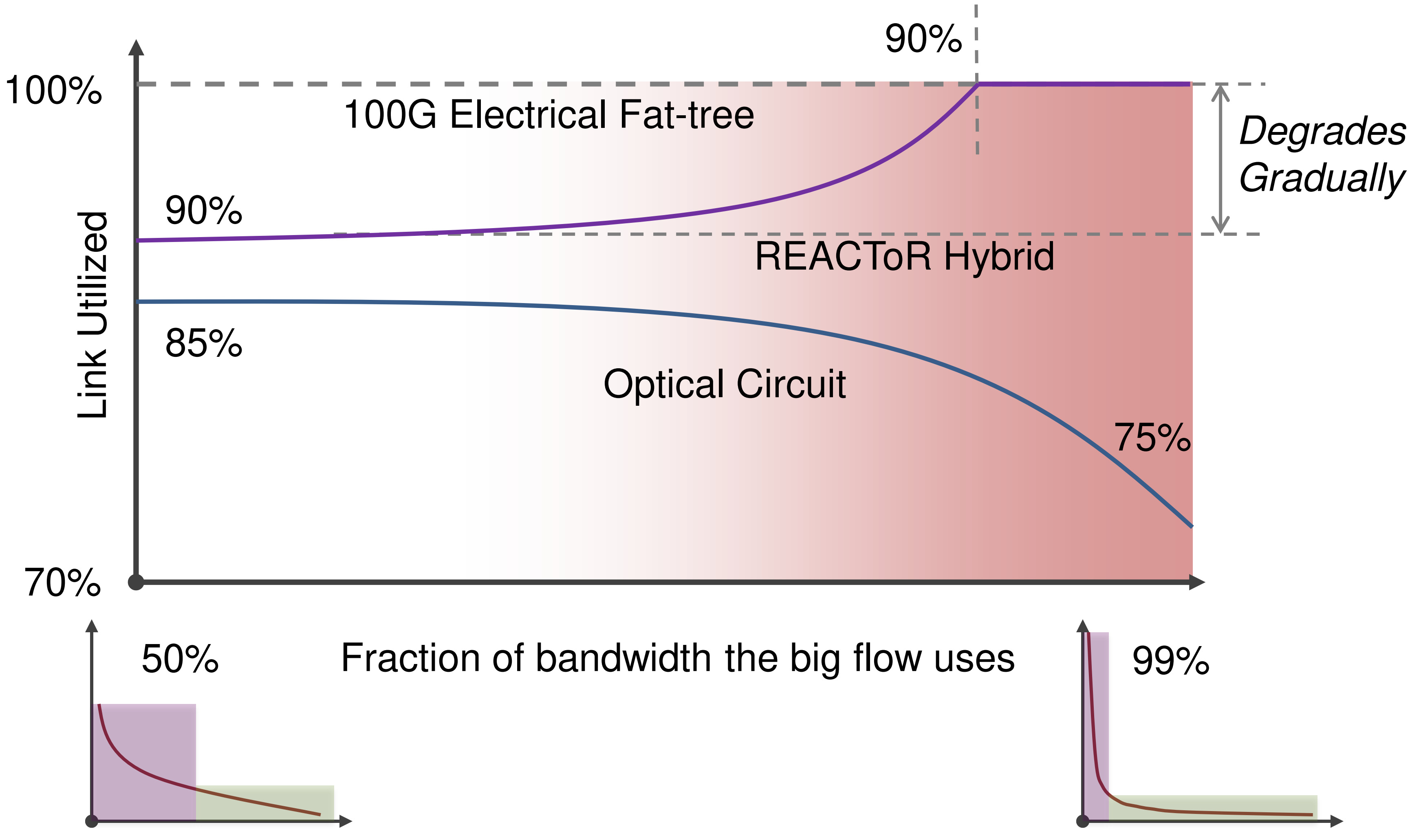
Optical Circuit Switching Not Enough



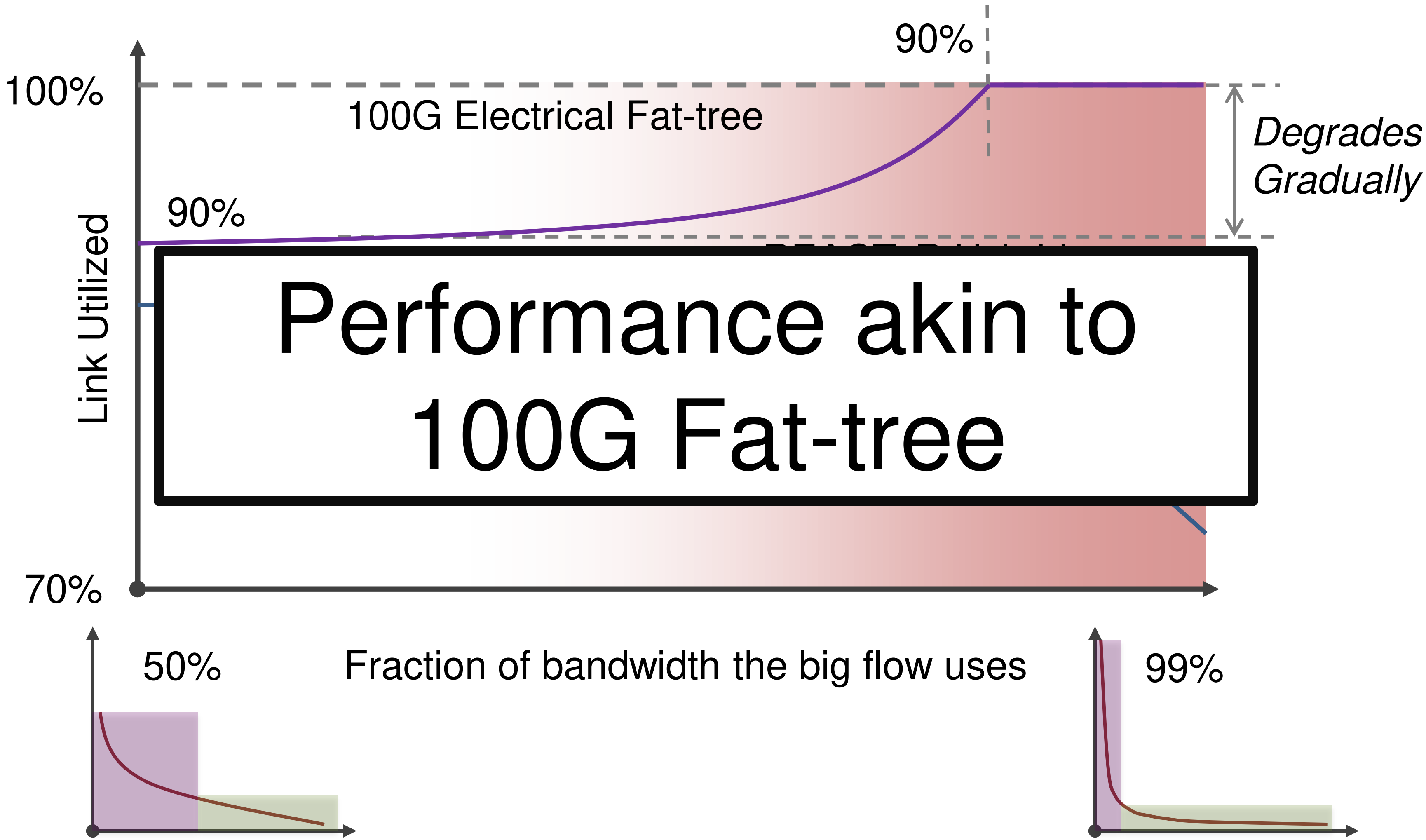
Hybrid Switching with REACToR



Hybrid Switching with REACToR

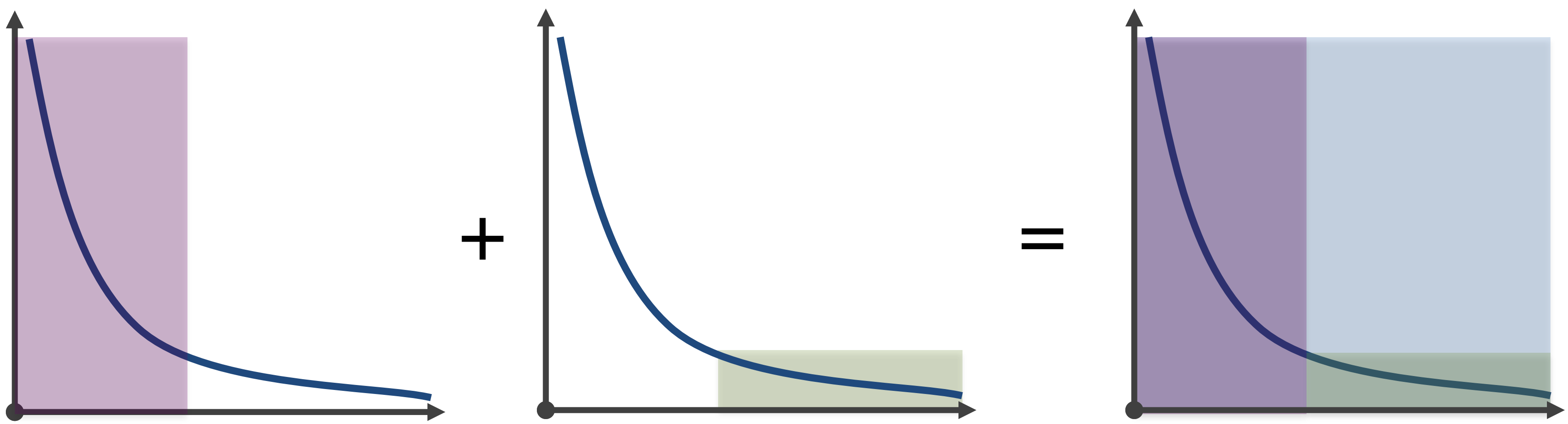


Hybrid Switching with REACToR



Conclusion

REACToR



100G Optical Circuit Switching

10G Electrical Packet Switching

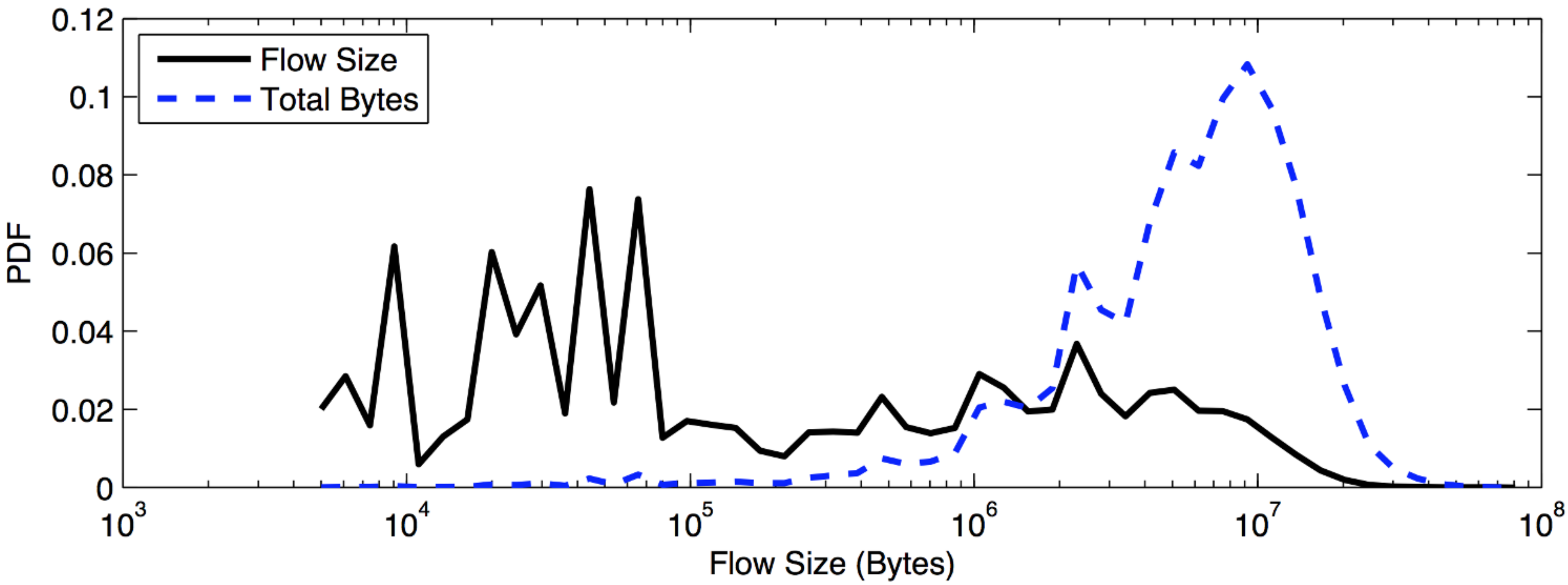
100G Electrical Packet Switching

For datacenter workloads
At a lower cost

Thank you!

DCTCP: Datacenter Workload

[SIGCOMM 2010]



Cost of Transceivers

- Cost of 10G Transceivers
  - Cost: $500 per pair
  - Power: 1Watt per pair
  - (100G costs even more)
- 3-Level Fat-tree: 27.6k hosts
- Transceivers per host:

Link rate	Full fat tree	Helios-like	REACToR
10 Gb/s	2 – 4	1 – 3	N/A
100 Gb/s	4	3	1†

Scheduling

- Problem: matrix decomposition
  - Similar to BvN, but must consider reconfiguration penalty
  - NP-complete problem
  - Goal: schedule all the big flows (90% of the demand)
- Greedy approach: e.g. iSLIP
  - Suboptimal
- Naïve BvN:
  - Fragmented by small elements and residuals
- A good algorithm should:
  - Prioritize the big flows
  - Perform full matrix decomposition (like BvN)
  - Minimize number of reconfigurations at the same time