Attaining the Promise and Avoiding the Pitfalls of TCP in the Datacenter

Glenn Judd

Morgan Stanley

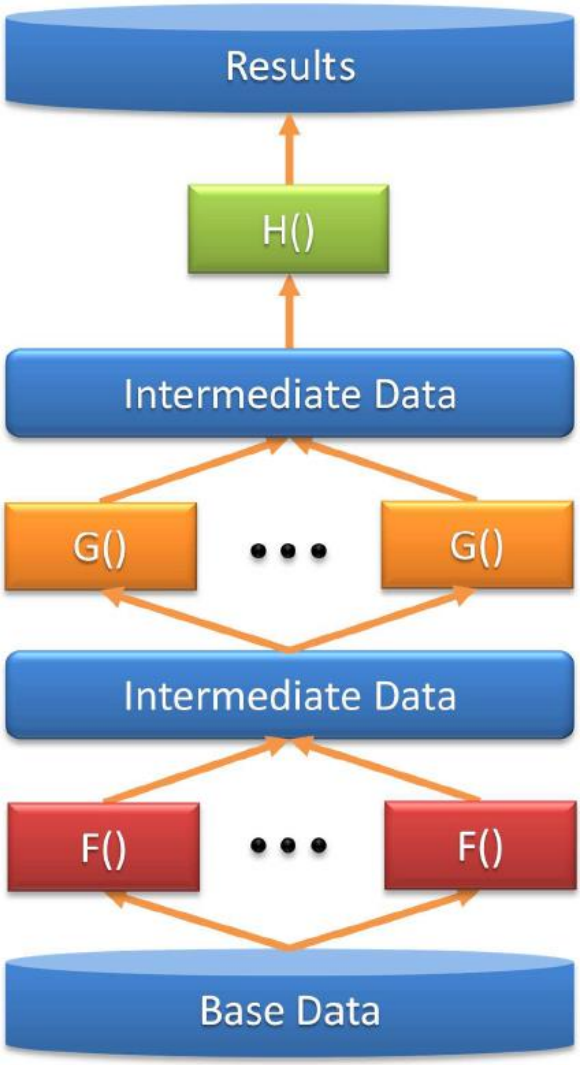
Introduction

- Datacenter computing pervasive
  - Beyond the Internet services domain
  - “BigData”, “Grid Computing”, “Cloud” etc.
- Communication is essential
  - Demand is increasing
- TCP
  - Leverage existing base of TCP applications
  - Avoid re-inventing flow control, congestion control, etc.
  - **Not designed with datacenter computing in mind**
    - Internet: propagation delay >>> queueing delay
    - Datacenter: propagation delay <<< queueing delay

Outline

- Background
- TCP Datacenter Challenges
  - Latency & Application Coupling
  - Incast
  - DCTCP
    - Deployment Challenges
    - Performance
  - Receive Buffer Tuning
- Related Work
- Conclusion

Application Types

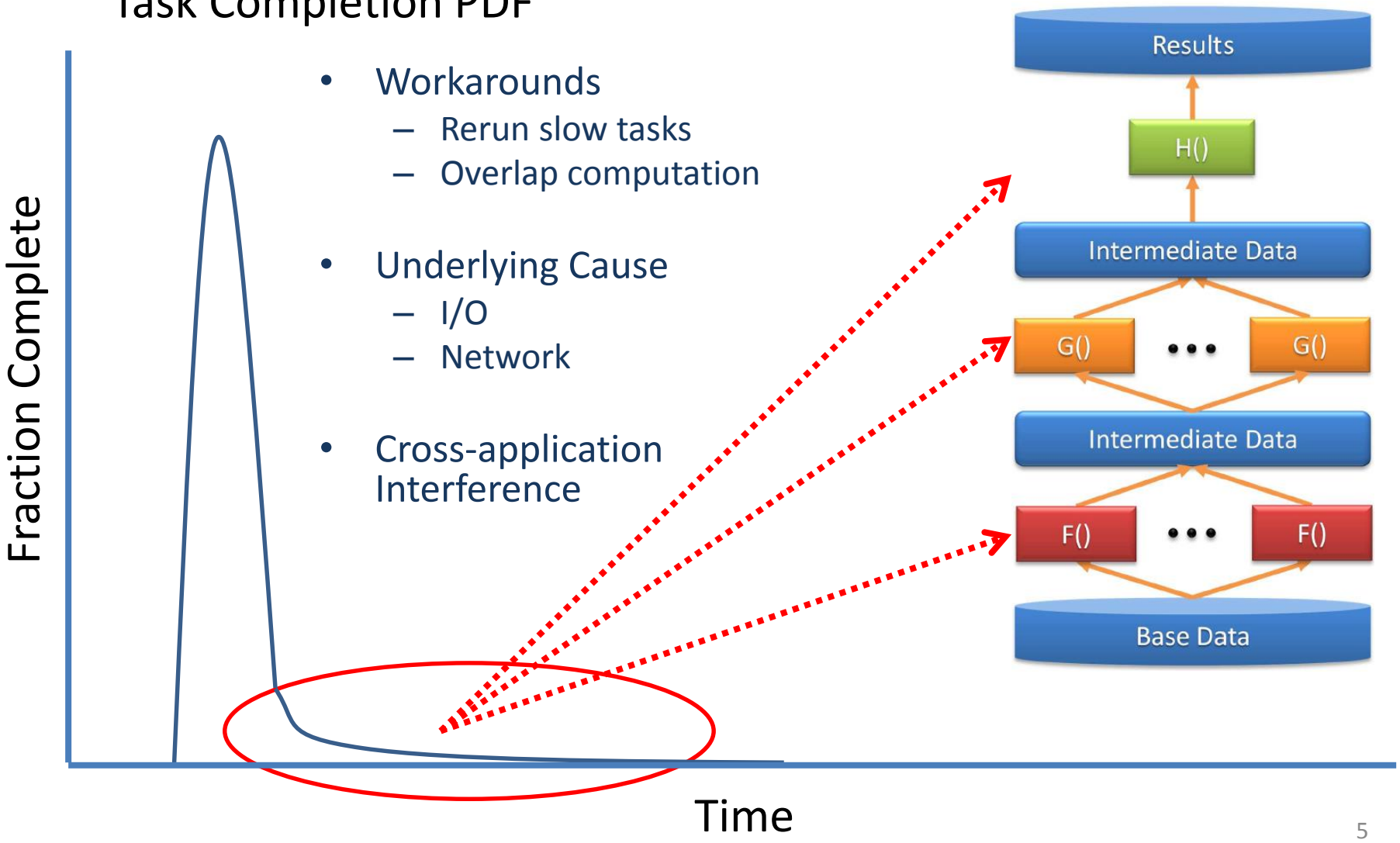


- Large-scale computation
  - Monte Carlo simulation
  - Data analysis
- Latency-sensitive computation
  - End-user-facing
- Multiple groups sharing the datacenter

Motivation

Task Completion PDF

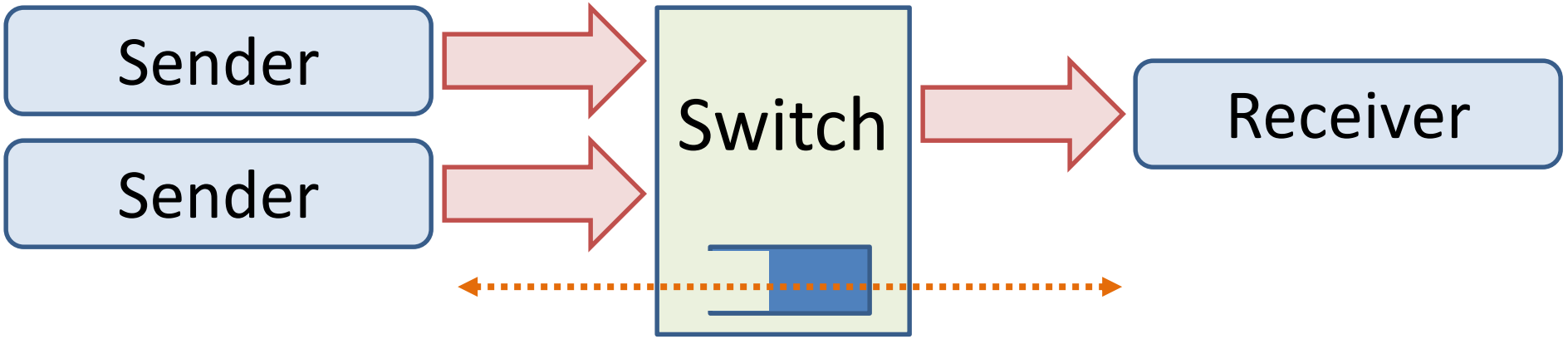
- Workarounds
  - Rerun slow tasks
  - Overlap computation
- Underlying Cause
  - I/O
  - Network
- Cross-application Interference



Outline

- Background
- TCP Datacenter Challenges
  - Latency & Application Coupling
  - Incast
  - DCTCP
    - Deployment Challenges
    - Performance
  - Receive Buffer Tuning
- Related Work
- Conclusion

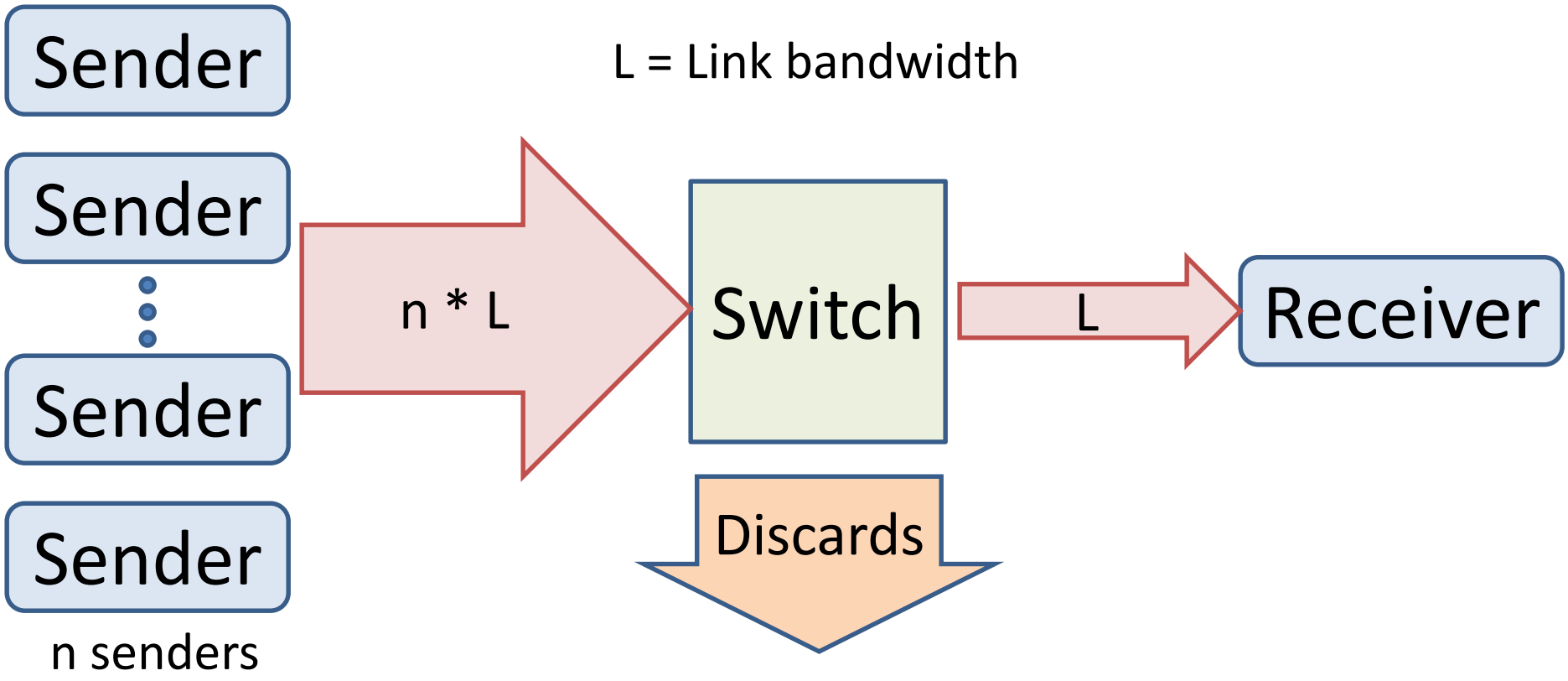
Latency & Application Coupling



All traffic transiting queue incurs latency

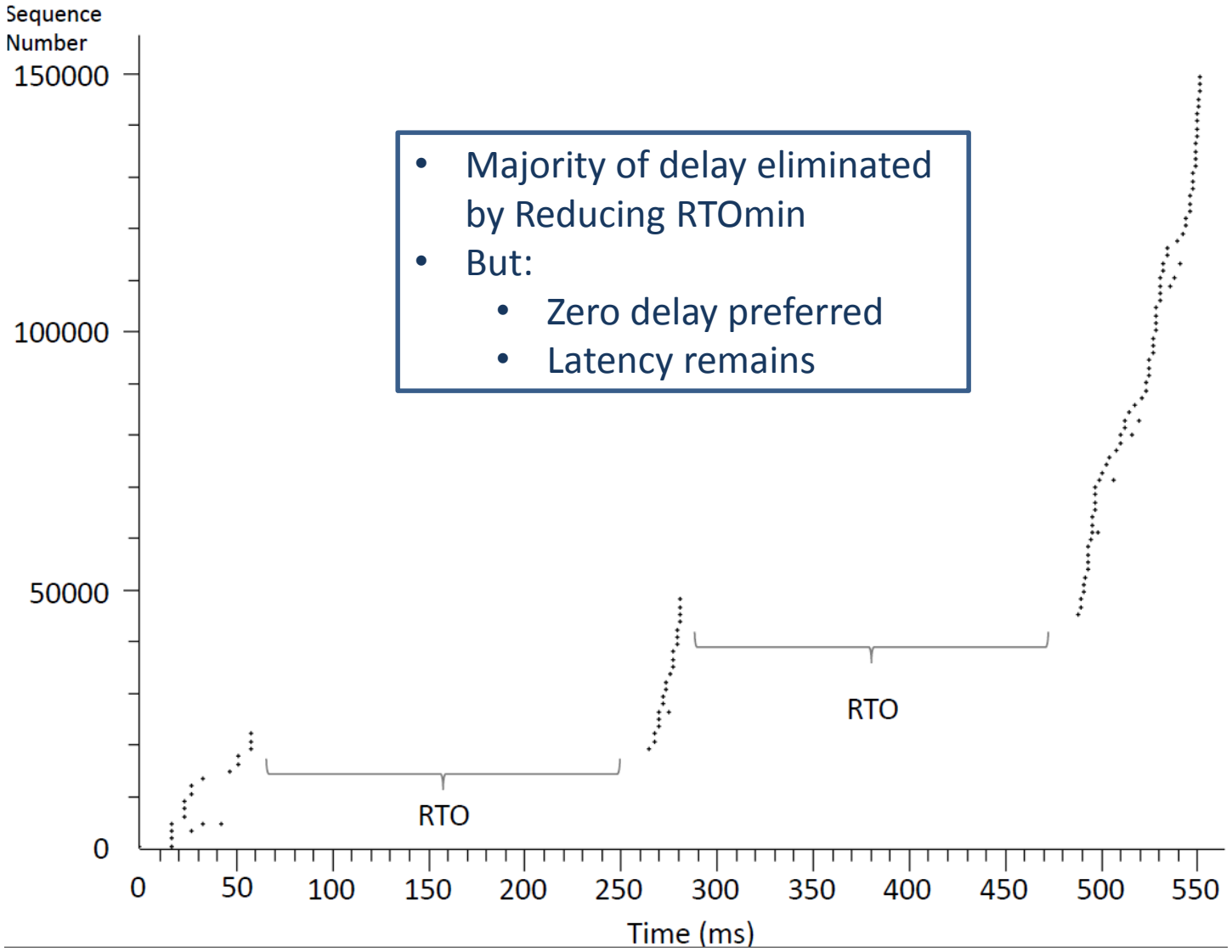
- By design, TCP fills queues
  - Even simple congestion
  - Latency
    - microseconds → milliseconds
- Application coupling
  - Shared congested link
  - Buffer pressure
  - Disparate groups coupled through network

Incast



- Triggers timeouts
- Exacerbated by synchronized requests
  - Common in distributed storage

Impact of Timeouts



- Majority of delay eliminated by Reducing RTOmin
- But:
  - Zero delay preferred
  - Latency remains

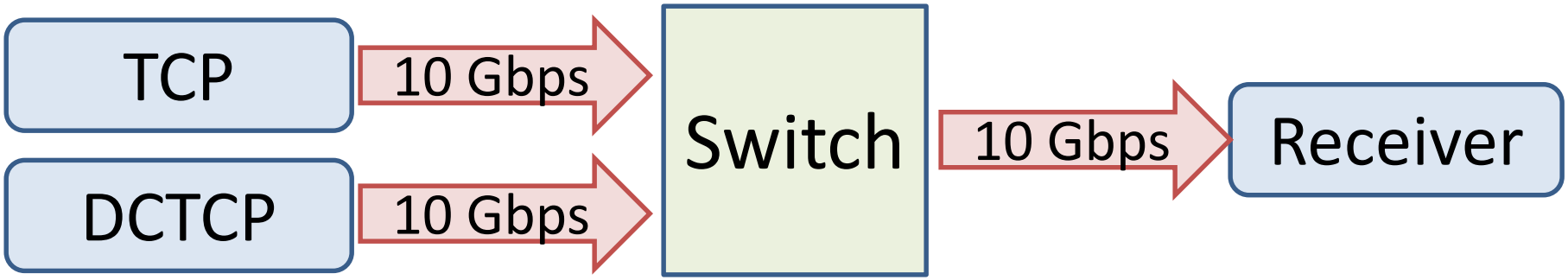
Outline

- Background
- TCP Datacenter Challenges
  - Latency & Application Coupling
  - Incast
  - DCTCP
    - Deployment Challenges
    - Performance
  - Receive Buffer Tuning
- Related Work
- Conclusion

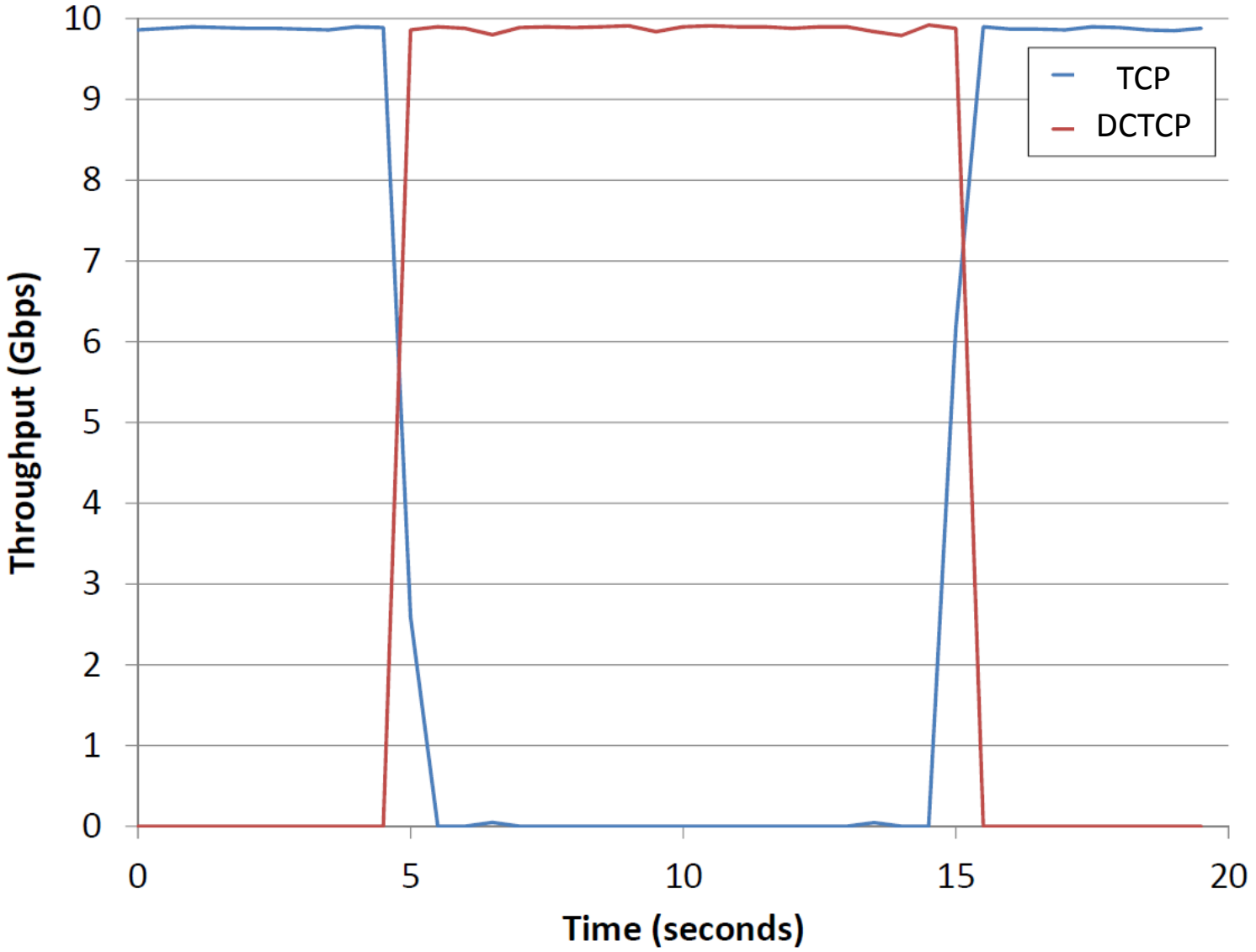
DCTCP

- Key idea: reduce queue occupancy
  - Reduced loss
  - Reduce timeouts
  - Reduced queueing delay
- Mechanism
  - Leverage ECN RED/AQM capabilities
- Available in current technology
- Can it work in production? At scale?

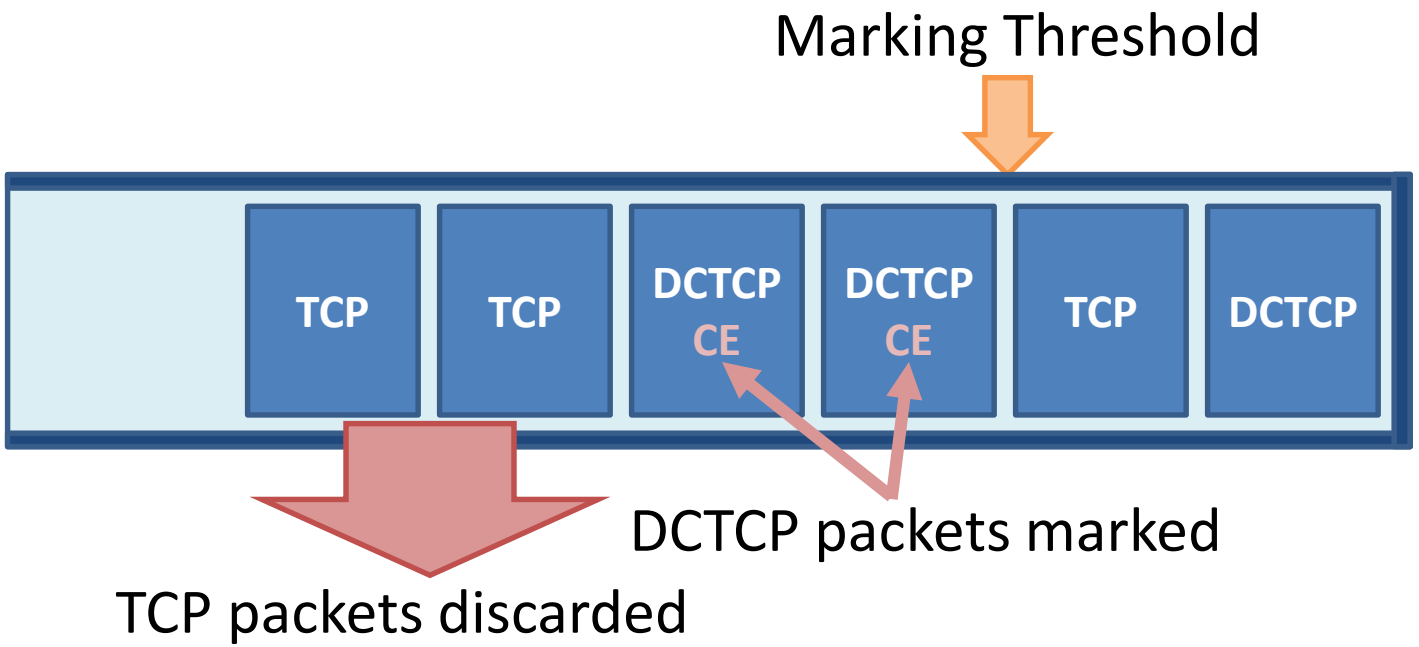
Coexistence with TCP



Coexistence with TCP

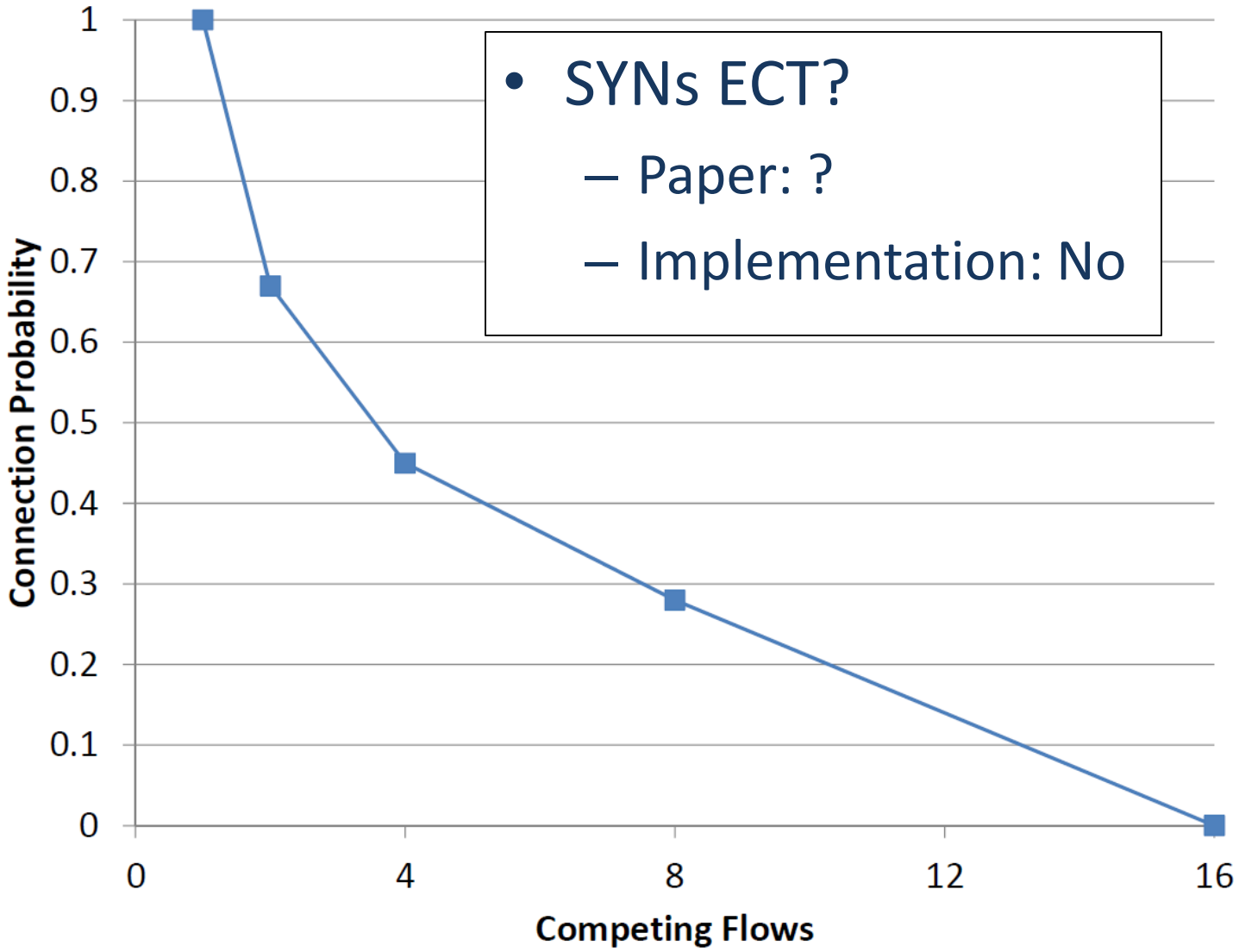


RED-ECN AQM Implementation



- Solution: Segregate DCTCP traffic using QoS

Connection Establishment



- SYNs ECT?
  - Paper: ?
  - Implementation: No

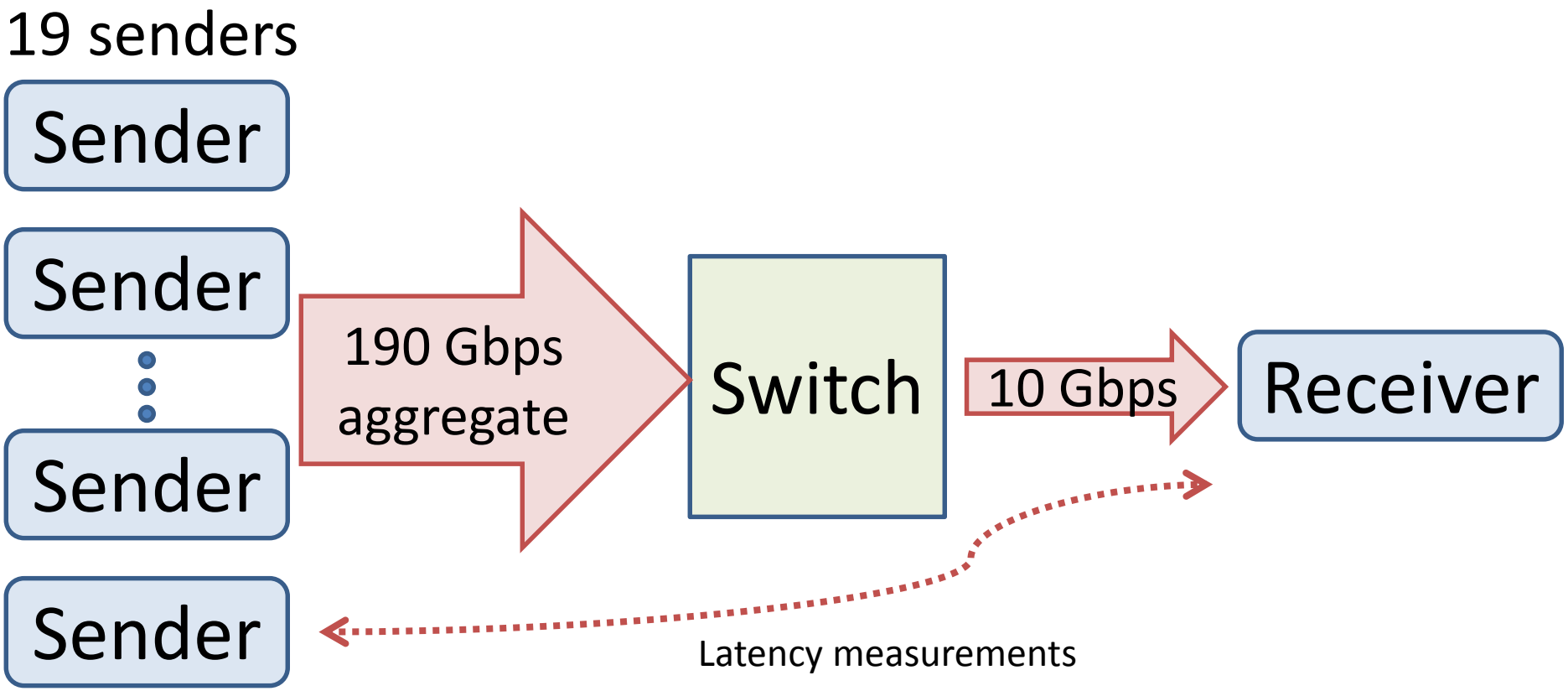
Non-technical Challenges

- Network administrators are necessarily risk averse
  - You want to do what?!
- Keys to deploying DCTCP
  - Demonstrated reduction in coupling
  - Primum non nocere
    - Support for TCP and DCTCP coeixistence
  - Timing

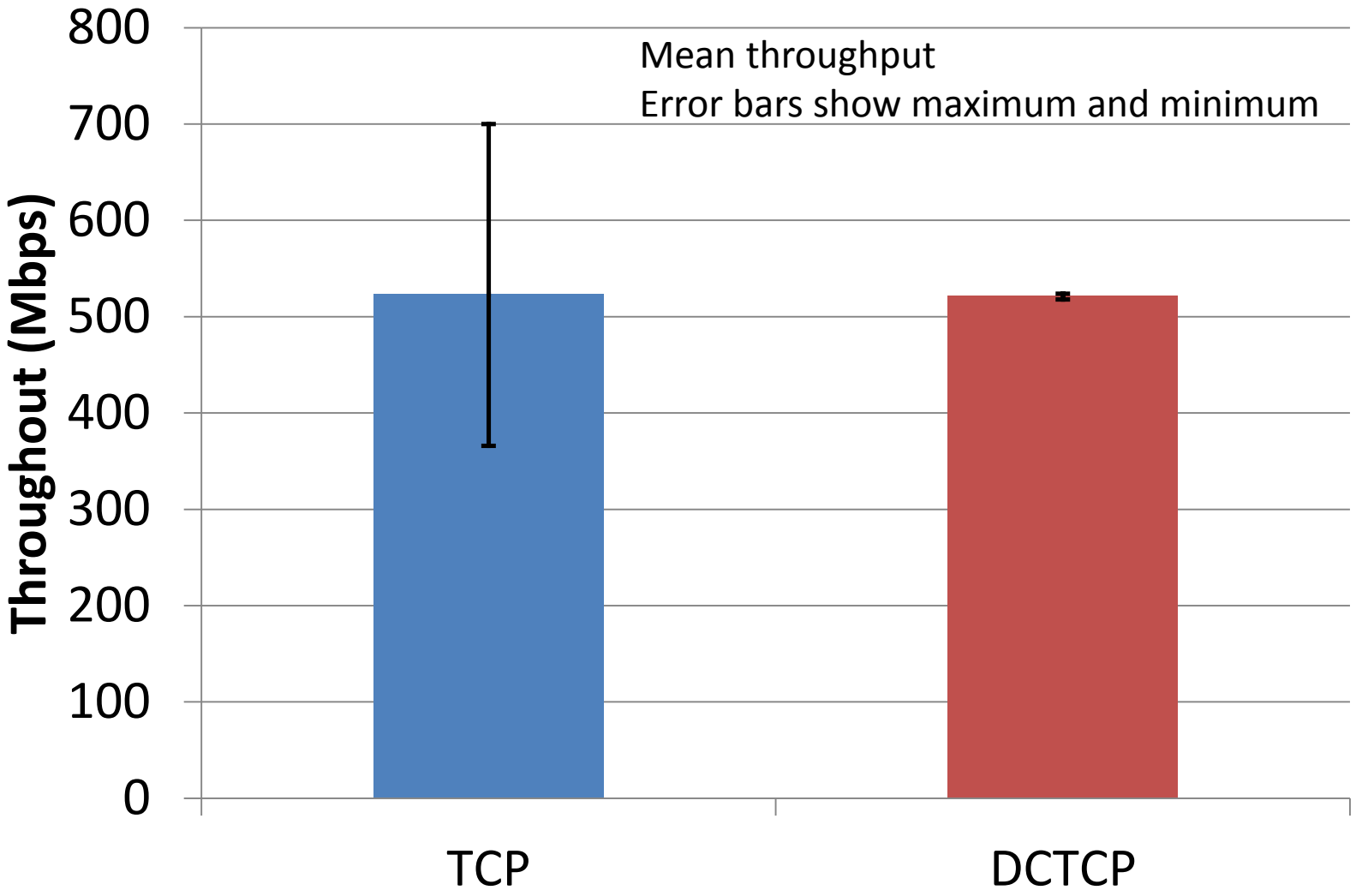
Outline

- Background
- TCP Datacenter Challenges
  - Latency & Application Coupling
  - Incast
  - DCTCP
    - Deployment Challenges
    - Performance
  - Receive Buffer Tuning
- Related Work
- Conclusion

Throughput, Fairness, & Latency



Throughput and Fairness



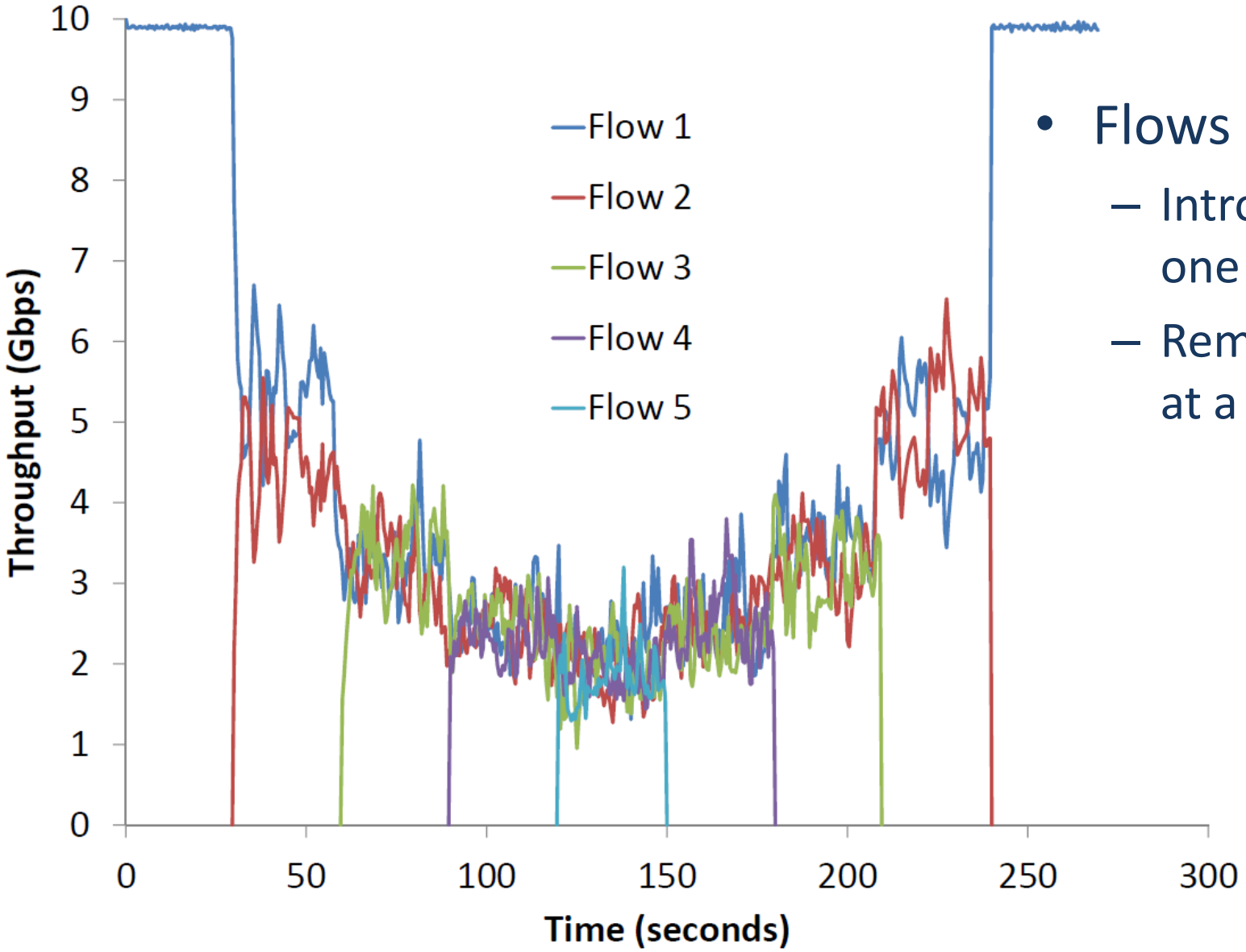
Latency

	TCP	DCTCP+
Mean	4.01	0.0422
Median	4.06	0.0395
Maximum	4.20	0.0850
Minimum	3.32	0.0280
σ	0.167	0.0106

Latency in ms

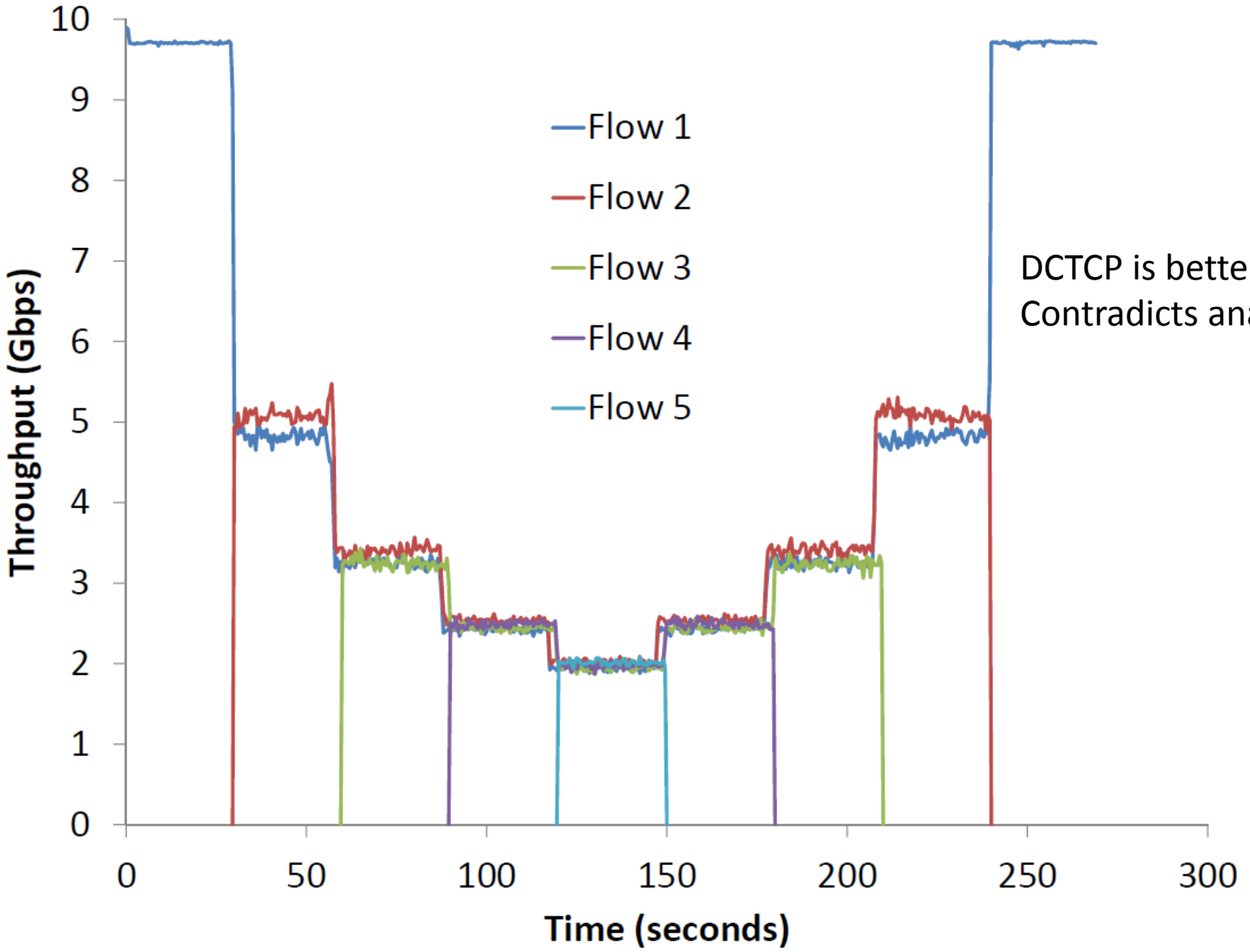
- DCTCP latency significantly lower than TCP

Stability, Convergence, & Fairness: TCP



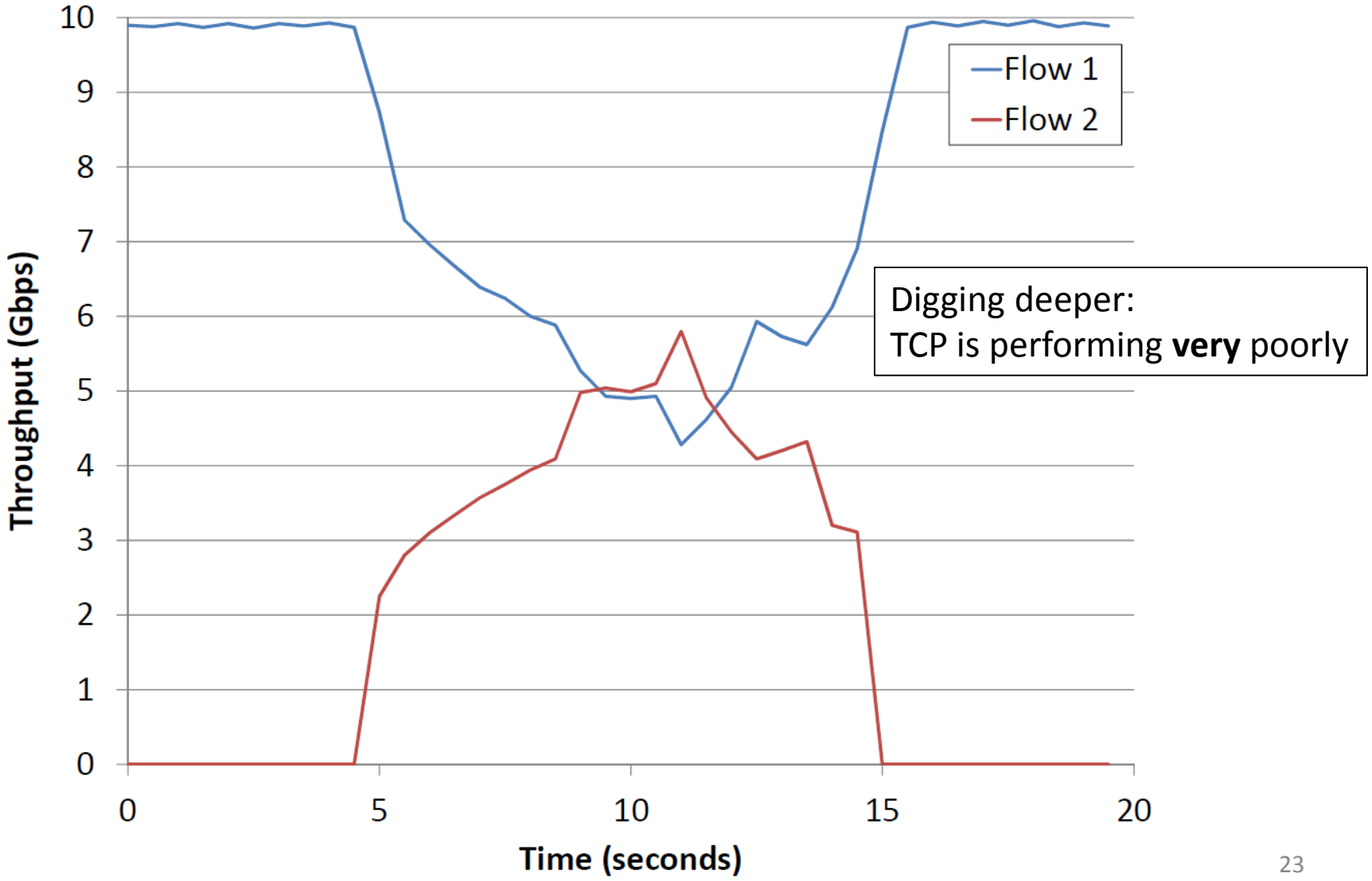
- Flows
  - Introduced one at a time
  - Removed one at a time

Stability, Convergence, & Fairness: DCTCP

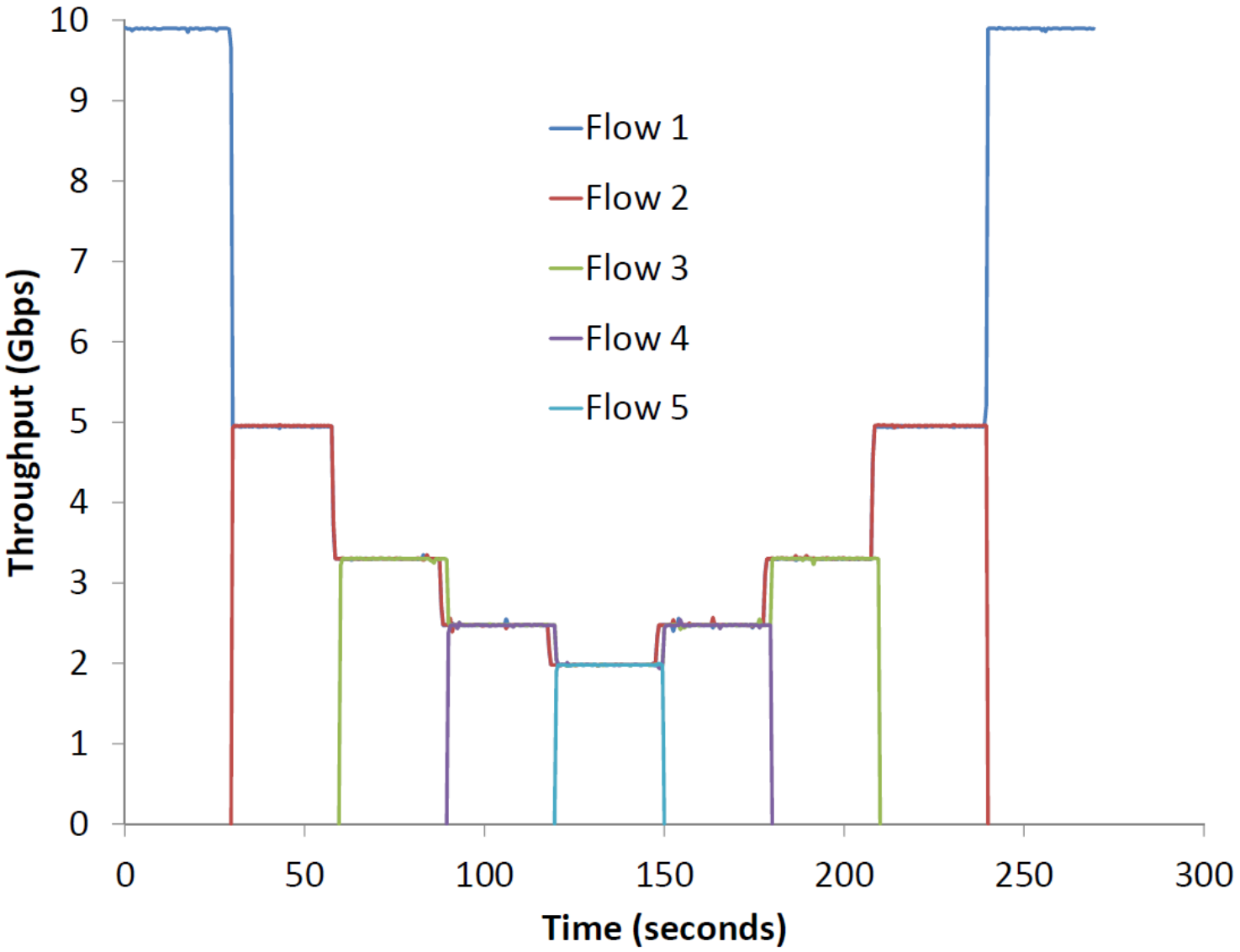


DCTCP is better!
Contradicts analytical results?

Stability, Convergence, & Fairness: TCP, 2 Flows



Receive Buffer Tuning Off: TCP



Scale

Senders	Total (Mbps)	RTT (ms)	Retransmissions		Timeouts	Min. Load (Gbps)
			Total	%		
100	9,901	1.6	0	0	0	3.27
200	9,900	3.11	0	0	0	6.55
300	9,901	4.38	3	0	0	9.82
400	9,894	4.42	702	4.6	274	13.09
500	9,895	4.44	1,110	8.7	655	16.36

Conclusion

- TCP has significant performance problems in datacenters
  - Latency & Application Coupling
  - Incast
    - Lowest feasible RTOmin
    - DCTCP
  - Receive buffer tuning
- DCTCP
  - Requires changes to work in production
    - Coexistence
    - Connection establishment
  - Has scaling limits, but significantly improves performance
  - In production
  - Coming soon to your favorite OS

Questions?