On the Data Path Performance of Leaf-Spine Datacenter Fabrics

Mohammad Alizadeh

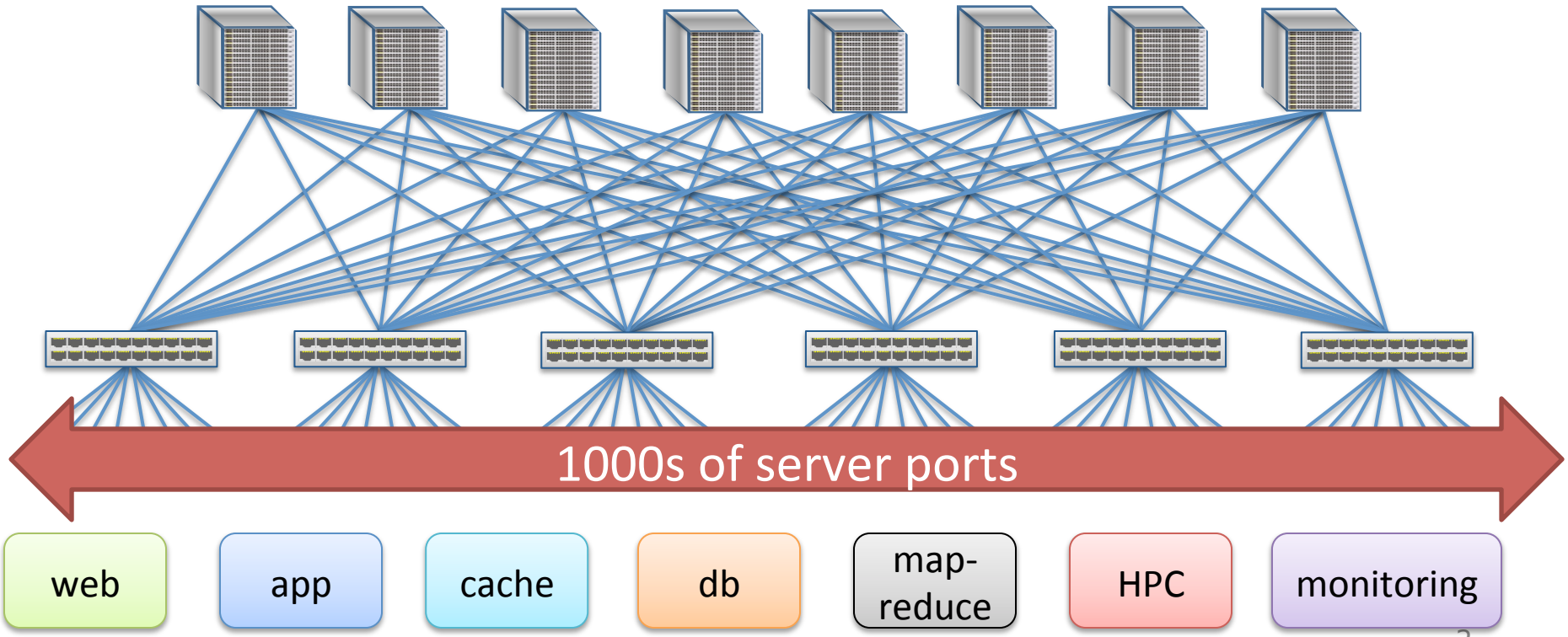
Joint with: Tom Edsall

1nsieme NETWORKS

insieme NETWORKS

Datacenter Networks

- **Must complete transfers or "flows" quickly**
  need very high throughput, very low latency



insieme NETWORKS

Leaf-Spine DC Fabric

Approximates ideal output-queued switch

- How close is Leaf-Spine to ideal OQ switch?
- What impacts its performance?
  - Link speeds, oversubscription, buffering

insieme NETWORKS

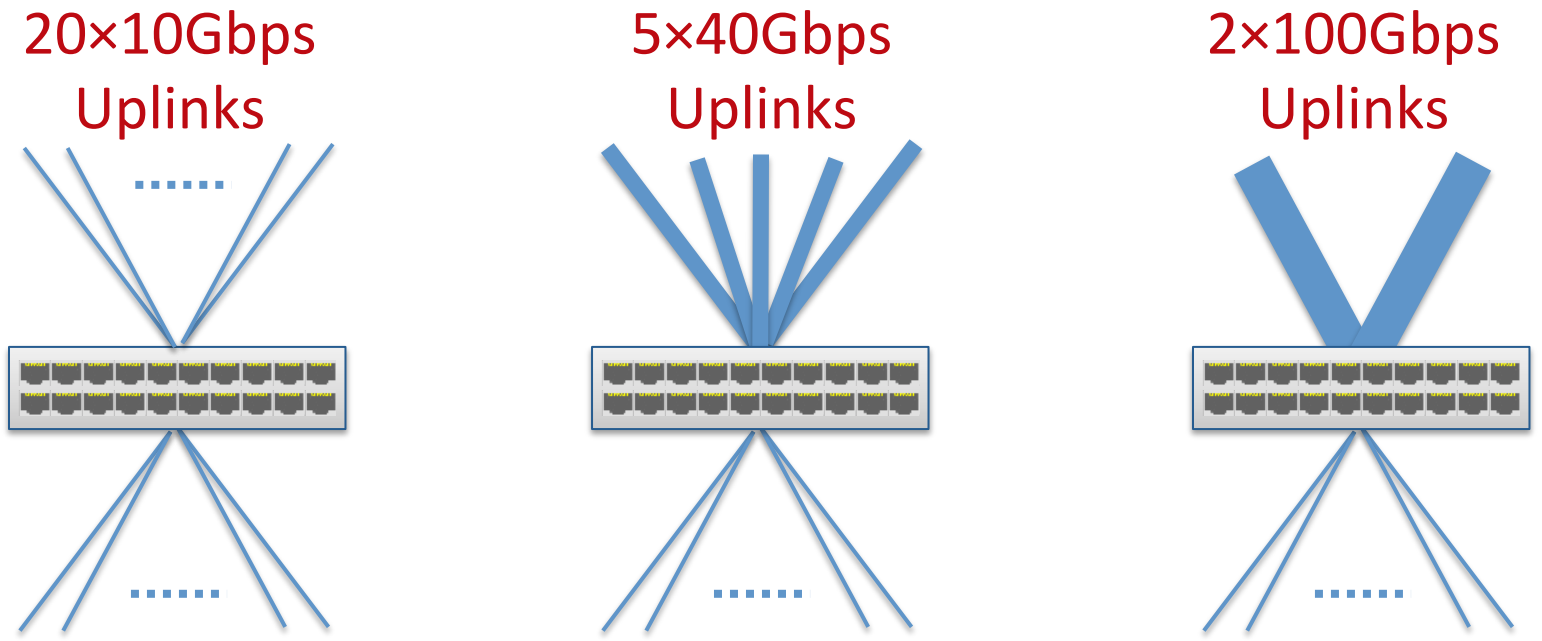
Methodology

- Widely deployed mechanisms
  - TCP-Reno+Sack
  - DropTail (10MB shared buffer per switch)
  - ECMP (hash-based) load balancing
- OMNET++ simulations
  - 100×10Gbps servers (2-tiers)
  - Actual Linux 2.6.26 TCP stack
- Realistic workloads
  - Bursty query traffic with Incast pattern
  - All-to-all background traffic: web search, data mining

Metric:
Flow completion time

Impact of Link Speed

Three non-oversubscribed topologies:



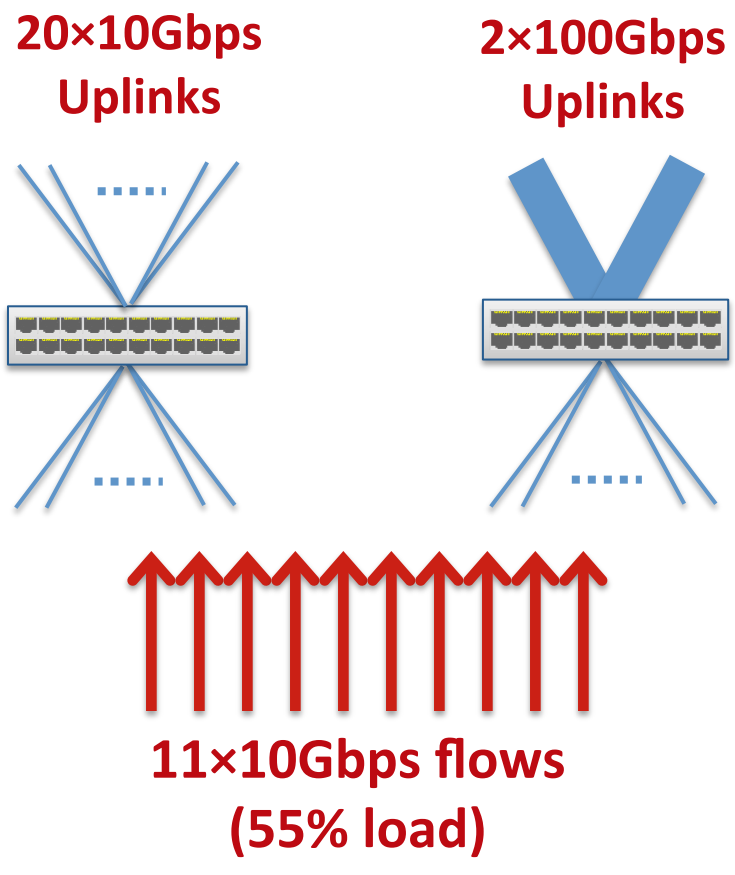
insieme NETWORKS

Impact of Link Speed

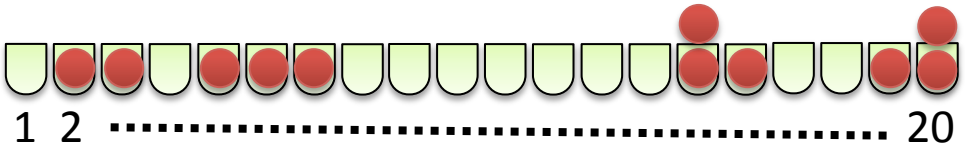
- 40/100Gbps fabric: ~ same FCT as OQ
- 10Gbps fabric: FCT up 40% worse than OQ

Intuition

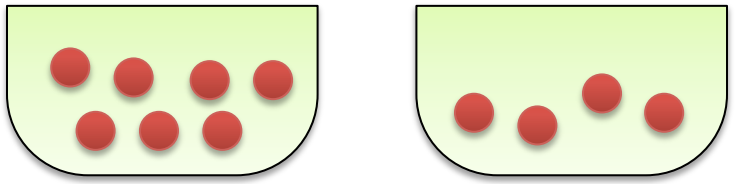
Higher speed links improve ECMP efficiency



Prob of 100% throughput = 3.27%



Prob of 100% throughput = 99.95%



Impact of Buffering

Where do queues build up?

2.5:1 oversubscribed fabric with large buffers

- Leaf fabric ports queues ≈ Spine port queues
  - 1:1 port correspondence; same speed & load

Impact of Buffering

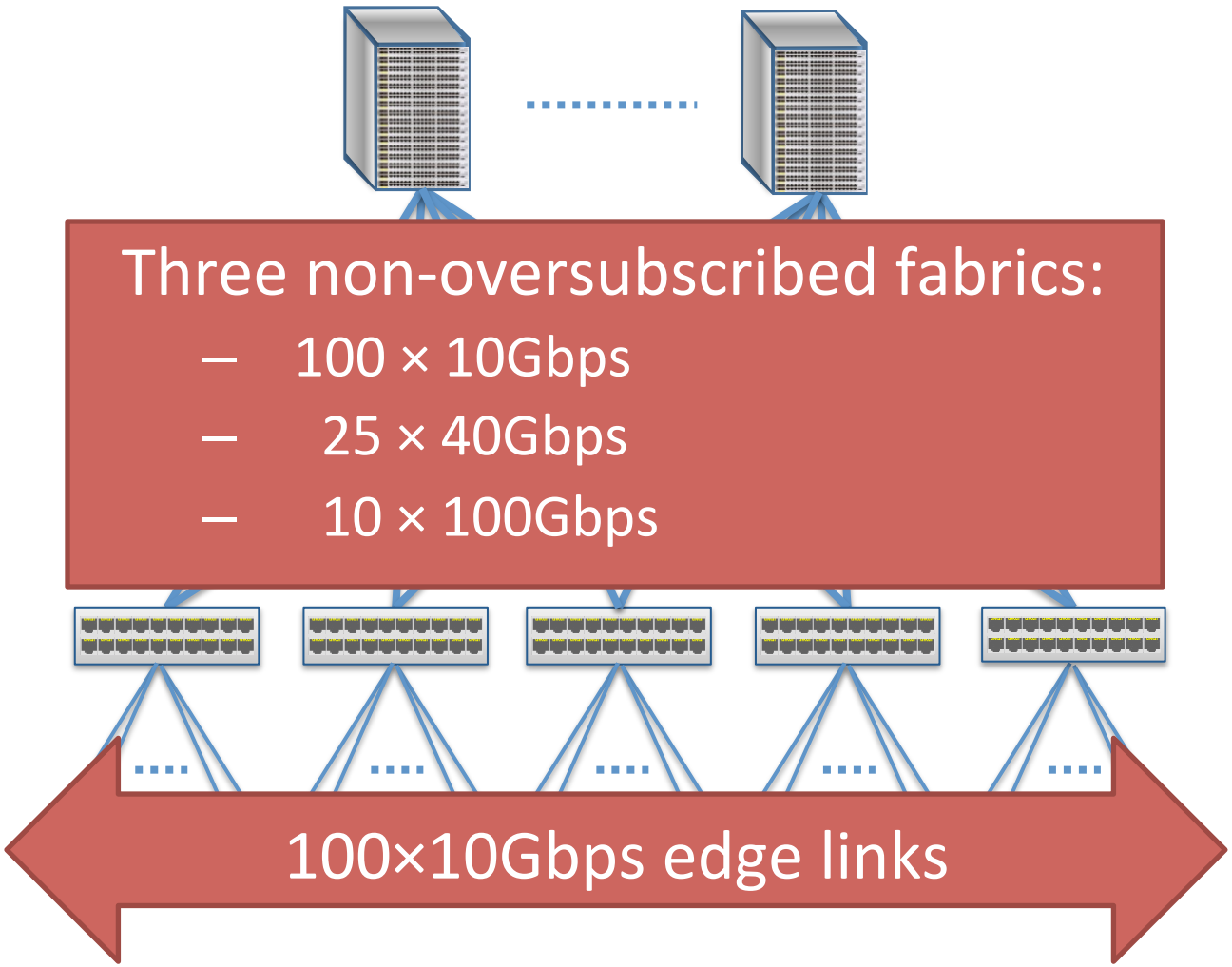
Where are large buffers more effective for Incast bursts?

- Larger buffers better at Leaf than Spine

Summary

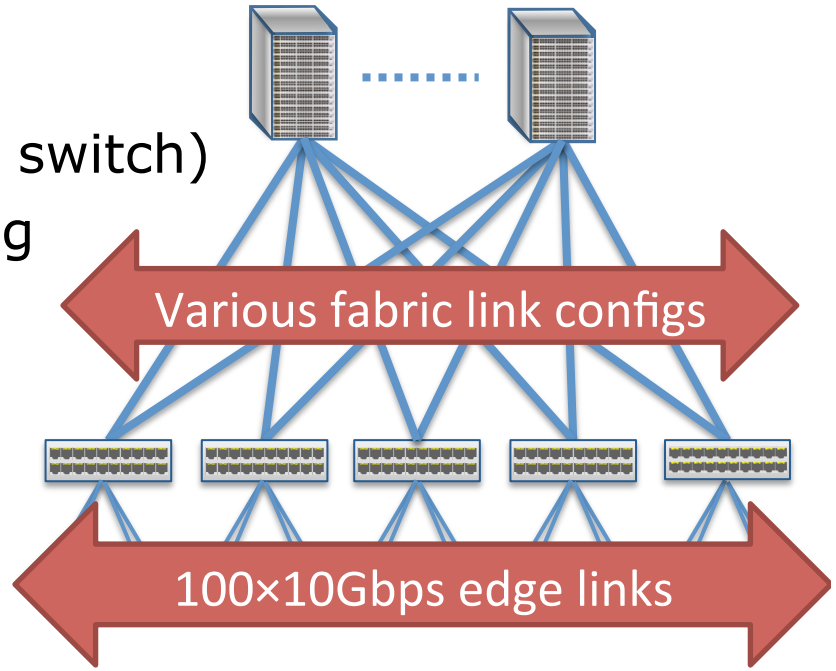
- 40/100Gbps fabric + ECMP ≈ OQ-switch; some performance loss with 10Gbps fabric
- Buffering should generally be consistent in Leaf & Spine tiers
- Larger buffers more useful in Leaf than Spine for Incast

Impact of Link Speed



Methodology

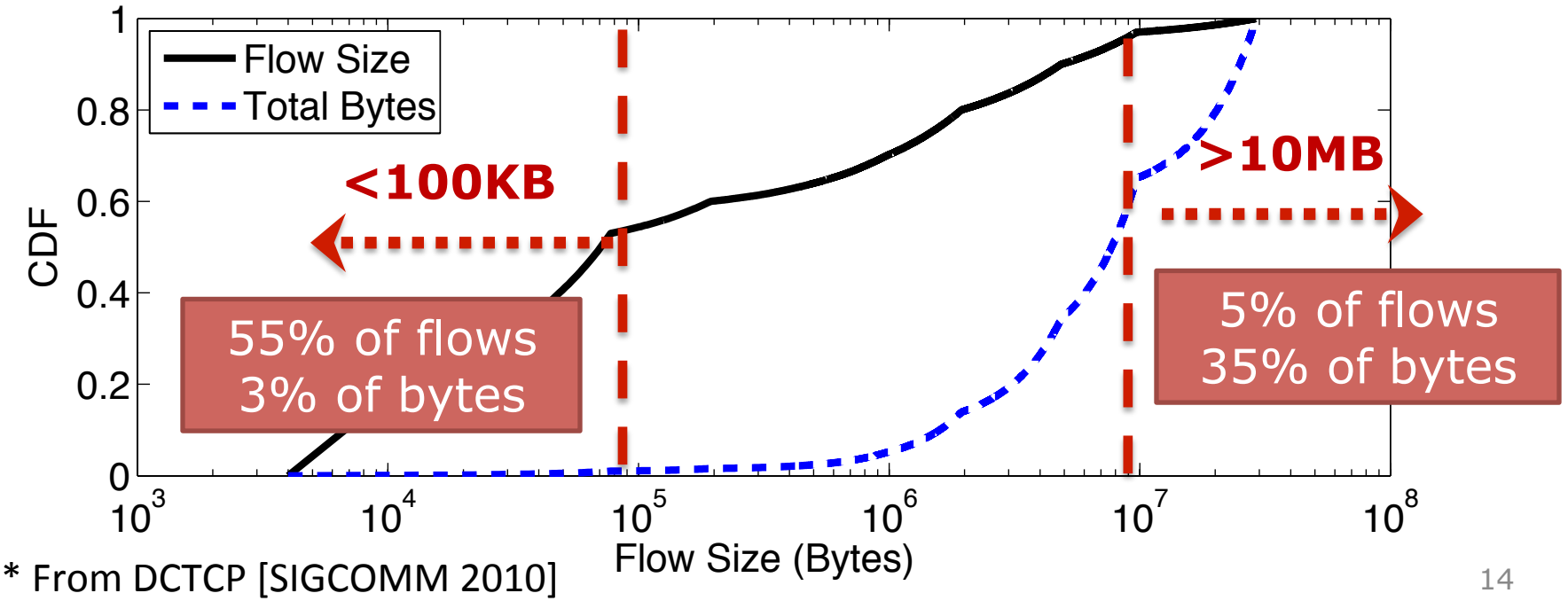
- Widely deployed mechanisms
  - TCP-Reno+Sack
  - DropTail (10MB shared buffer per switch)
  - ECMP (hash-based) load balancing
- OMNET++ simulations
  - 100×10Gbps servers (2-tiers)
  - Actual Linux 2.6.26 TCP stack
- Realistic workloads
  - Bursty query traffic
  - All-to-all background traffic: web search, data mining



Metric:
Flow completion time

Workloads

- Realistic workloads based on empirical studies
  - Query traffic with Incast pattern
  - All-to-all background traffic



* From DCTCP [SIGCOMM 2010]

Intuition

Incast events are most severe at receiver