Microsoft

Larry: Practical Network Reconfigurability in the Data Center

Andromachi Chatzieleftheriou, Sergey Legtchenko, Hugh Williams, Antony Rowstron

Microsoft Research

Large-scale data center networking

Increasing bandwidth per server: 40-100Gbps

Full bisection bandwidth not cost efficient

Oversubscription

- tradeoff between performance and cost

Do we need to provision for peak performance?

Network flexibility

Highly skewed rack-level traffic demand [ProjectToR'16, Facebook'15]

Dynamically adapt to traffic demand

- Full reconfigurability [OSA, ProjecToR, Helios, etc.]
- Traffic engineering [Jellyfish, Xpander, etc.]

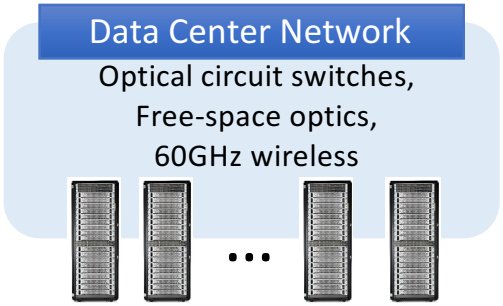
Network flexibility

Highly skewed rack-level traffic demand [ProjectToR'16, Facebook'15]

Dynamically adapt to traffic demand

Full reconfigurability [OSA, ProjecToR, Helios, etc.] → Hard deployment and operation

Traffic engineering [Jellyfish, Xpander, etc.]



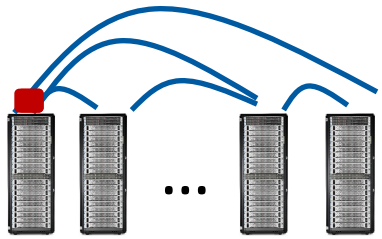
Network flexibility

Highly skewed rack-level traffic demand [ProjectToR'16, Facebook'15]

Dynamically adapt to traffic demand

Full reconfigurability [OSA, ProjecToR, Helios, etc.] → Hard deployment and operation

Traffic engineering [Jellyfish, Xpander, etc.] → Often higher end-to-end latency



Network flexibility

Highly skewed rack-level traffic demand [ProjectToR'16, Facebook'15]

Dynamically adapt to traffic demand

- Full reconfigurability [OSA, ProjecToR, Helios, etc.] → Hard deployment and operation
- Traffic engineering [Jellyfish, Xpander, etc.] → Often higher end-to-end latency

→ Real-world implementation challenges

Our solution: local reconfigurability

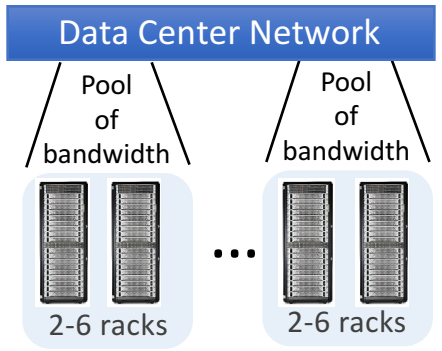
Reconfiguration at a scale of *2 to 6 racks*

Transparency and incremental deployment

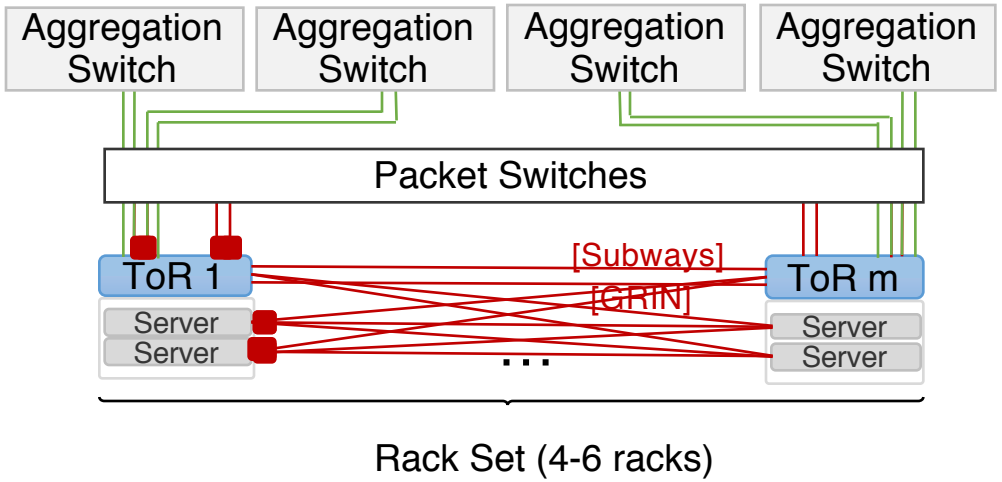
Use *spare* uplink bandwidth from adjacent racks

Key insights

- Spare ToR ports
- Rack-level traffic: loosely corelated & bursty



First approach: Non-shortest path routing



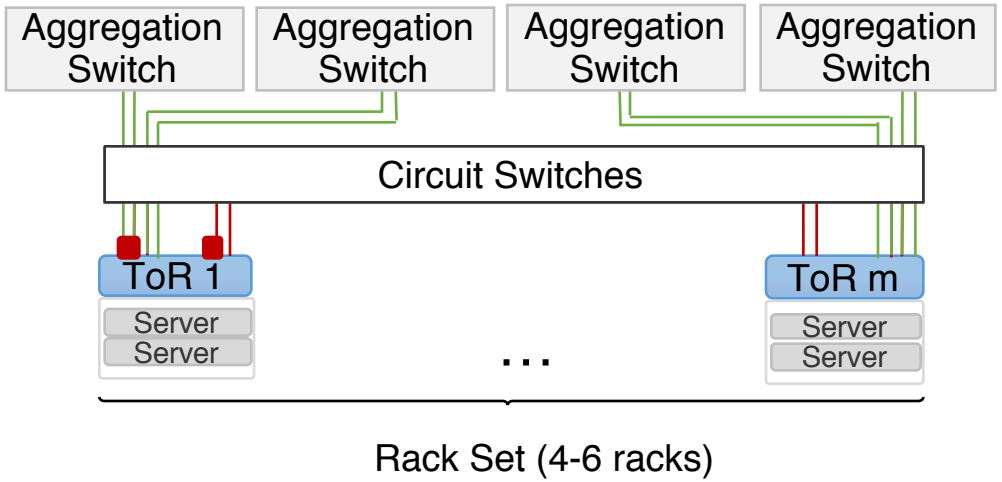
extra forwarding latency

complicated control plane

additional ports

possible fate sharing

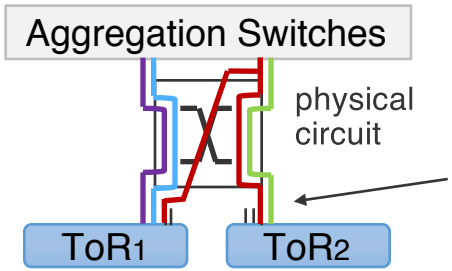
Larry: reconfiguration at the physical layer



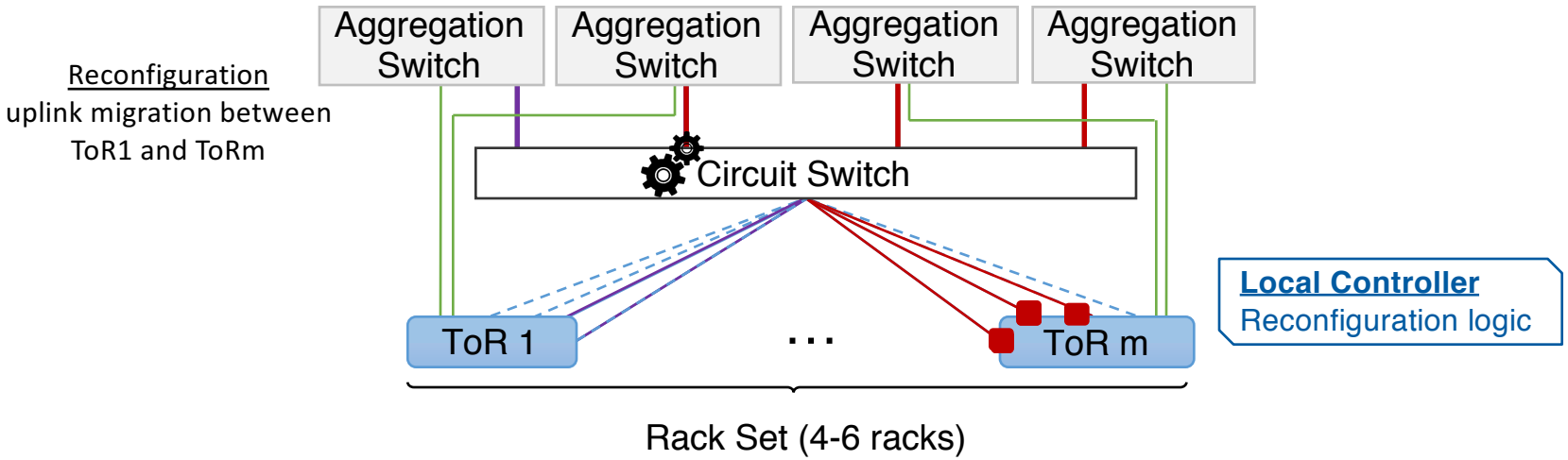
Low-cost electrical circuit switch

- electrical signal forwarding
- no buffering or packet processing

→ low and predictable forwarding latency overhead



Larry architecture



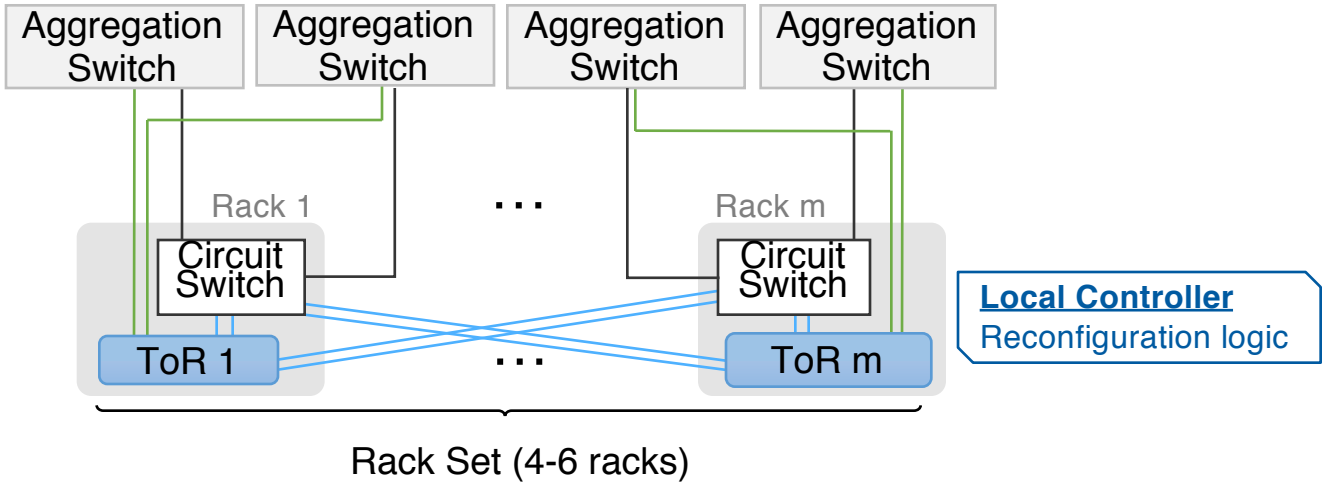
Uplink migration within small sets of racks

No additional cables to aggregation switches

→ But, single switch not practical

- ✓ Transparency
- ✓ Incremental deployment
- ✓ Failure resilience
- ✓ Cost efficiency

Larry architecture



Distributing the circuit switch

- small-radix circuit switch per rack (32 ports)

- ✓ Transparency
- ✓ Incremental deployment
- ✓ Failure resilience
- ✓ Cost efficiency

Control plane

Small *local* soft-state controller

- traffic demand monitoring
- reconfiguration management

Uplink migration to decongest ToRs

- byte count metrics per port received from ToRs
- conservative by design

Explicit or implicit link failure management

Performance evaluation

Small testbed using prototype circuit switch

Flow-based simulations

- 32-port @100Gbps ToR and circuit switches
- 32 servers per rack @50Gbps
- 4 racks per rack set

Workloads: Microbenchmarks, storage & production traces

Metrics: switching time, flow completion time, performance/$

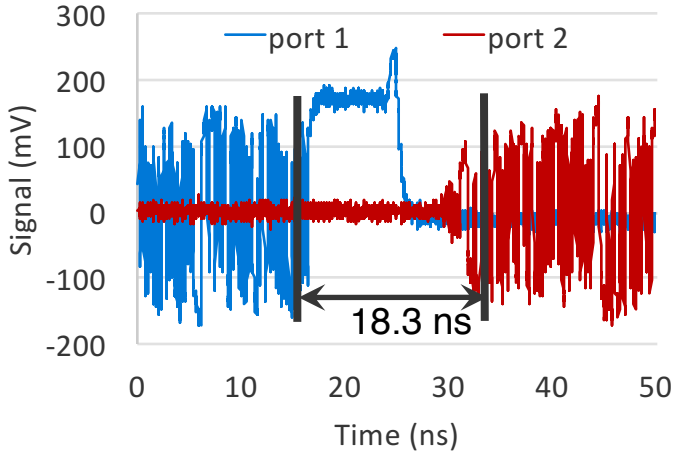
Prototype of rack-scale circuit switch

40 ports @ 40Gbps

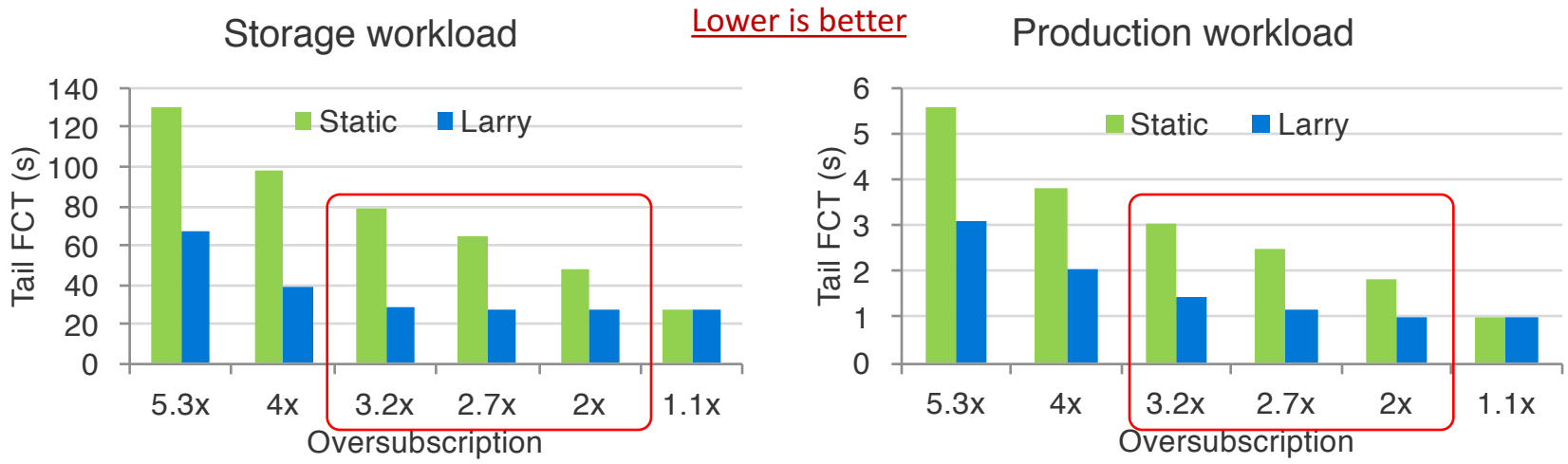
Performance evaluation

Overhead to establish a circuit

Micro-benchmarking



99th percentile flow completion time

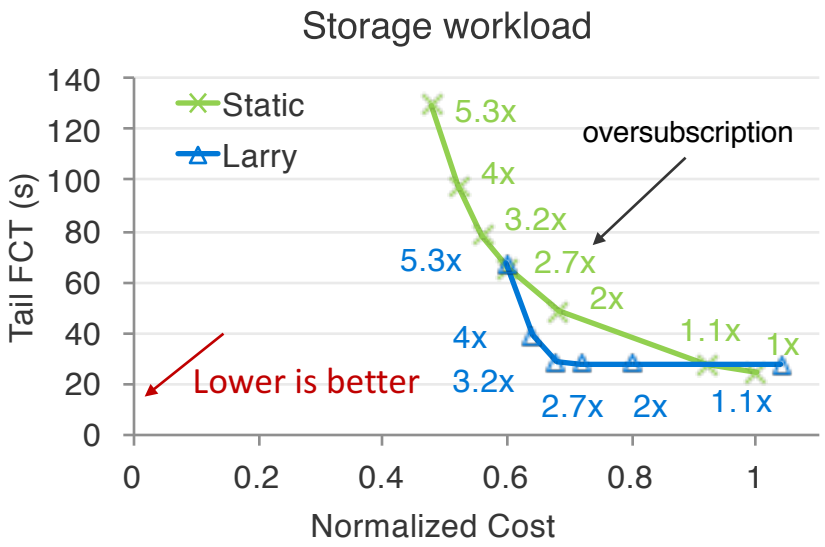


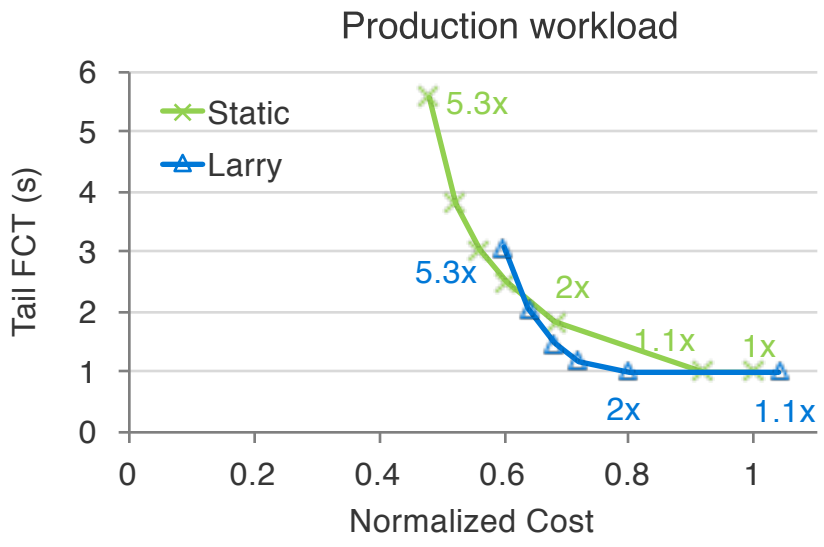
Tail FCT as a function of oversubscription

Sweep # of uplinks to aggregation switches provisioned per rack

→ Improved tail FCT even at high oversubscription

Performance per $





DC network deployment cost normalized to fully provisioned topology

Improved performance up to 2.3x at the same network cost

Conclusion

Local reconfigurability at the physical layer within small sets of racks

Incremental and transparent deployment

Predictable low forwarding overhead

Improved performance per $