FRANCIS

Fast Reaction Algorithms for Network Coordination In Switches

Presenter: Wenchen Han (UCL)

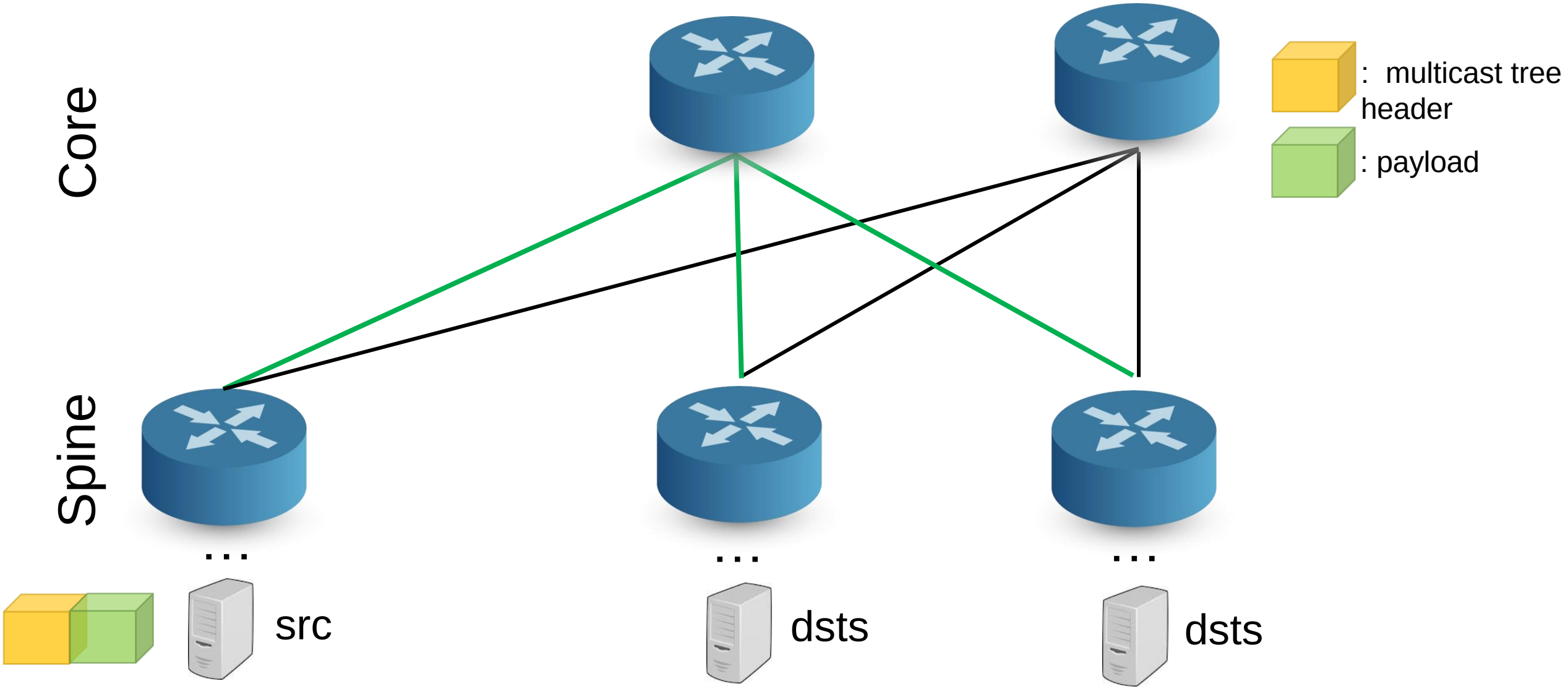
Joint work with: Vic Feng, Gregory Schwartzman, Yuliang Li,

Michael Mitzenmacher, Minlan Yu, **Ran Ben Basat**

The need of real-time network changes handling

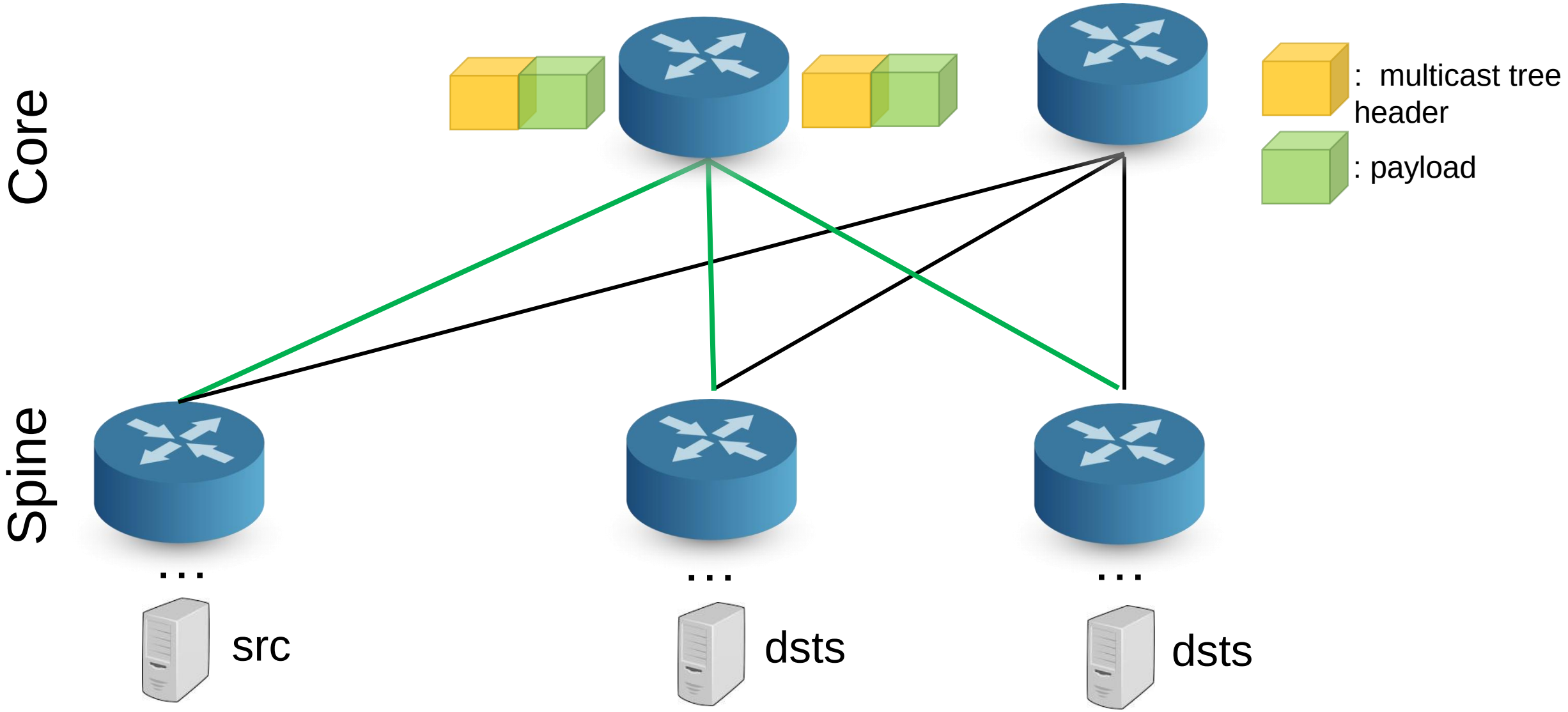
- **What** are network changes: link/switch failures, congestions, short traffic bursts, etc.
- **Why** real-time: minimizing their impact to applications (multicast, clock-sync, unicast).
- **SDN approaches**: O(100ms), inherently too slow for many applications.

For example: Elmo [1]'s source-routed multicast



[1] Shahbaz et al. Elmo: Source Routed Multicast for Public Clouds. In SIGCOMM 2019.

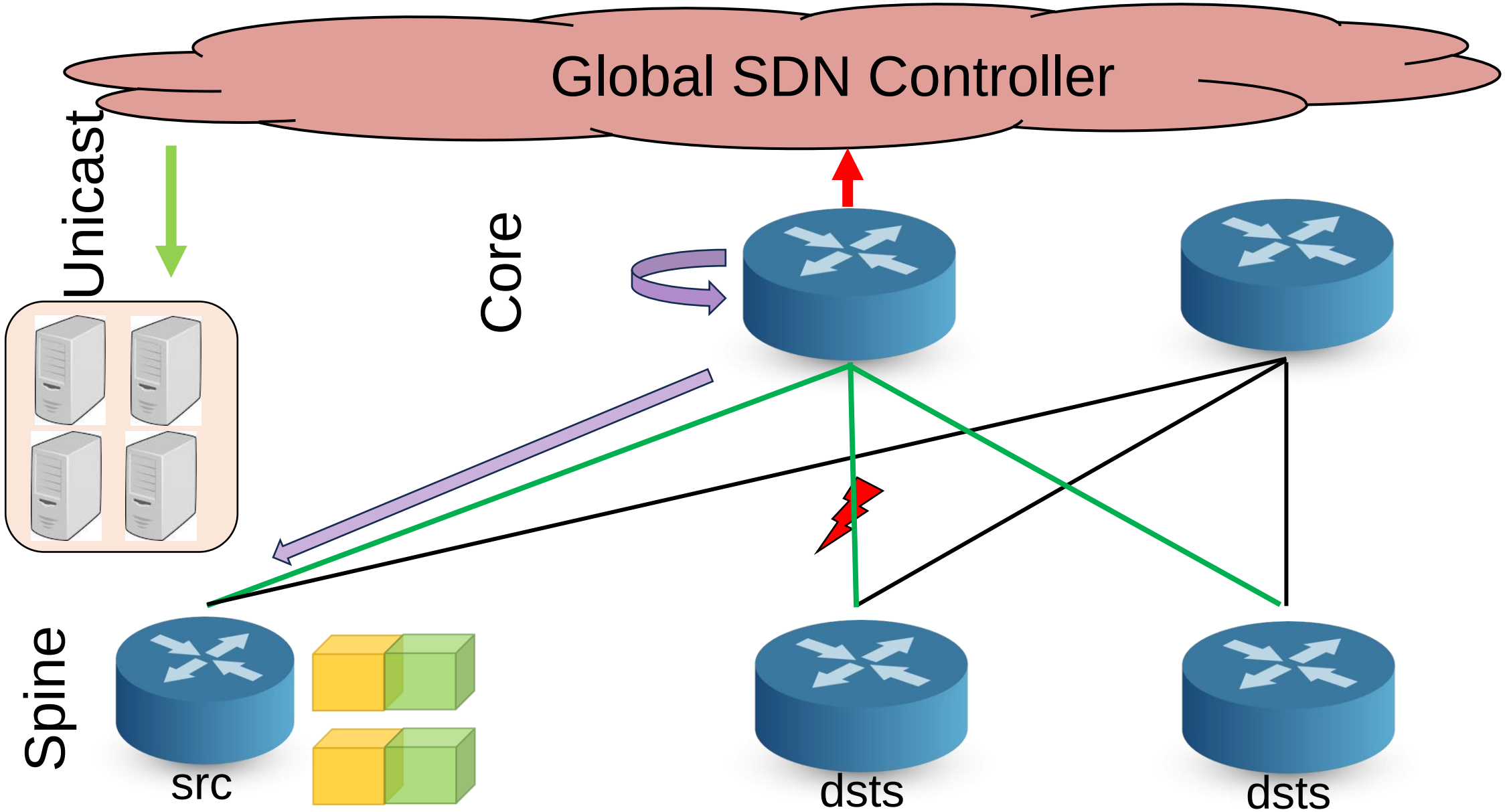
For example: Elmo [1]'s source-routed multicast



[1] Shahbaz et al. Elmo: Source Routed Multicast for Public Clouds. In SIGCOMM 2019.

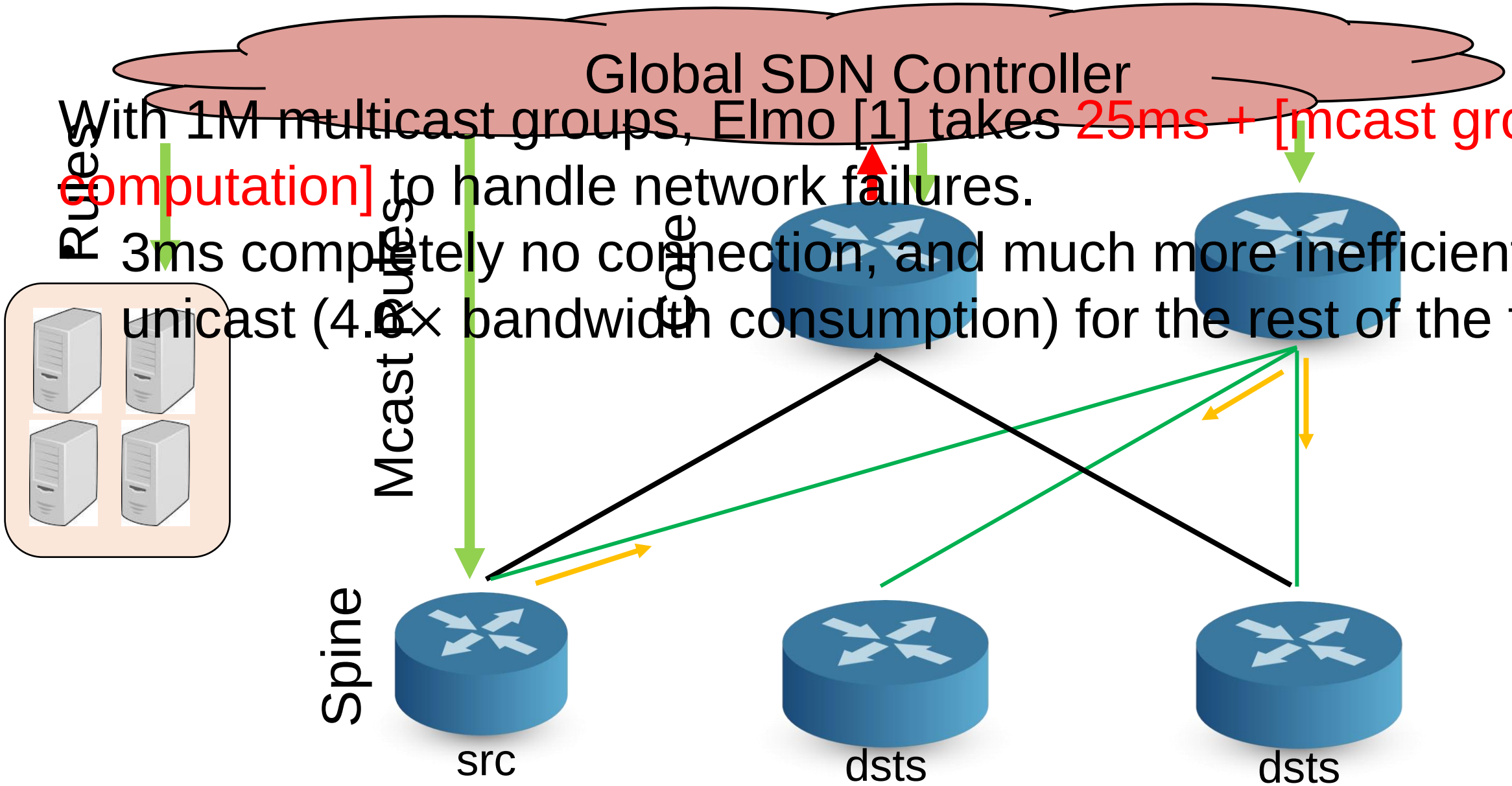
Elmo's SDN-based failure recovery

3ms (DP based)



Elmo's SDN-based failure recovery

O(25ms)

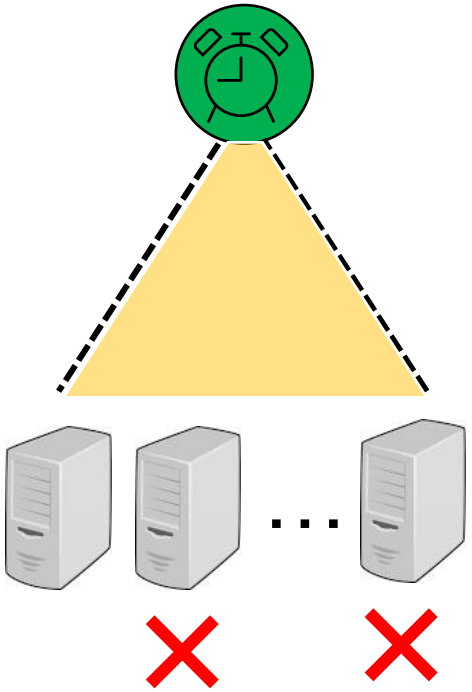


With 1M multicast groups, Elmo [1] takes 25ms + [mcast group computation] to handle network failures.

3ms completely no connection, and much more inefficient unicast (4.5× bandwidth consumption) for the rest of the time.

[2] Li et al. Sundial: Sundial: Fault-tolerant Clock Synchronization for Datacenters. In OSDI 2020.

Clock Synchronization

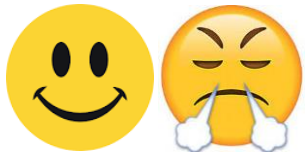


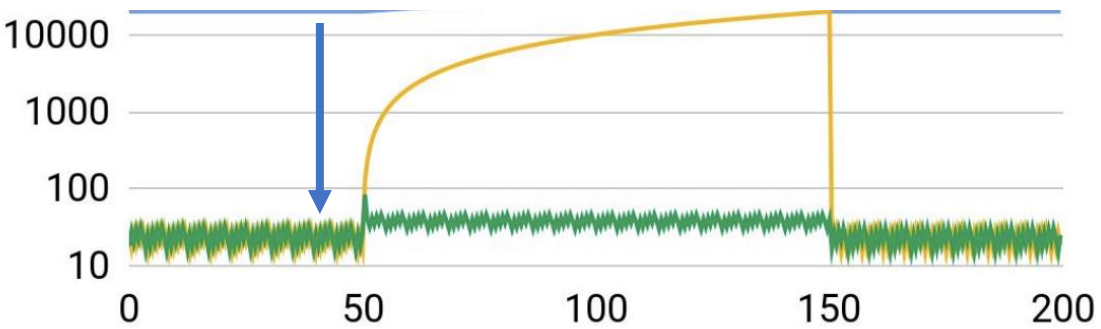
Clock Sync

Distributed Transactional Database

Network Telemetry

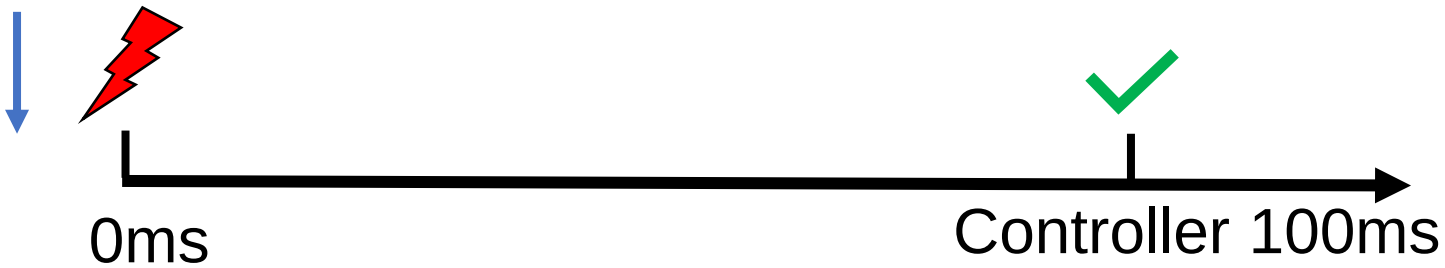
…



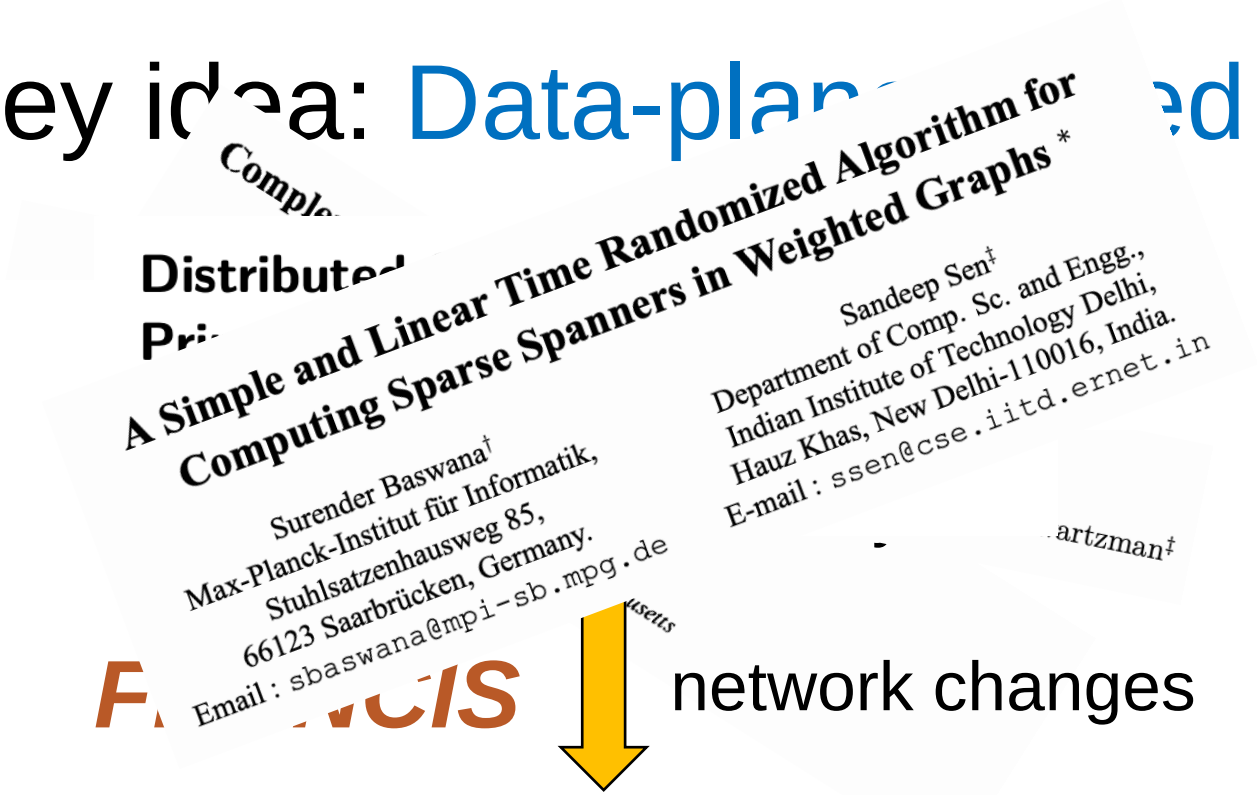


[2]

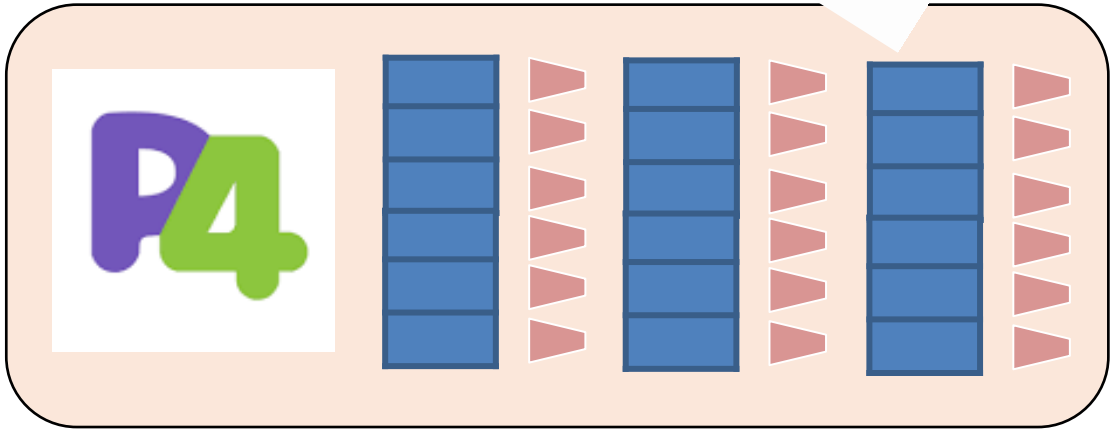
Needs uniformly small ϵ!



Key idea: Data-plane distributed algorithms



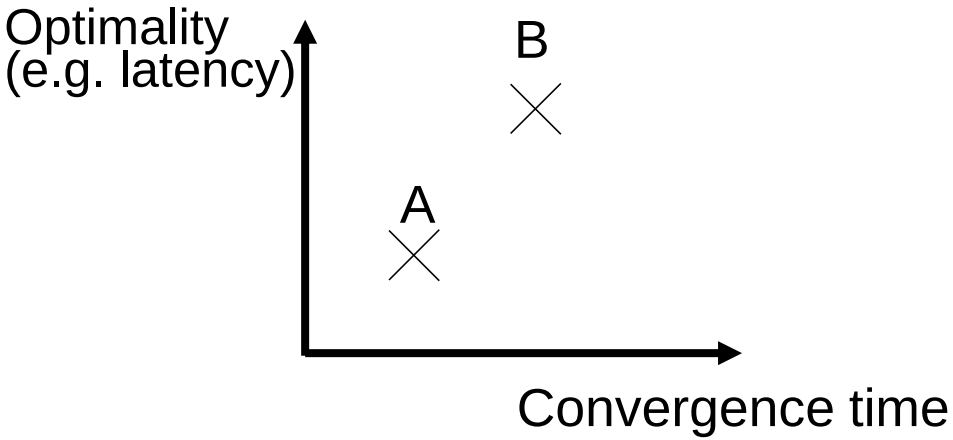
network changes



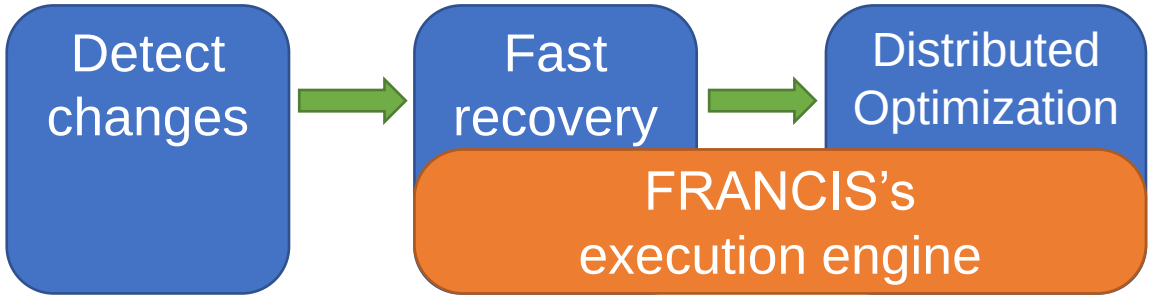
- **Benefits**: Fast reaction.
  - O(100) ms → O(1) ms
- **Bridging** the gap between theory community and system in practice (bidirectional).

The principle of choosing algorithms

- Fast recovery for reactivity, and dist optimization for optimality.

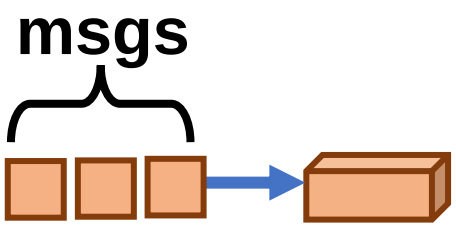


	Conv time	latency
A: One spanning tree for ***all*** groups.	$289\mu s$	$1.4\mu s$
B: One mcast tree for ***each*** group.	$1481\mu s$	$0.9\mu s$



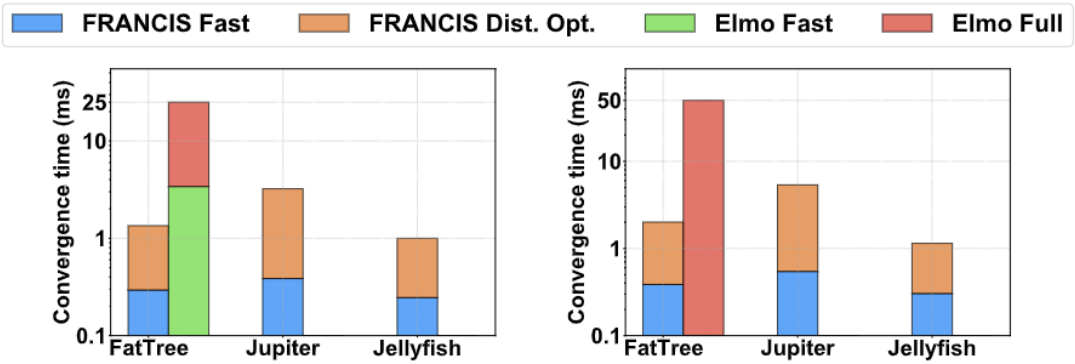
FRANCIS's Design overview

Achieving generality	☑ Supporting both Async and Sync algorithms
	☑ $\mathrm{Bw} \leq B, \mathrm{Mem} \leq M$ B and M are user-specified.
Execution on switches	☑ Overcoming Tofino switch restrictions
	☑ Packet loss handling
	☑ Message-packing for reducing bandwidth consumption



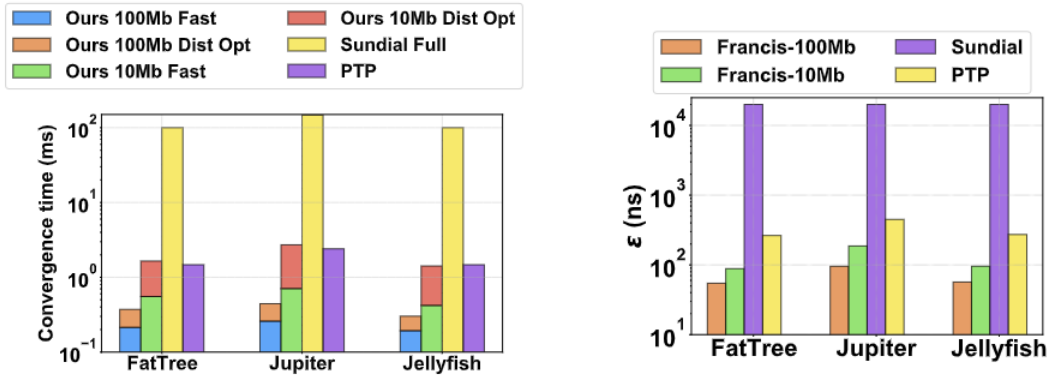
Benefits: fast reaction

- FRANCIS's Data-plane-based approaches converge 1-2 orders of magnitude faster!



(a) 1 link failure. (b) 2 link failures.

multicast



(a) Convergence time. (b) Peak ϵ.

Clock synchronization

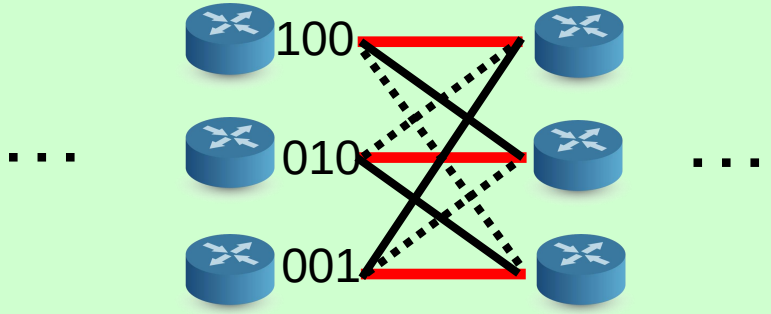
100Mbps bandwidth, 1.5% SRAM usage

Bridging the gap between theory and practice

Theory

- Distributed spanner algorithm: find $G' \subset G$ with minimum #edges such that $d_{G'}(x, y) \leq \lambda d_G(x, y)$.

Source-routed multicast



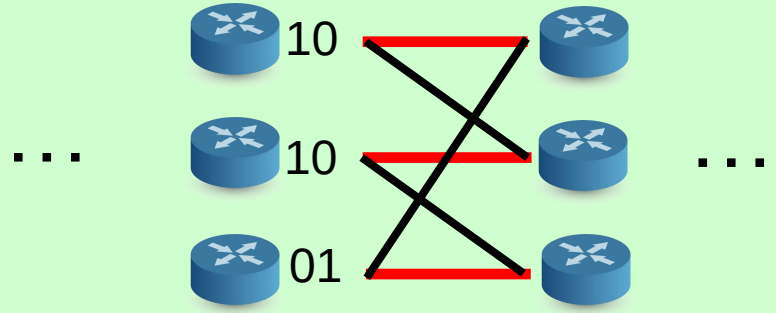
Bridging the gap between theory and practice

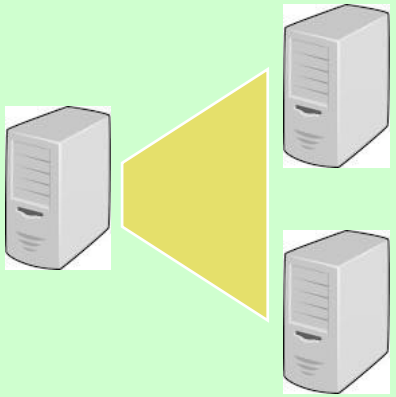
Theory

- Distributed spanner algorithm: find $G' \subset G$ with minimum #edges such that $d_{G'}(x, y) \leq \lambda d_G(x, y)$.

- Minimum size shortest path tree.

Source-routed multicast





- ↓ #Links, s.t.
- Lowest latency

Thank you for your attention!

Email: wenchen.han.22@ucl.ac.uk

Q&A?