Resolving Consensus

Benchmarking distributed key value stores on arbitrary network configurations

Chris Jensen (University of Cambridge, cjj39@cam.ac.uk)

Prior Work

- Specific installations tested
- Homogenous hosts
- Singular failure trace

Aims

Evaluate arbitrary deployments

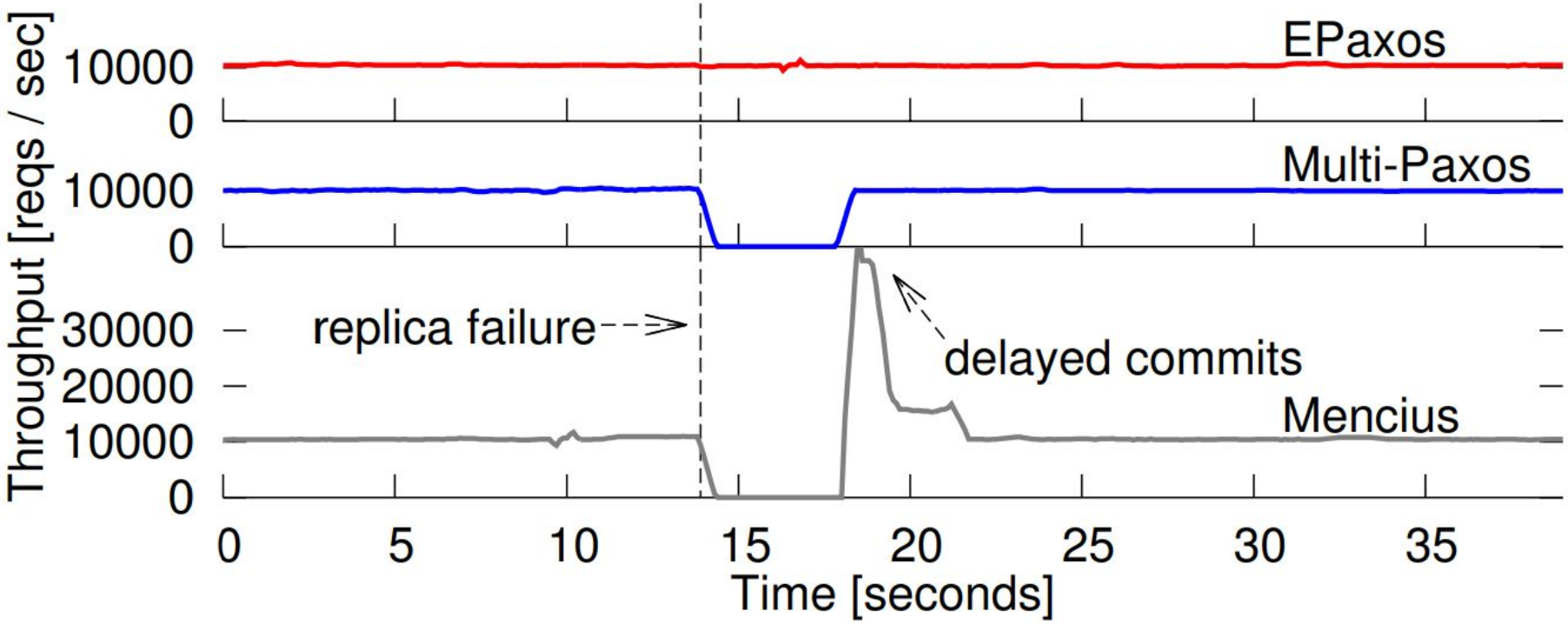
- Homogenous hosts
- Singular failure trace

Aims

Evaluate arbitrary deployments

Heterogeneous hosts

- Singular failure trace



Moraru, Iulian, David G. Andersen, and Michael Kaminsky. "There is more consensus in egalitarian parliaments." *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*. 2013.

Aims

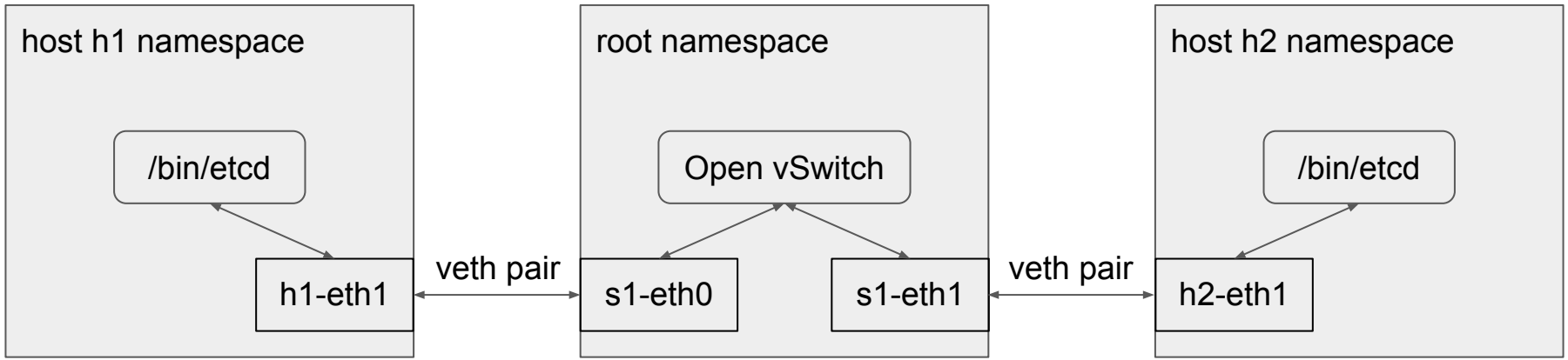
Evaluate arbitrary deployments

Heterogeneous hosts

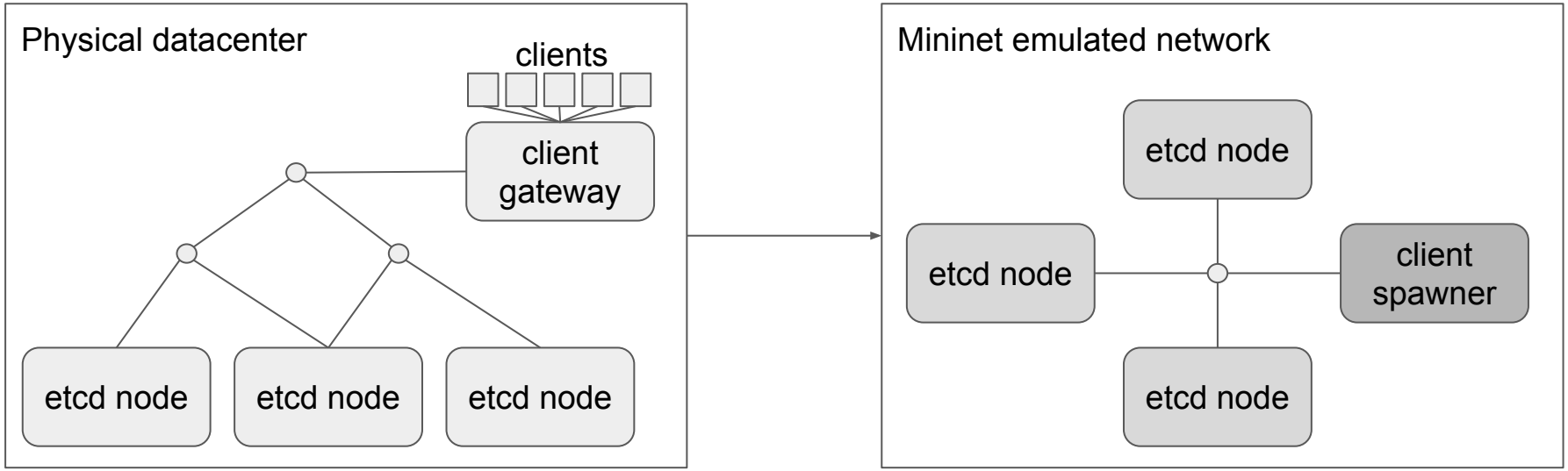
Comprehensive failure analysis

Mininet deployment emulation

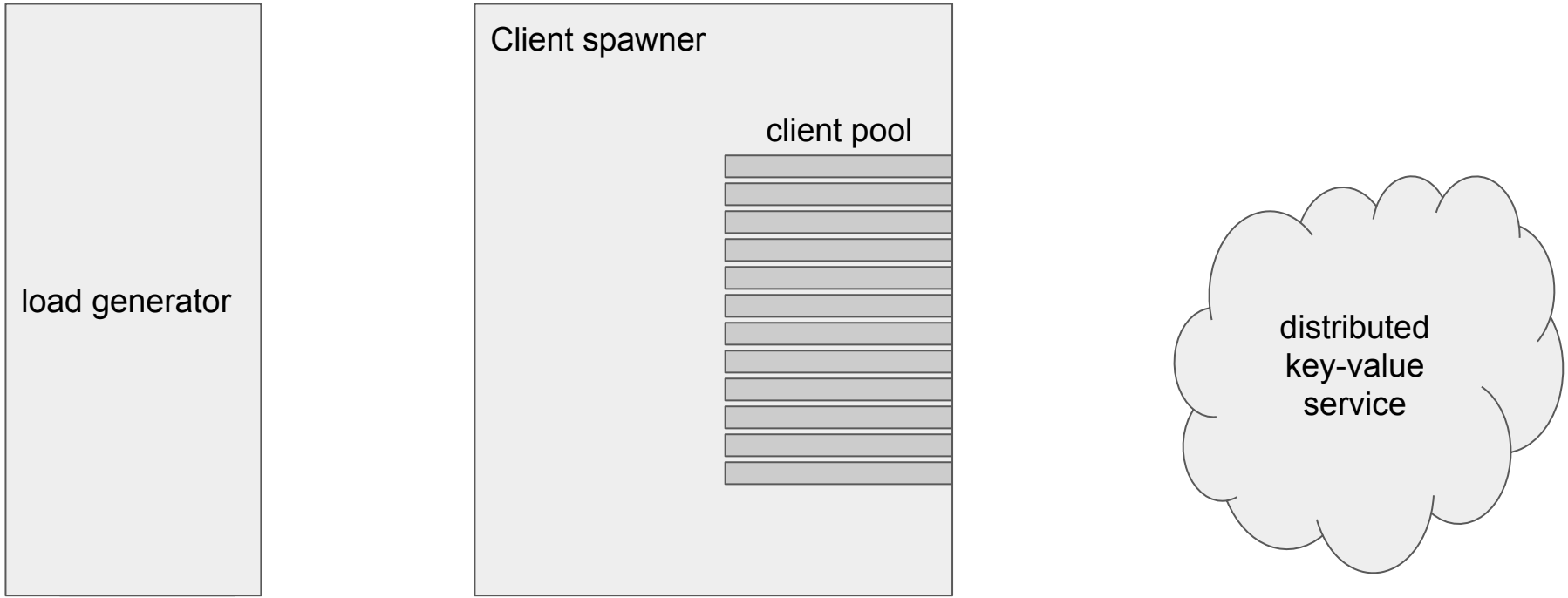
- Network emulation via network namespaces and Open vSwitch
- Heterogeneous host emulation via cgroups



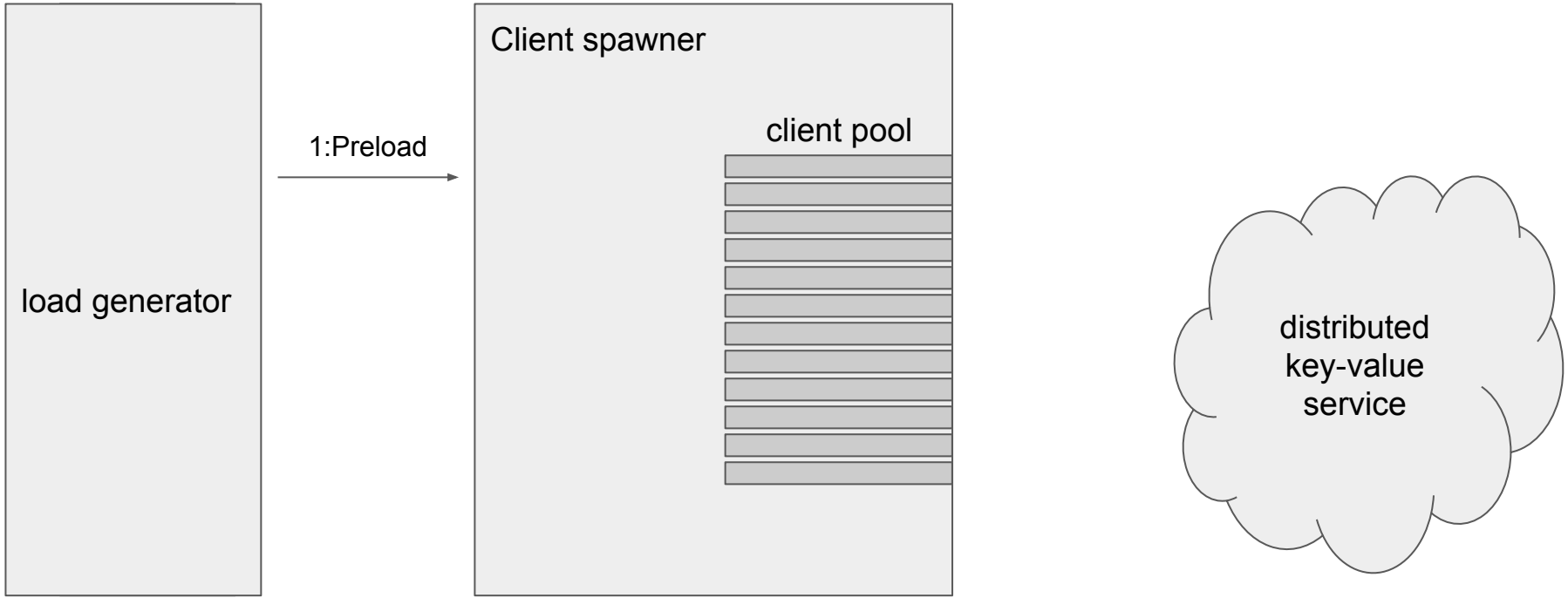
How we emulate topologies



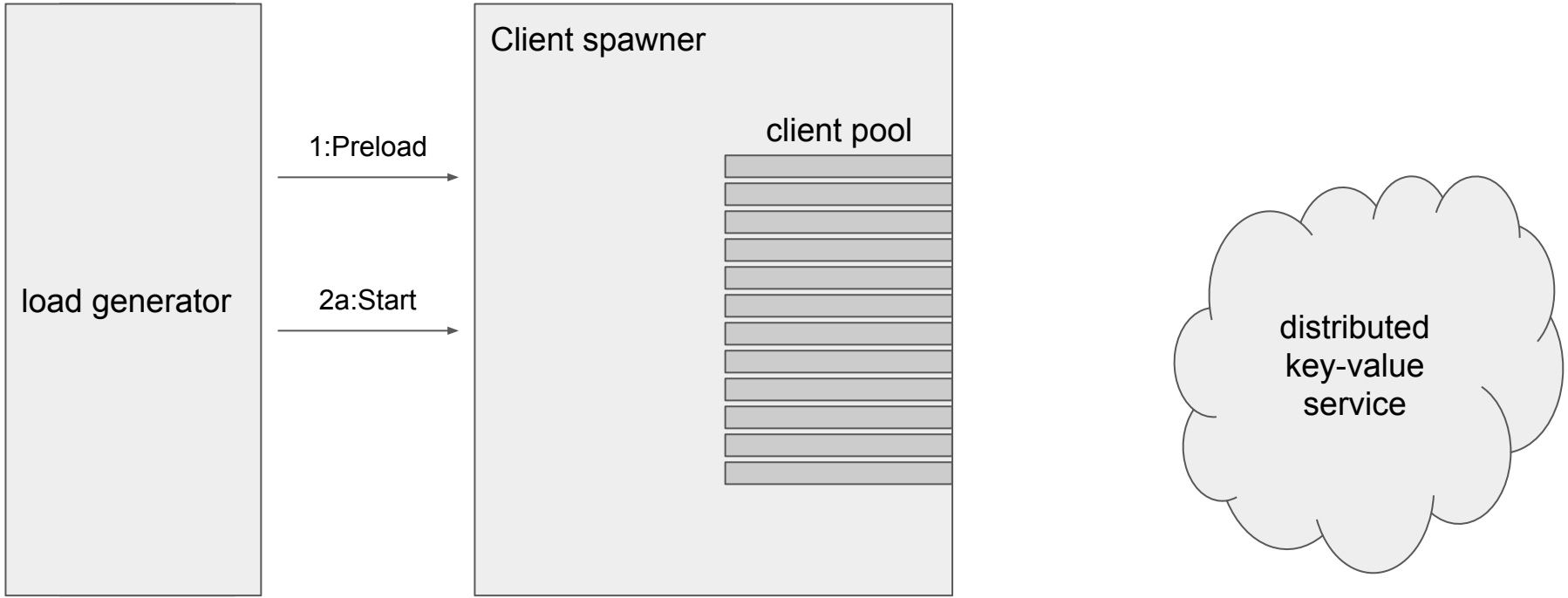
Load generator and client spawner



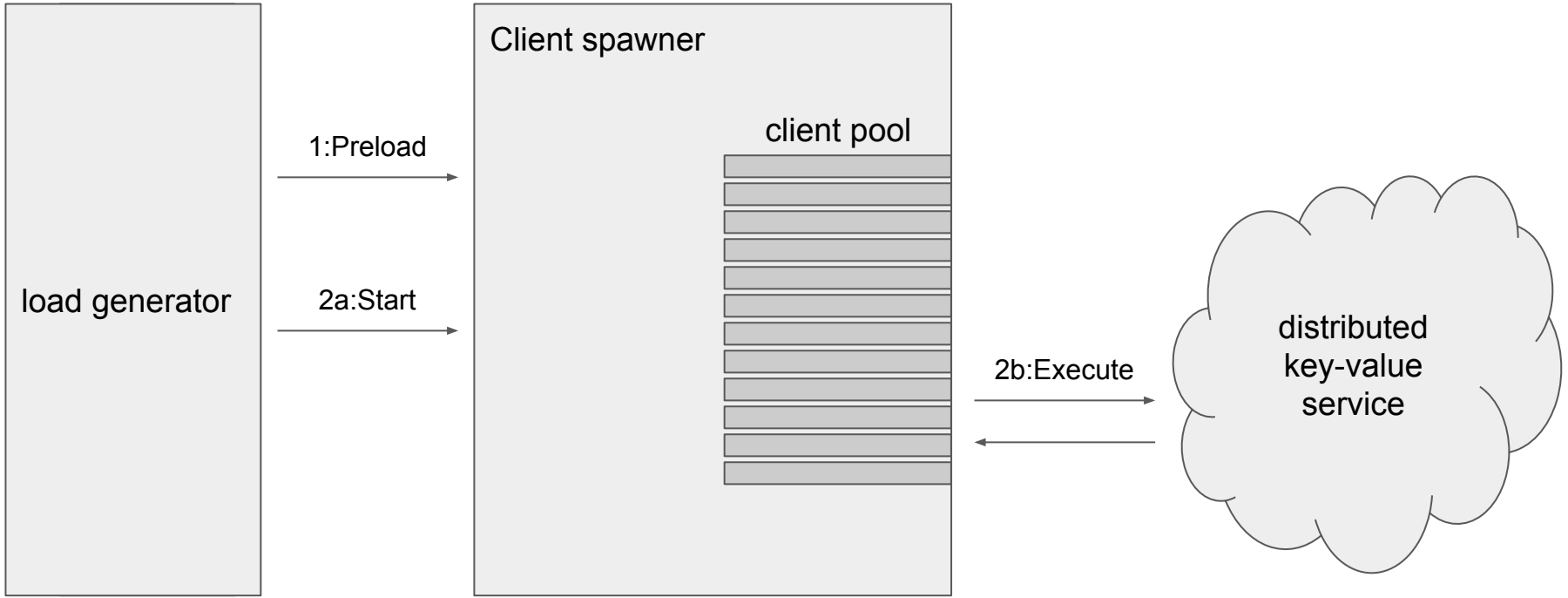
Load generator and client spawner



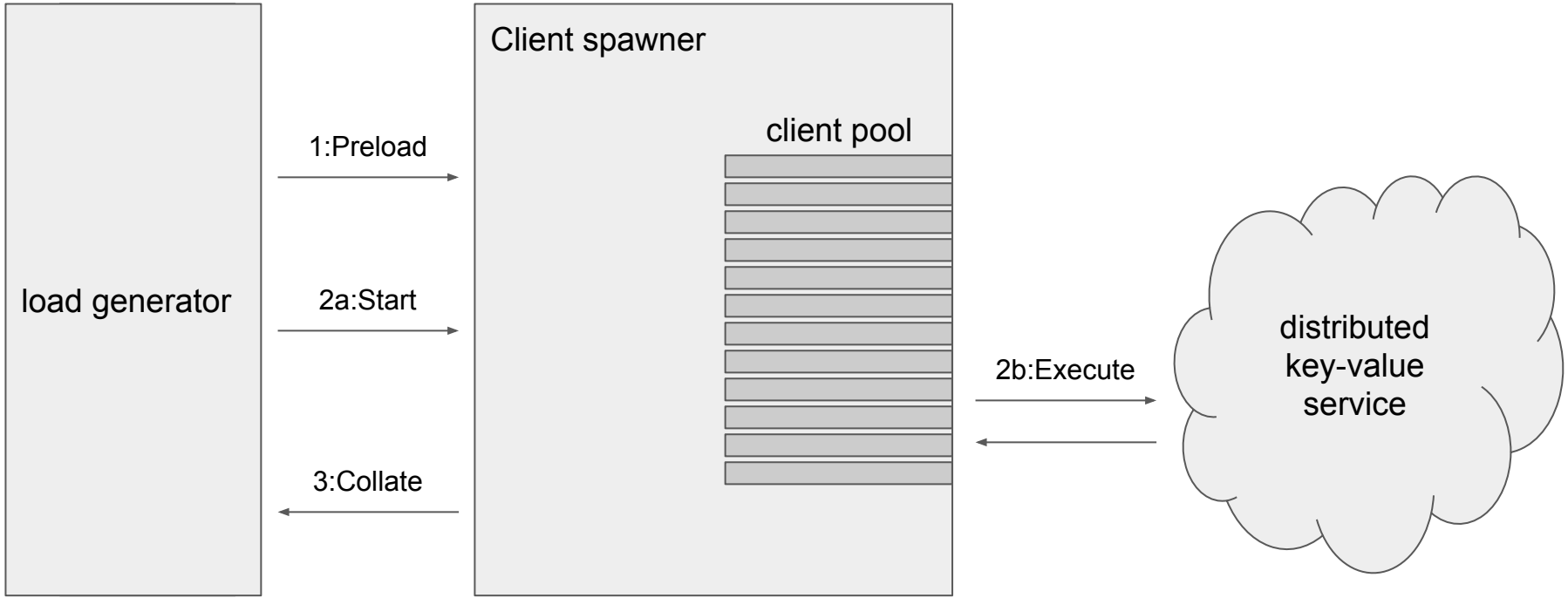
Load generator and client spawner



Load generator and client spawner

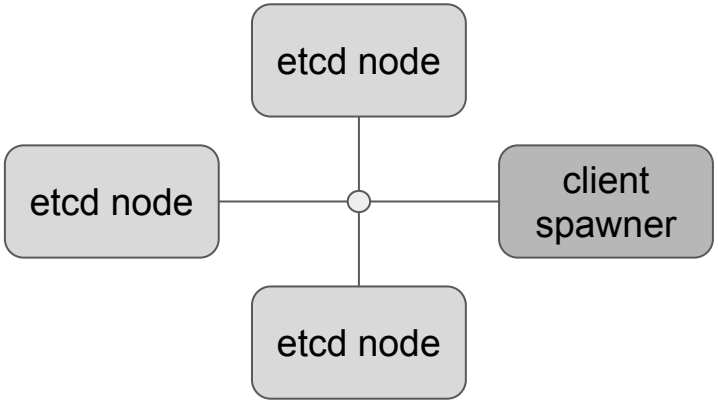


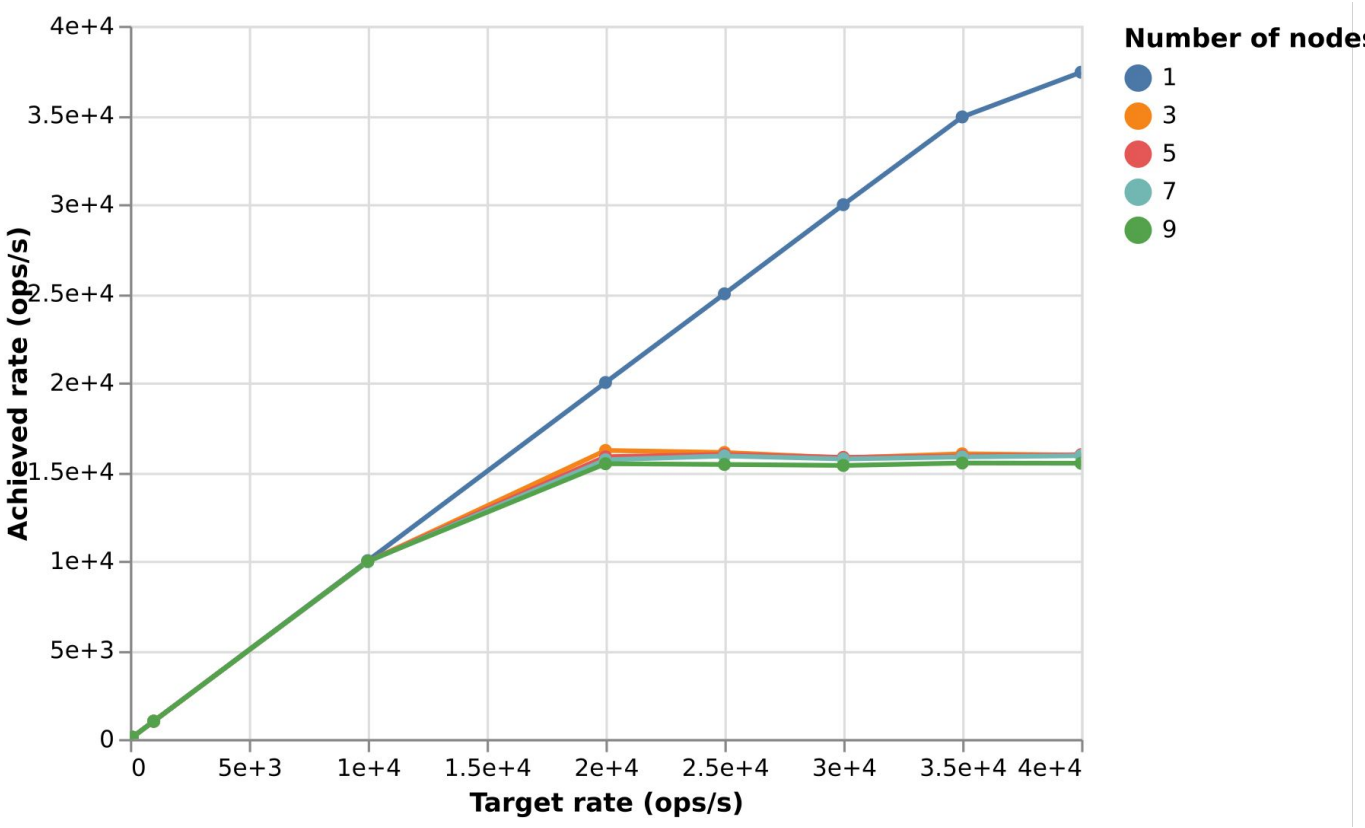
Load generator and client spawner



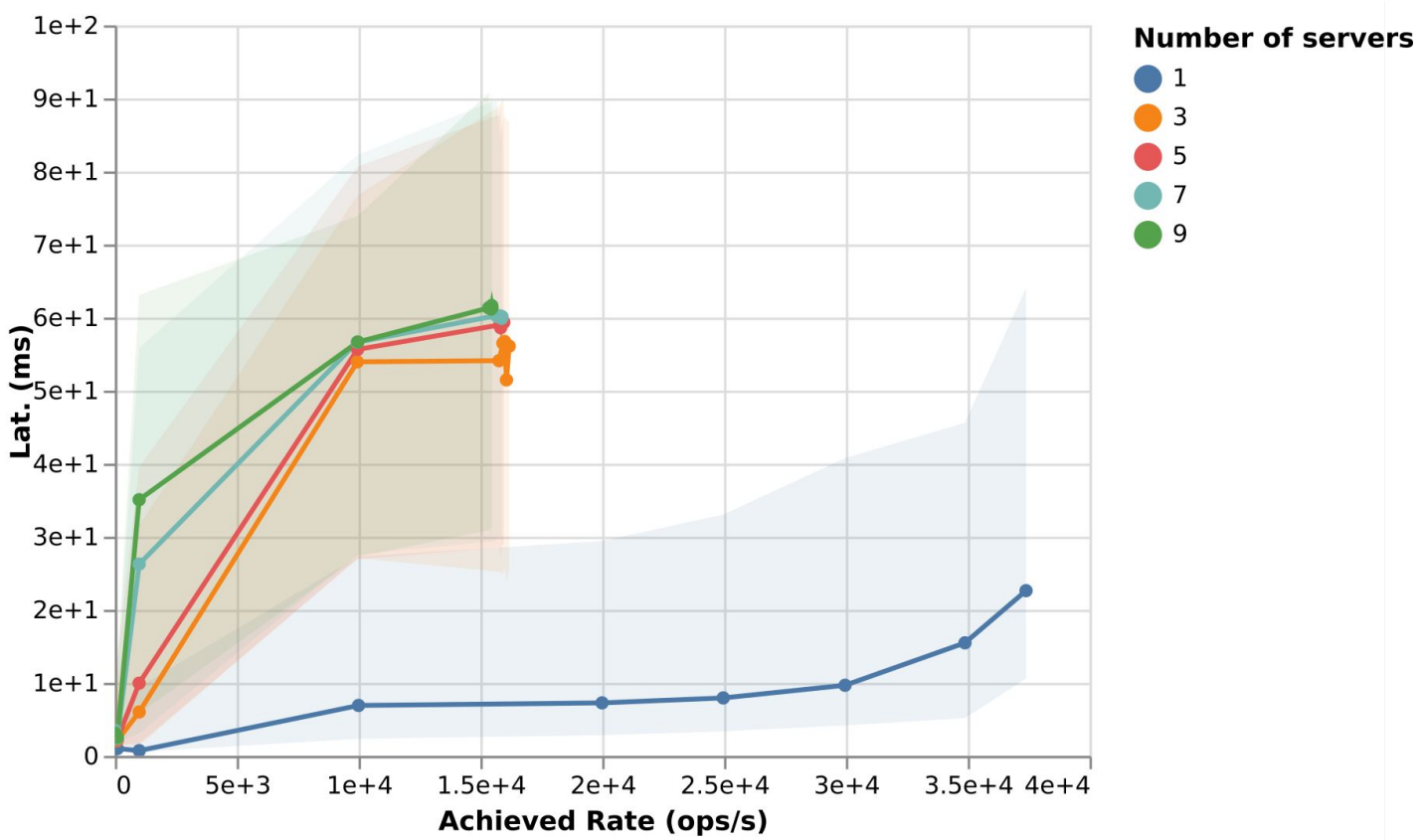
Validation test setup

- Tested using etcd v3.5.2
- N etcd nodes, one client spawner all connected to a central switch
- No limits on bandwidth/latency
- Just write requests
- Keys in the range 1-10 uniformly distributed
- 10 Byte keys and values
- Leader failure via SIGKILL
- 1000 closed loop clients in client pool

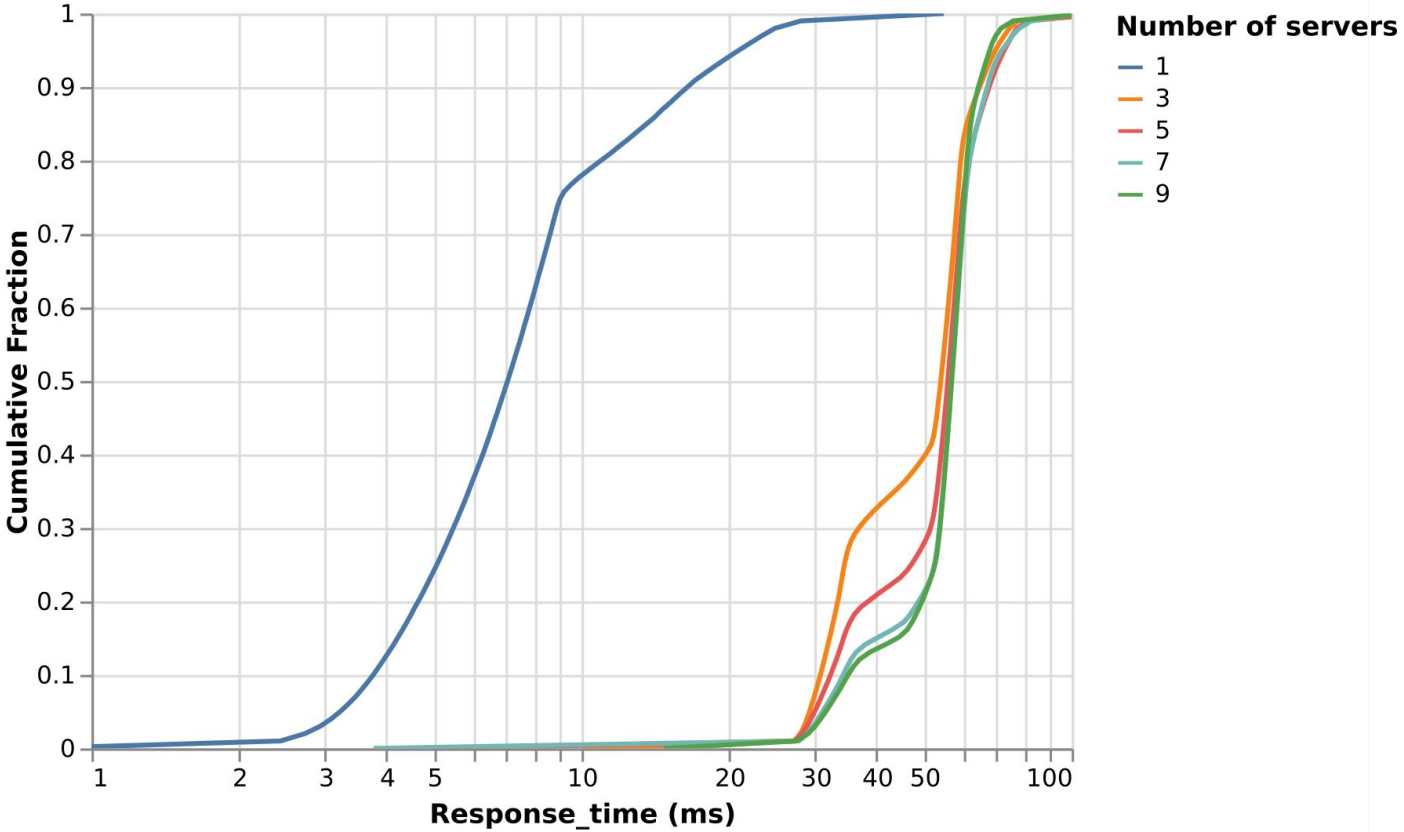




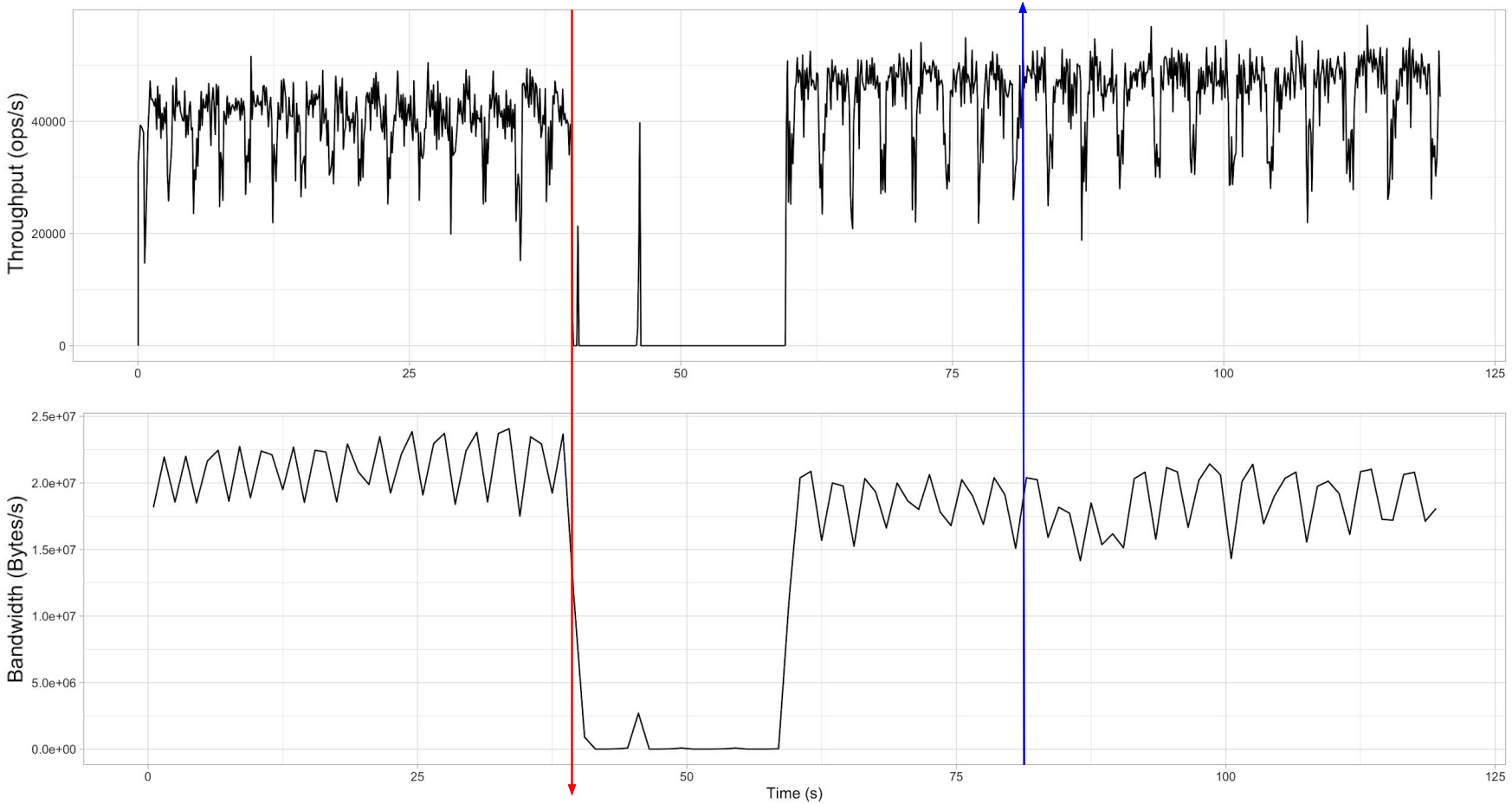
The target throughput of etcd against what it actually achieved, for [1,3,5,7,9] nodes.



etcd: Latency at achieved throughputs using [1,3,5,7,9] servers
5th, 50th and 95th percentiles shown.



etcd: Cumulative density plot of latency for [1,3,5,7,9] nodes at 5k ops/s



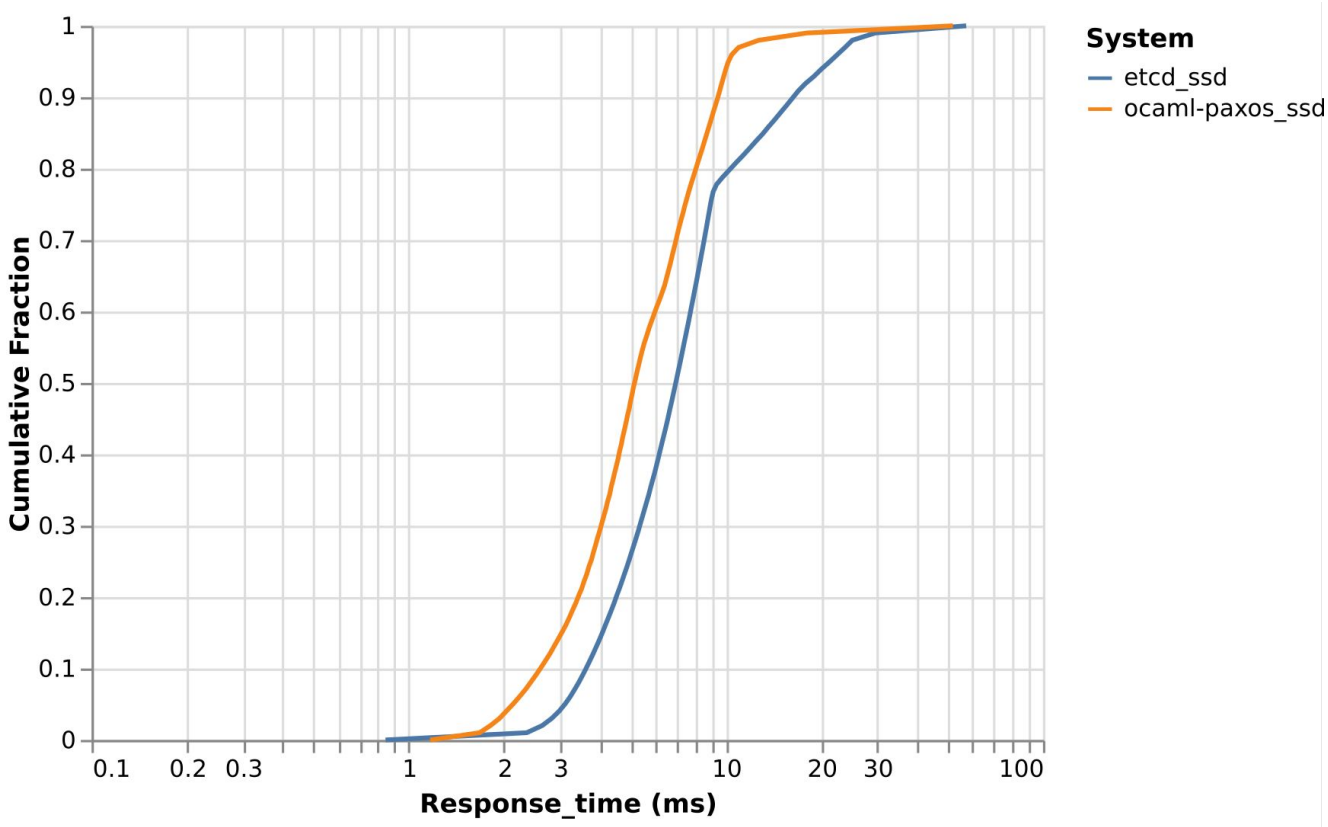
etcd: trace of leader failure for a 3 node configuration

Currently out of scope / Limitations

- Data dependencies between requests
- Failure trace artifacts from client pool approach
- Limited to tree topology

Upcoming work

- Extend to new systems:
  - Custom Multi-Paxos and Raft implementation
- Failure and recovery analysis (In general cases and specific case studies)
- New network topologies
- Heterogeneous host deployments
- Data dependencies
- Other workloads



Latency cdf of ocamlpaxos vs etcd v3.5.2 on an ssd at 8k ops/s

Thanks for listening!
Any questions?

https://github.com/Cjen1/Resolving-Consensus

Chris Jensen (cjj39@cam.ac.uk)
Daniel Sääw (dks28@cam.ac.uk)
Heidi Howard (hh360@cam.ac.uk)
Richard Mortier (richard.mortier@cl.cam.ac.uk)