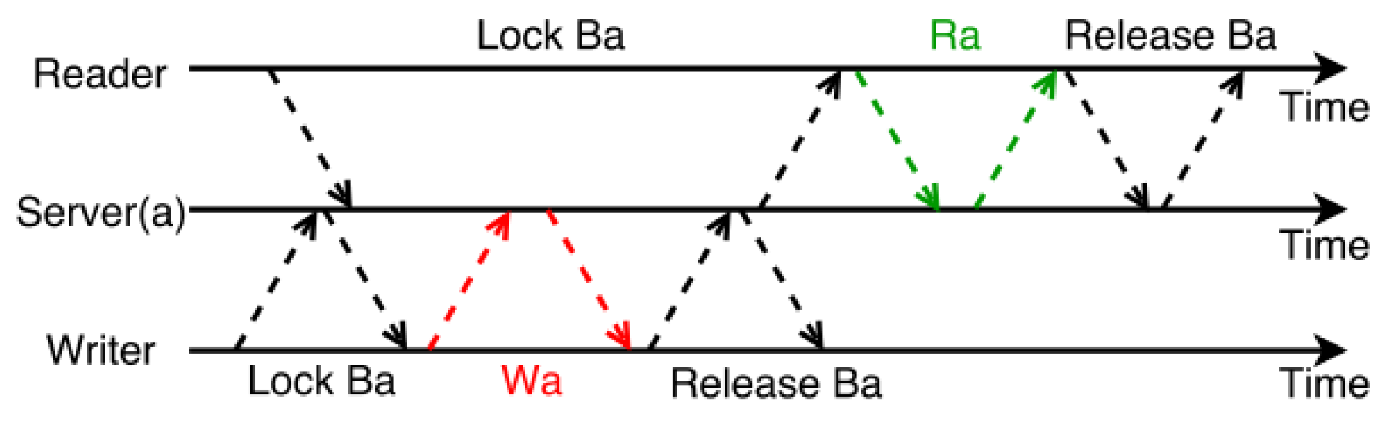
Efficient and Scalable Total-Order Message Scattering in Data Center Networks

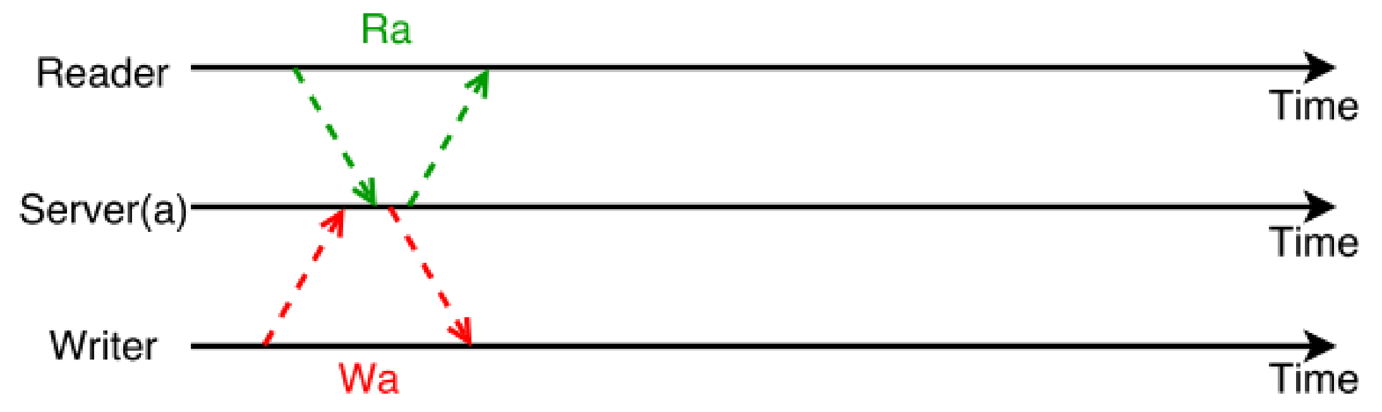
SOSP’17 SRC #34

Gefei Zuo, Bojie Li, Lintao Zhang

Motivation: Transactions



Lock-based concurrency control



Timestamp-based concurrency control

Total-Order Message Scattering

- Each *host* has a monotonic *timestamp* clock
- Each *event* is assigned a *timestamp*
- An event *scatters* messages to other hosts
- Each host *delivers* received messages in monotonic timestamp order

Existing work

- Centralized sequencers: not scalable
- Receiver-side synchronization: latency and network overhead

Design goals

- Scalable, fault tolerant, incremental deploy
- Low network and CPU overhead

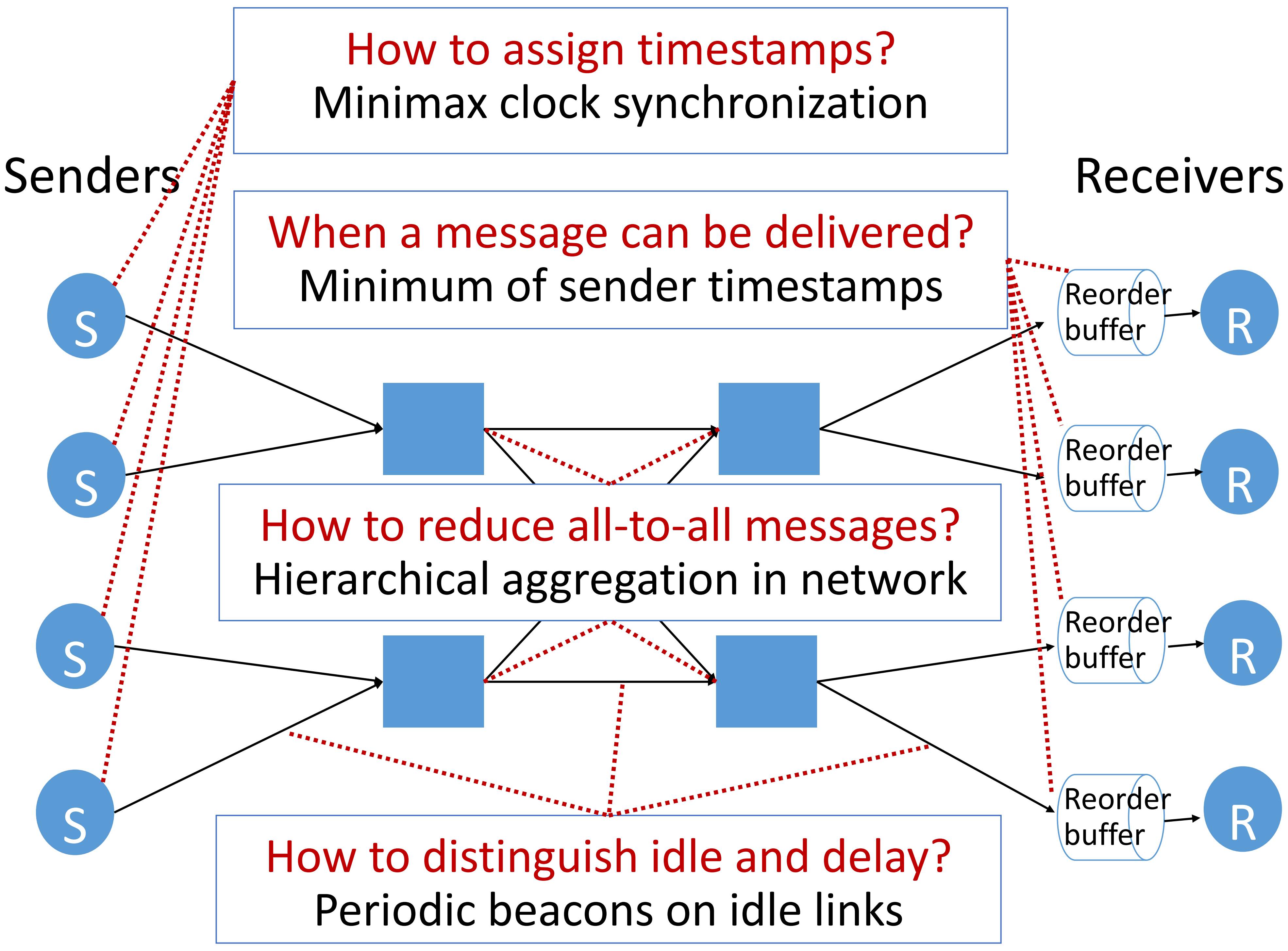
Where to ensure total ordering?

Network switches	End hosts
Wide visibility	Narrow visibility
Small buffer	Large memory
Low programmability	High programmability

Principle: Separate control plane from data plane

Control plane: *Aggregate* ordering information in network

Data plane: *Reorder buffer* in end-host receiver



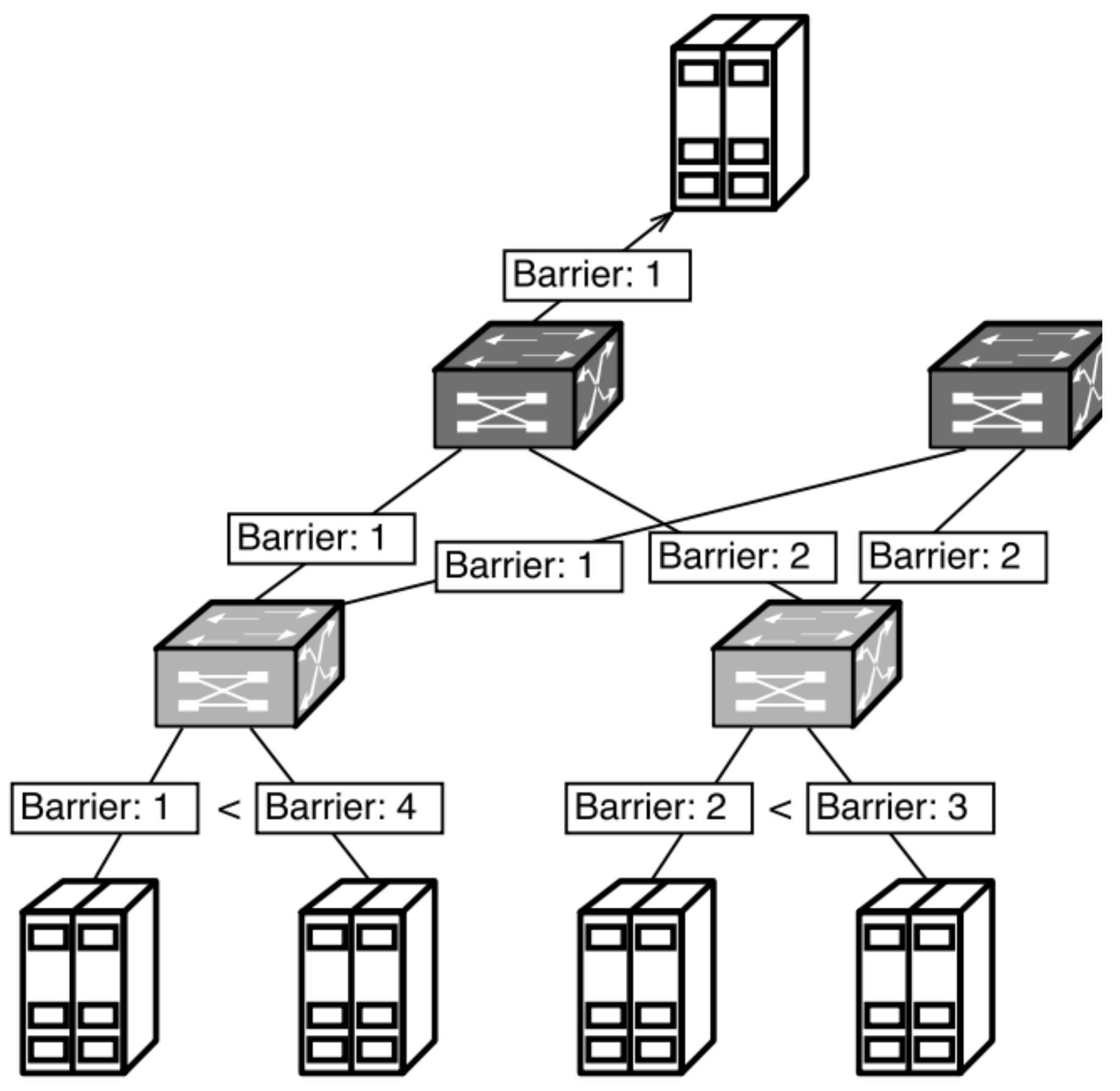
How to assign timestamps?
Minimax clock synchronization

When a message can be delivered?
Minimum of sender timestamps

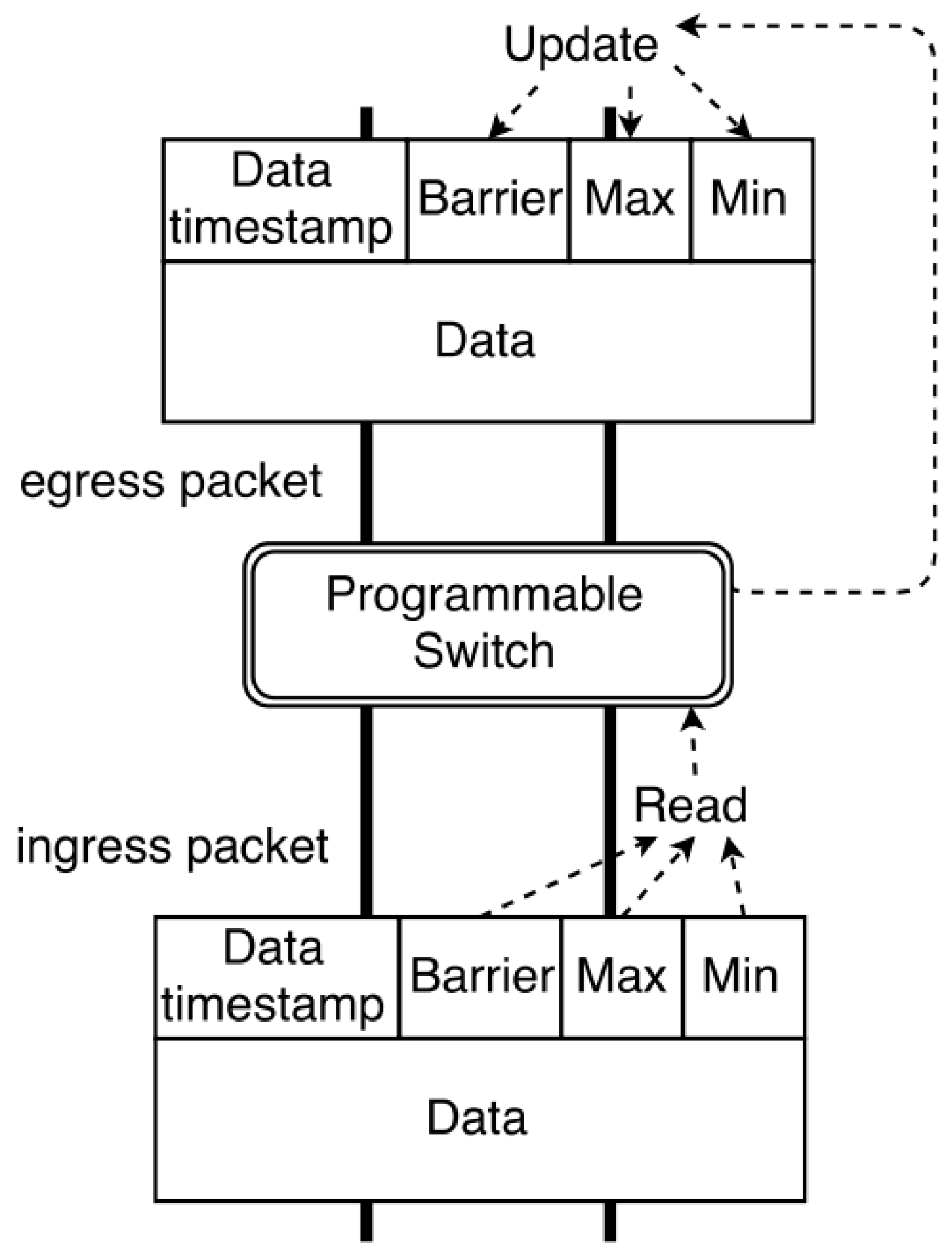
How to reduce all-to-all messages?
Hierarchical aggregation in network

How to distinguish idle and delay?
Periodic beacons on idle links

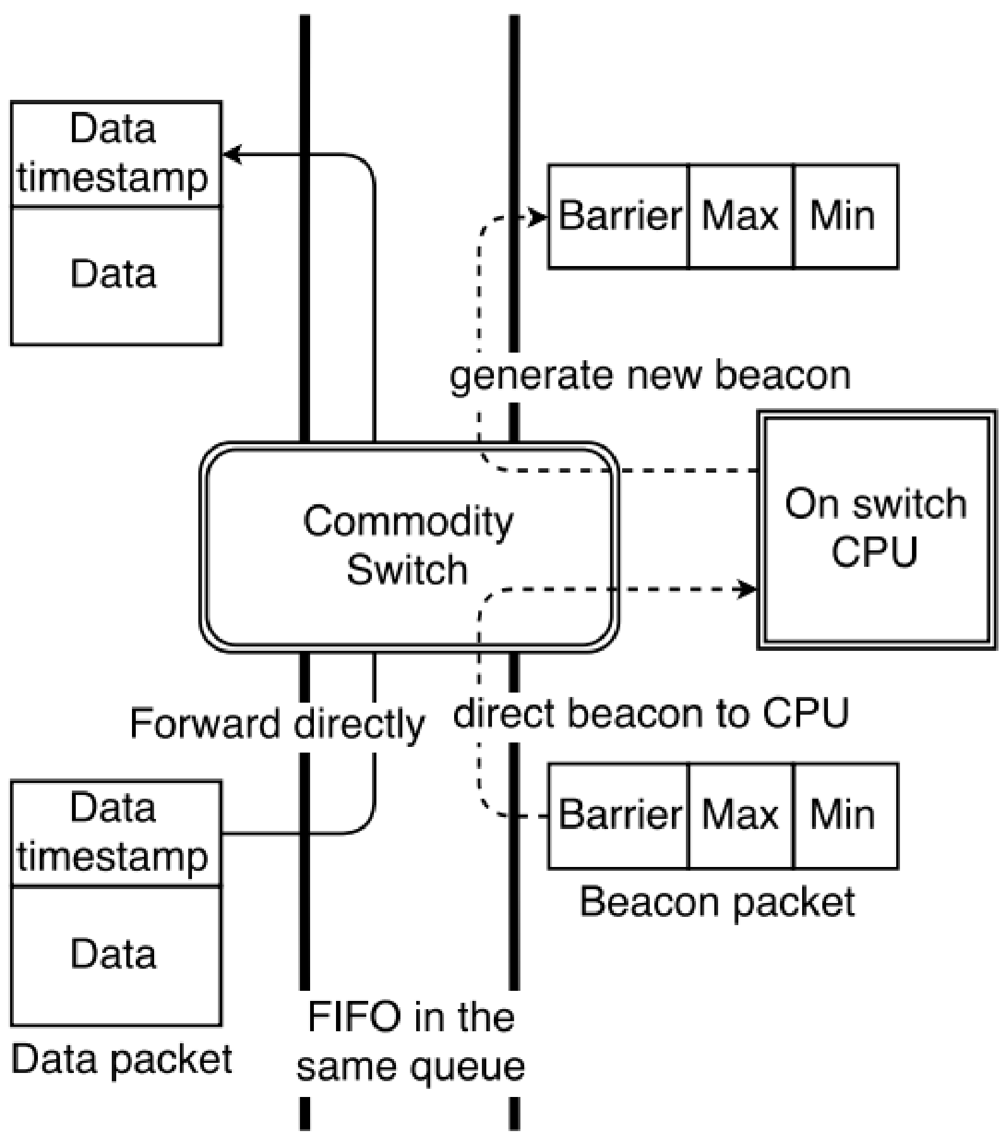
Aggregate timestamp barrier and sync in network switches



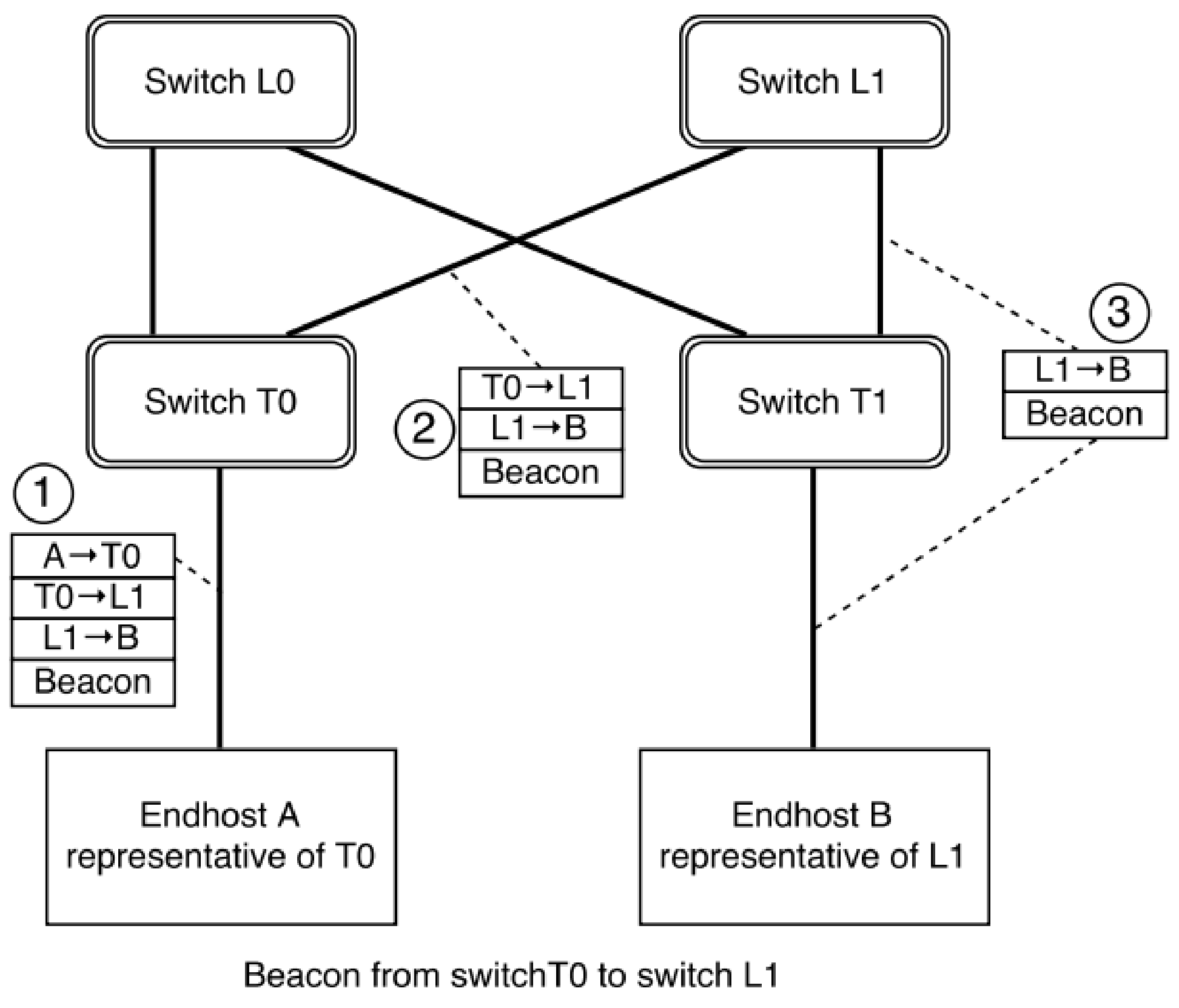
Hierarchical aggregation of timestamp barriers



Data-plane (P4) programmable switch



Commodity switch CPU


Switch L0
Switch L1
Switch T0
Switch T1
1
A→T0
T0→L1
L1→B
Beacon
2
T0→L1
L1→B
Beacon
3
L1→B
Beacon
Endhost A representative of T0
Endhost B representative of L1
Beacon from switchT0 to switch L1

End-host representative of switches

Evaluation with YCSB+T transactional key-value store

10M
YCSB+T transactions/s per server
50x lock-based

1
RTT per transaction
50% lock-based

10
microsecond reorder delay
+50% RTT

1%
Network traffic overhead

8
Per-port states in switch

2M
Reorder ops per CPU core

Low overhead

Scalability

- Simulation with 10K servers
- Both inside DC and across inter-DC WAN

Fault tolerance

- Event timestamps re-converge in 1 RTT.
- Incrementally deployable: add new host/link/switch in 1 RTT.