Fast Total Ordering for Modern Data Centers

Amy Babay, Yair Amir – {babay, yairamir}@cs.jhu.edu
Johns Hopkins University Distributed Systems and Networks Lab – www.dsn.jhu.edu

Background: Totally Ordered Multicast

- Agreed Delivery (Total Order): all group members deliver messages in the same order
- Safe Delivery (Stability): a group member only delivers a message after all other members have received it (and will deliver it, unless they crash)

Background: Token-based Protocols

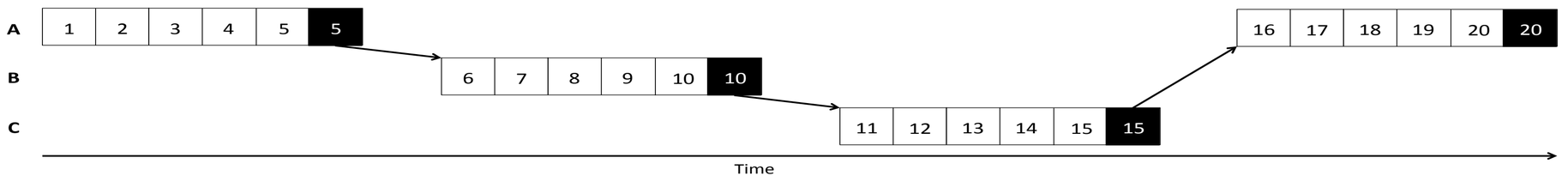
- Participants pass a token in a logical ring
- The token carries the sequence number of the last message sent
- A participant multicasts messages while it holds the token, then updates the token and passes it on

Why Another Protocol?

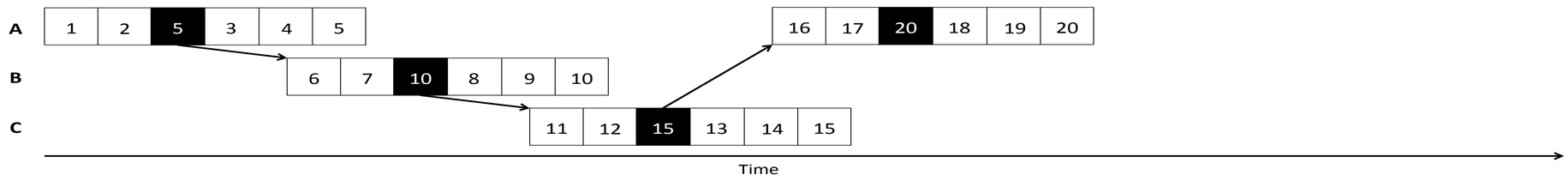
- Network trade-offs changed
  - Throughput improvements outpaced latency improvements
  - Buffering in switches
- Existing token-based protocols don't fully utilize modern networks

Accelerated Ring Protocol

- Key difference from previous protocols: participants can pass the token before they finish multicasting
  - Correct semantics maintained through careful protocol modifications



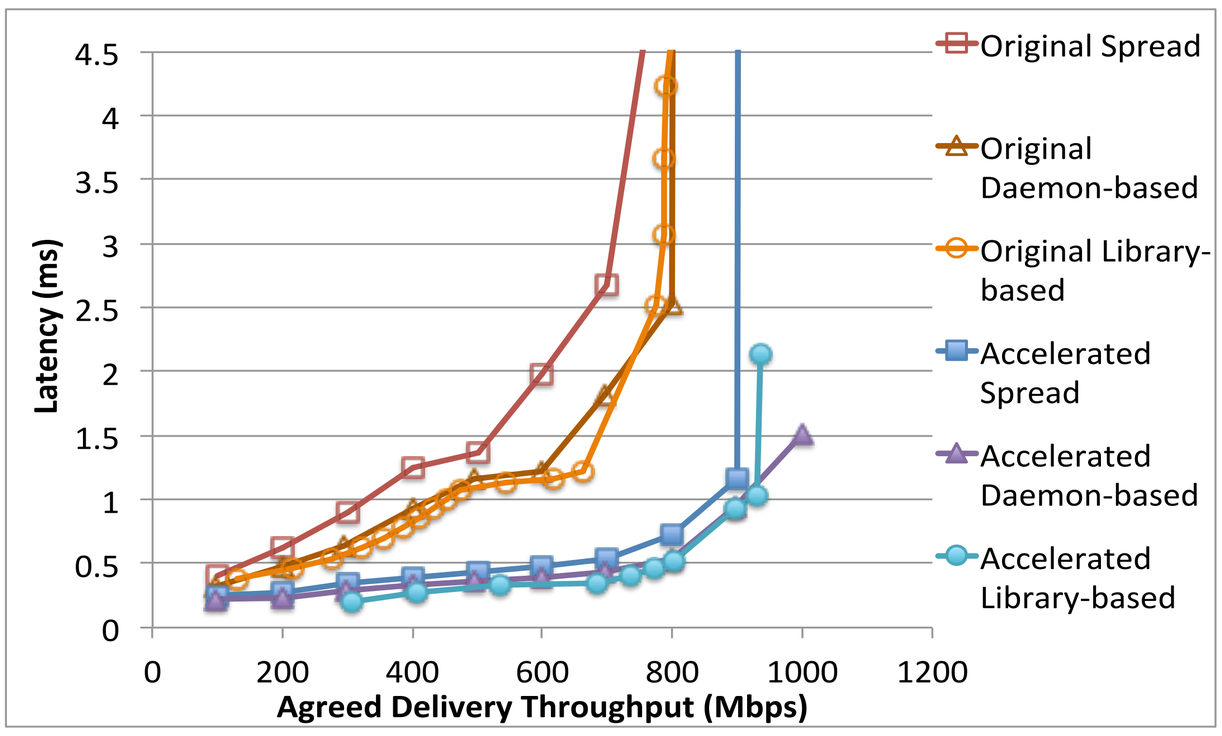
(a) Original Ring Protocol



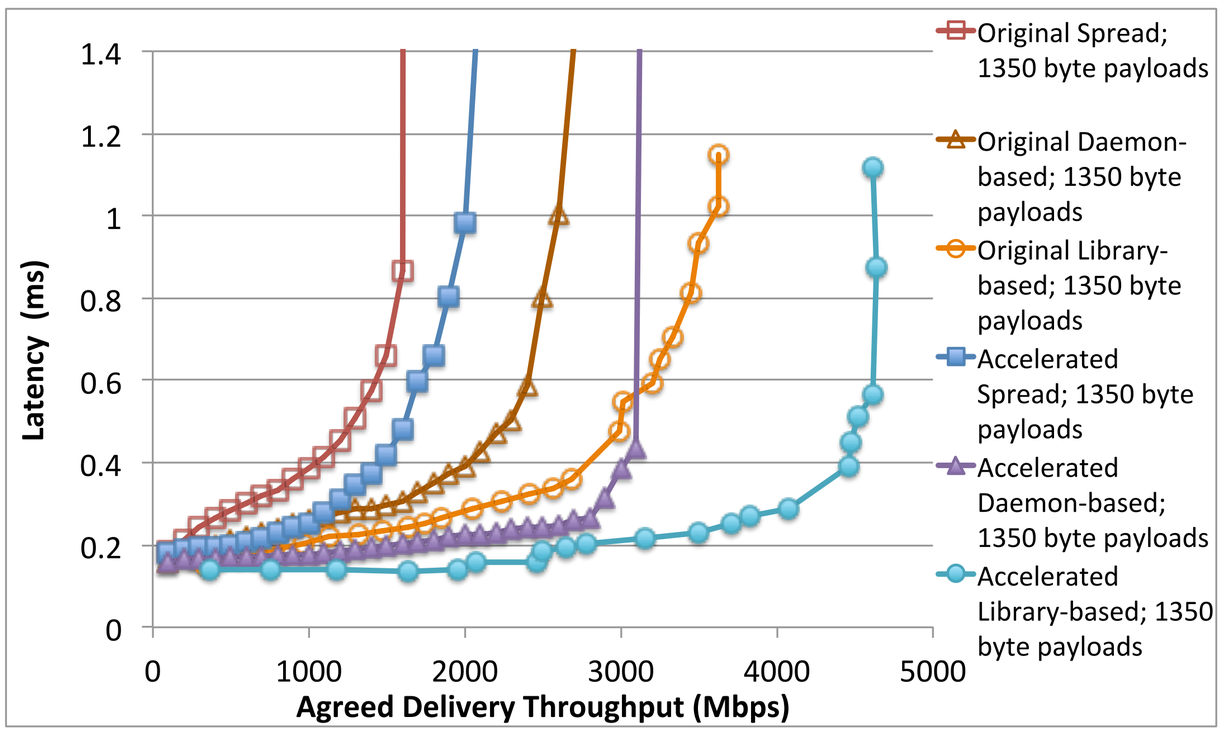
(b) Accelerated Ring Protocol

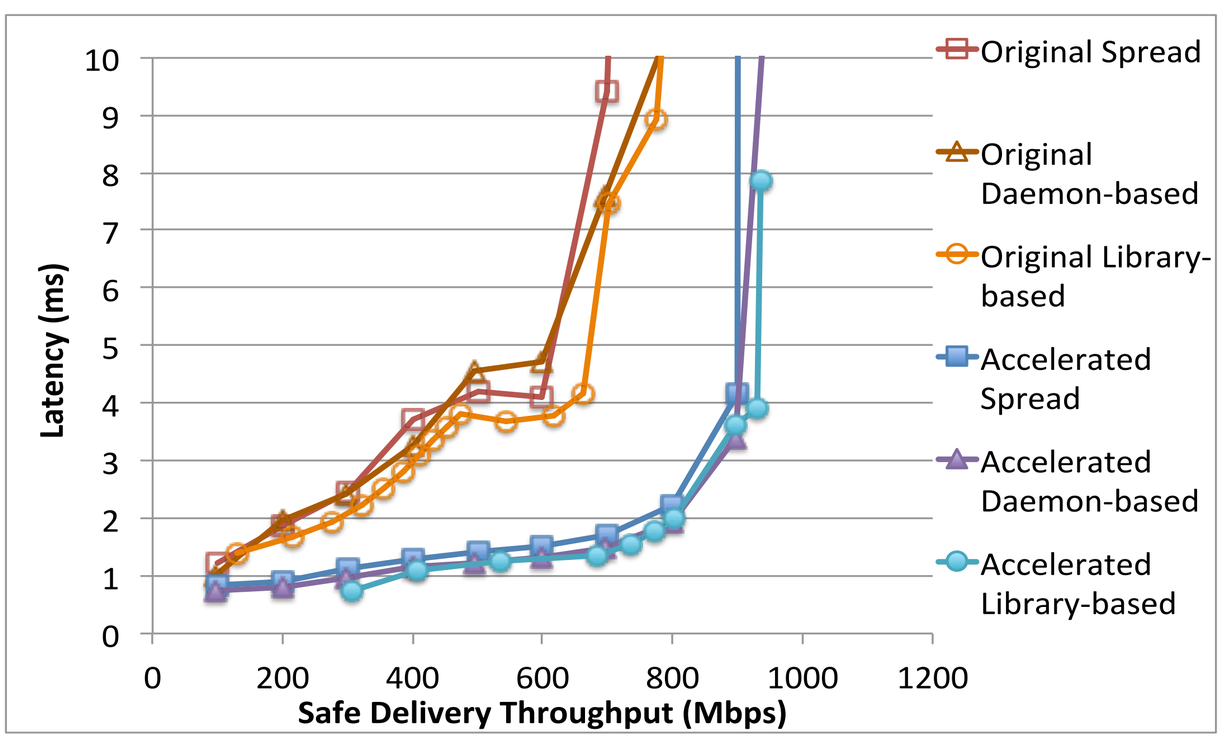
A simple, powerful improvement that allows a useful protocol to scale three orders of magnitude over 20 years. (www.spread.org)

1-Gigabit Network Results

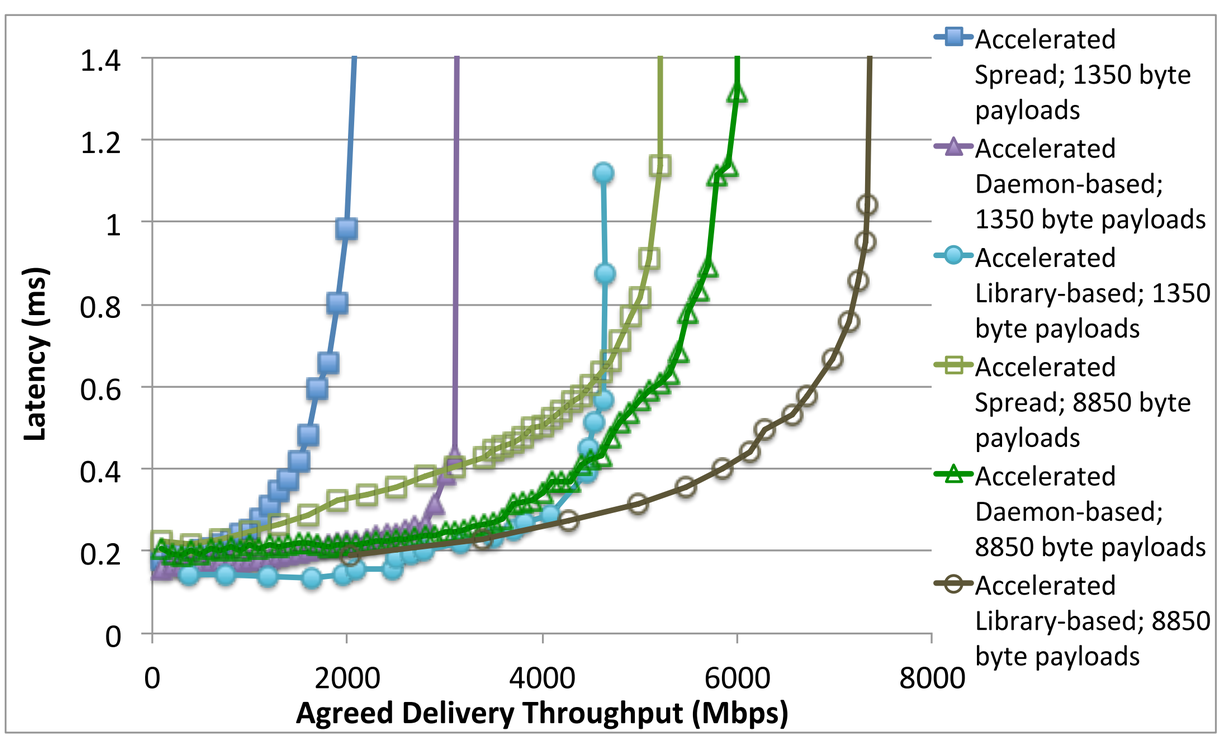


10-Gigabit Network Results





Can simultaneously improve throughput by 45% and latency by 30%; reaches network saturation



Can simultaneously improve throughput by 25-40% and latency by 25-40%; reaches 6 Gbps with 8850-byte datagrams (daemon-based)