

Atomic Multicast: do Skeen ao Pacheco

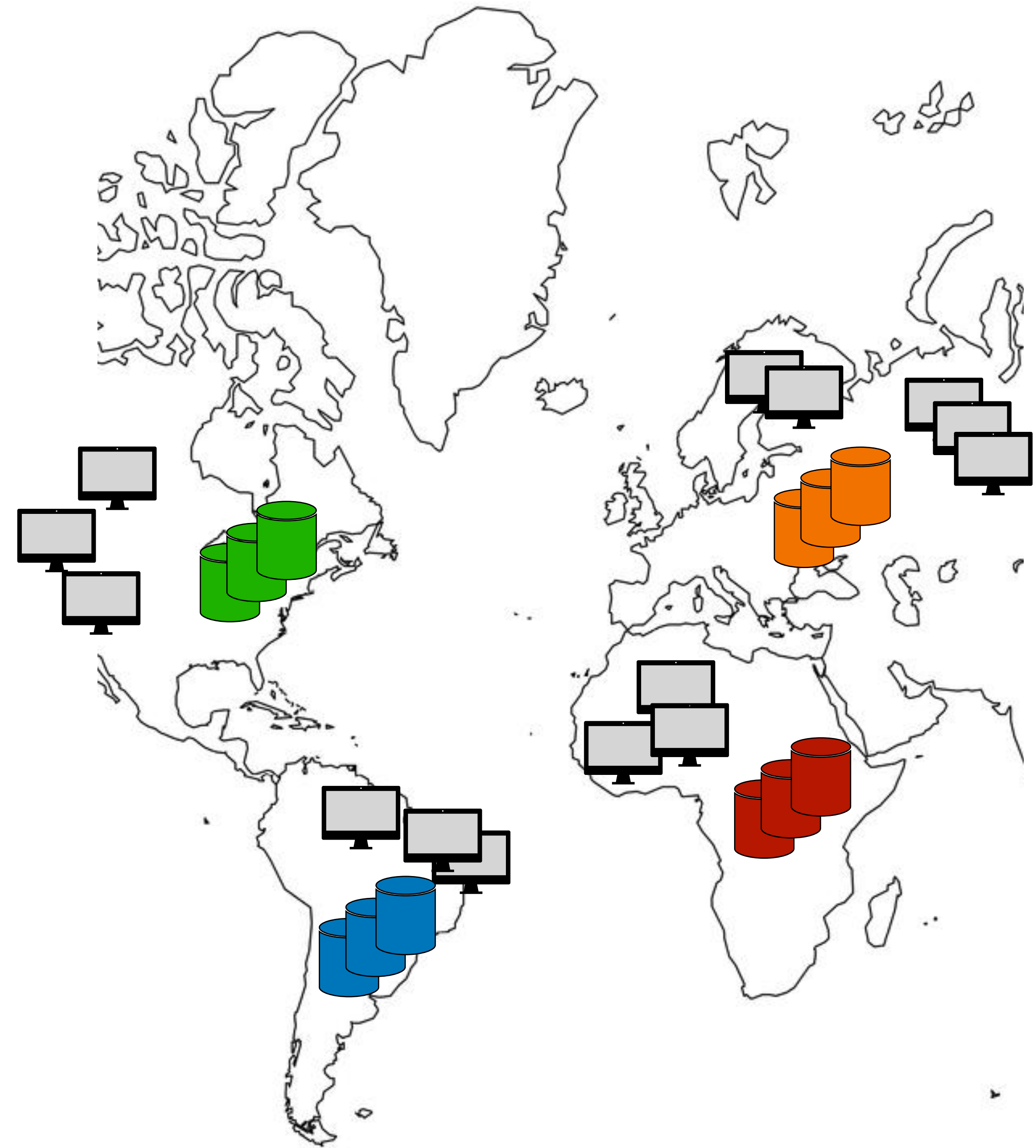
**Paulo Coelho
Universidade Federal de Uberlândia**

Abril/2025 - Porto Alegre/RS

Motivation

Global distributed systems

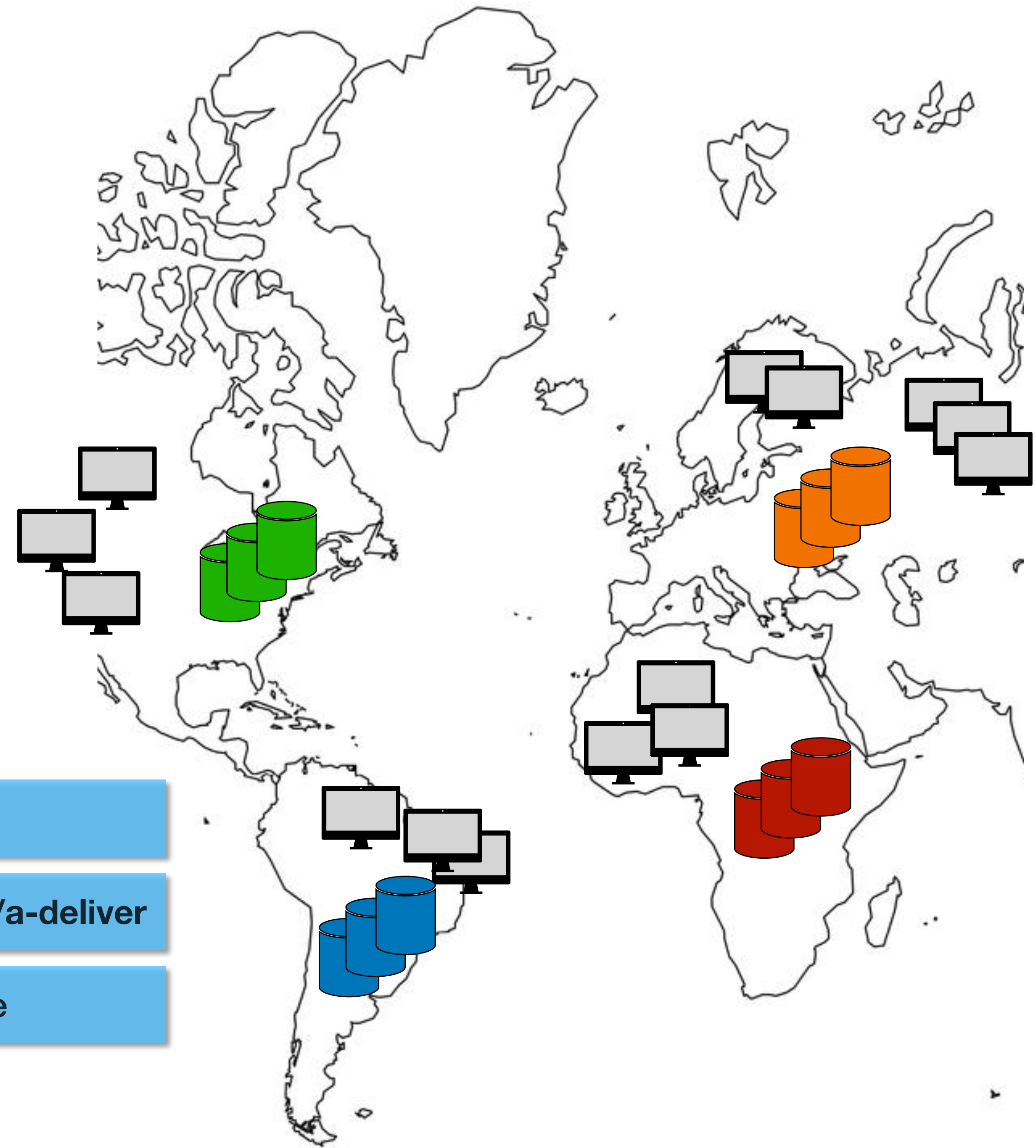
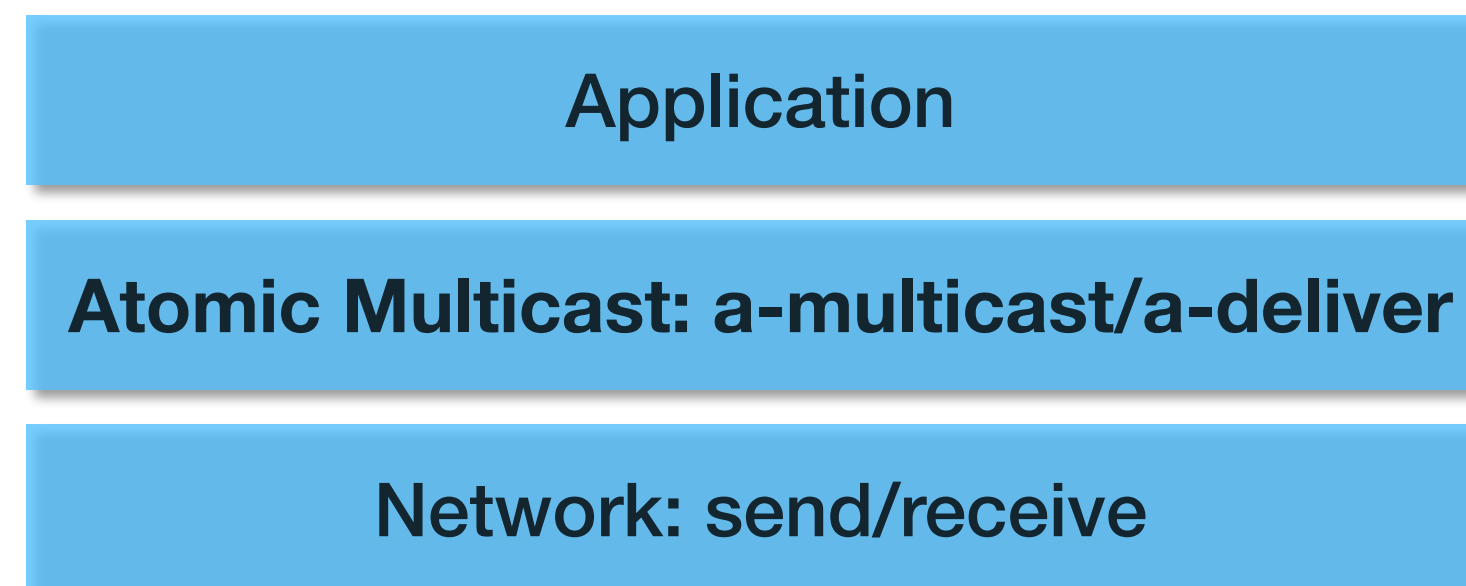
- **Replication** for fault-tolerance
- **Data partitioning** for scalability and locality
- Challenges:
 - Scale with the number of partitions
 - Good performance (latency) with global distribution



Motivation

Atomic Multicast

- Communication abstraction for partitioned systems
 - Messages addressed to groups of replicas (partitions)
- Strong ordering guarantees, simple API
 - `a-multicast(m) :`
multicasts `m`
to groups in `m.dst`
 - `a-deliver(m) :`
delivers ordered `m`
to groups in `m.dst`

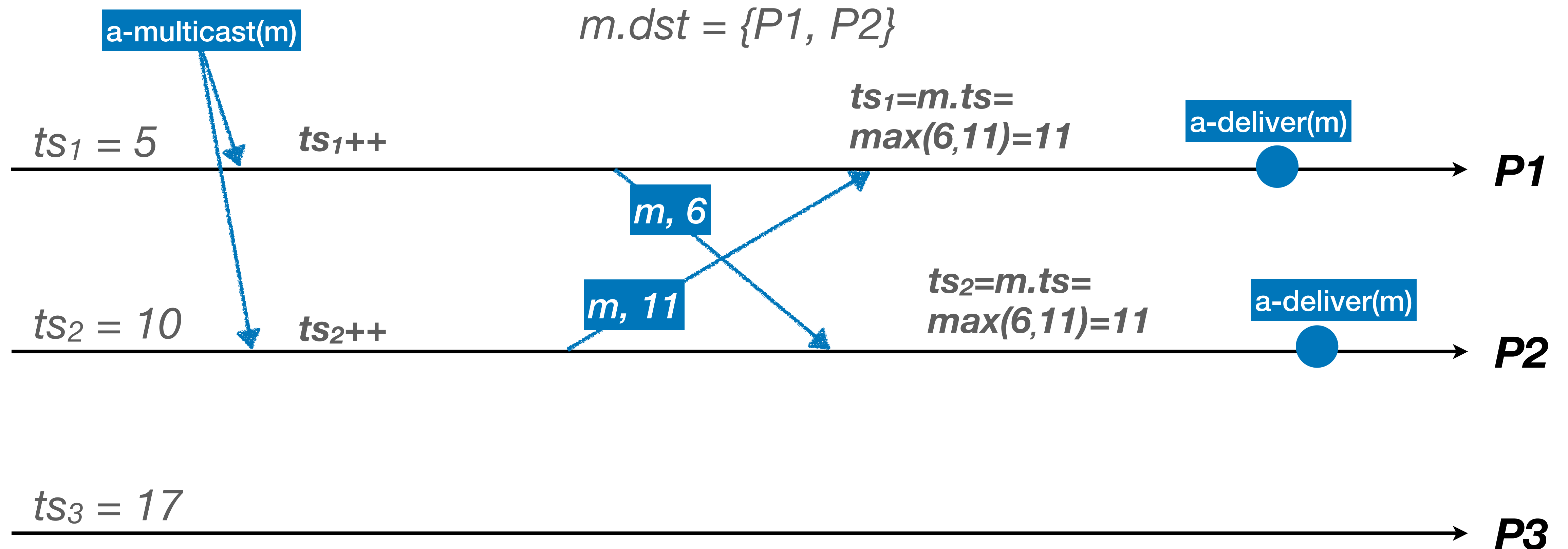


System Model and Definitions

- **N** server processes (replicas) organized in **m** disjoint groups
- Partially synchronous system
- Leader election oracle for each group **g** (Ω_g)
- **Atomic Multicast** properties:
 - Global total order
 - Prefix order
 - Genuineness

From Skeen to Pacheco

Skeen's Algorithm (1987)

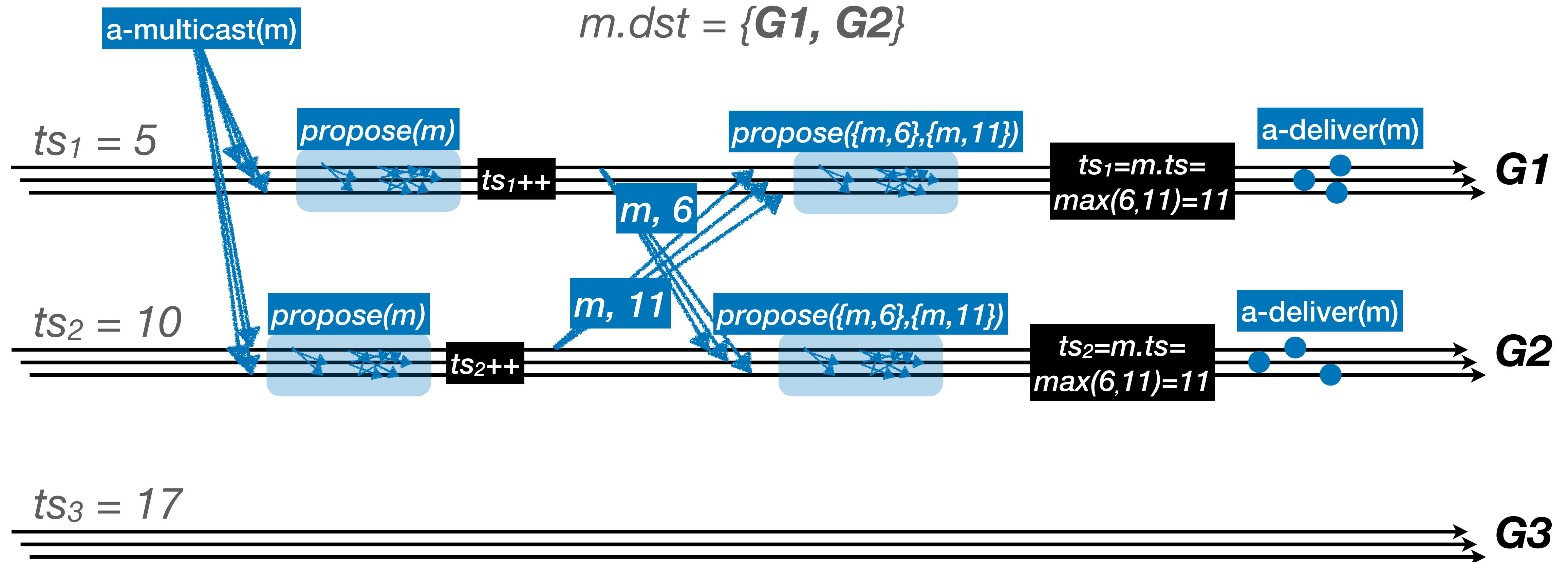


Only 2 communication steps! :-)

NO FAULT TOLERANCE! :-)

From Skeen to Pacheco

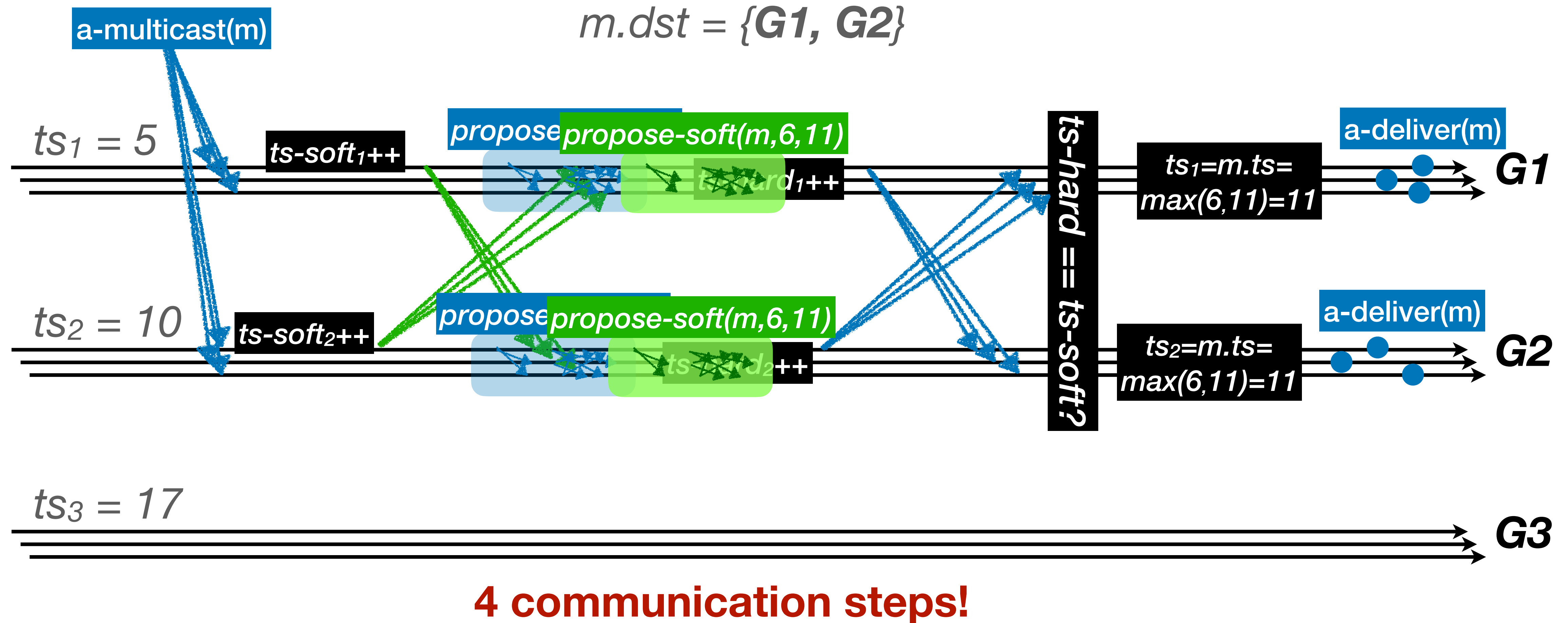
Skeen's Algorithm - Fault-tolerant



6 communication steps! :-)

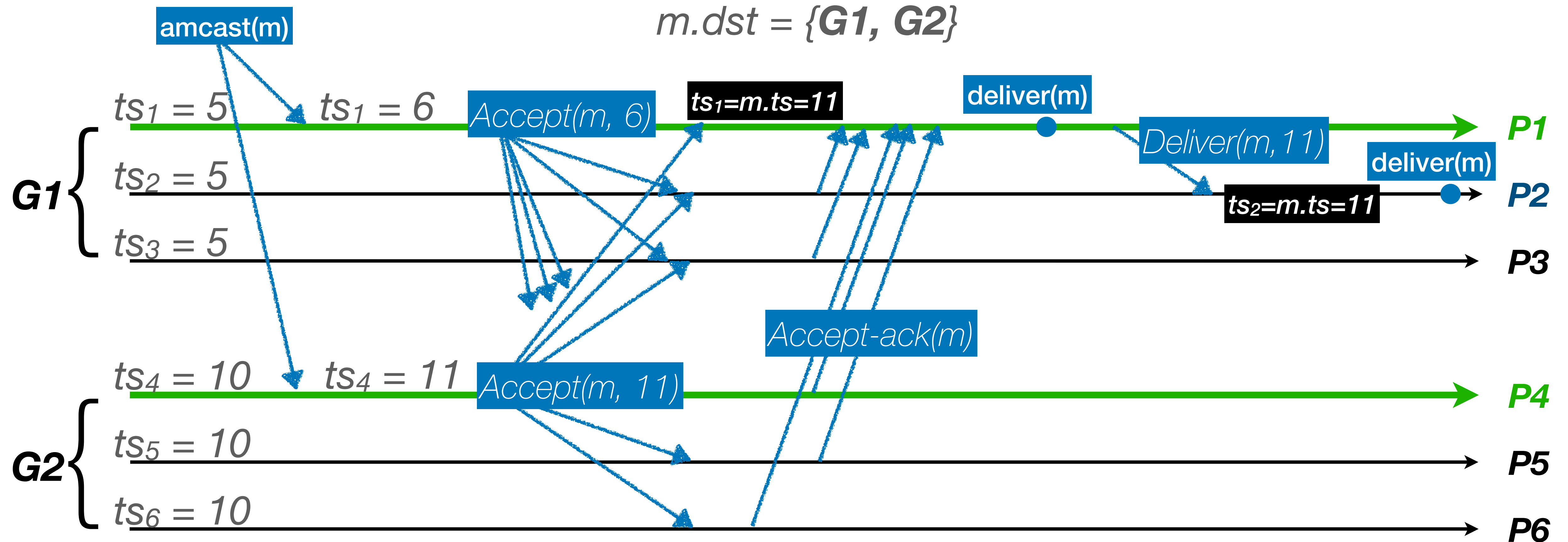
From Skeen to Pacheco

Coelho's Algorithm - FastCast (2017)



From Skeen to Pacheco

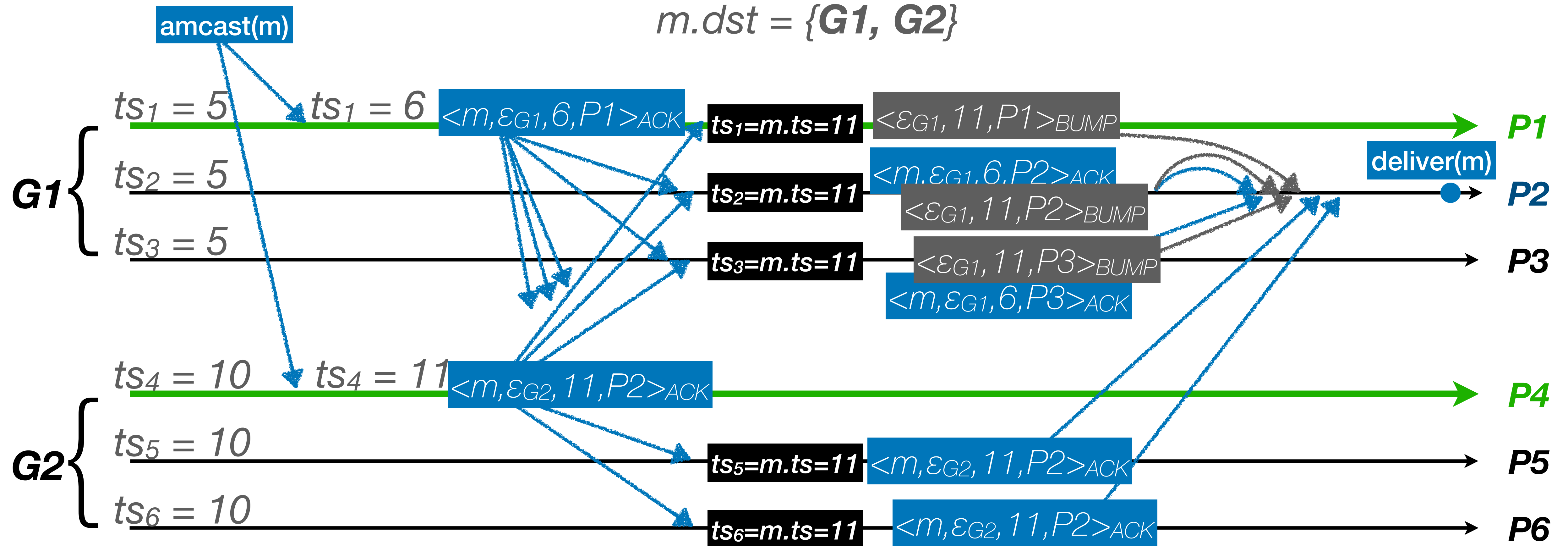
Gotsman's algorithm - WhiteBox (2019): Delivery in processes **P1** and **P2**



Leader: 3 communication steps!
Followers: still 4 communication steps!

From Skeen to Pacheco

Pacheco's algorithm - PrimCast (2023): Delivery in process **P2**



3 communication steps! :-)

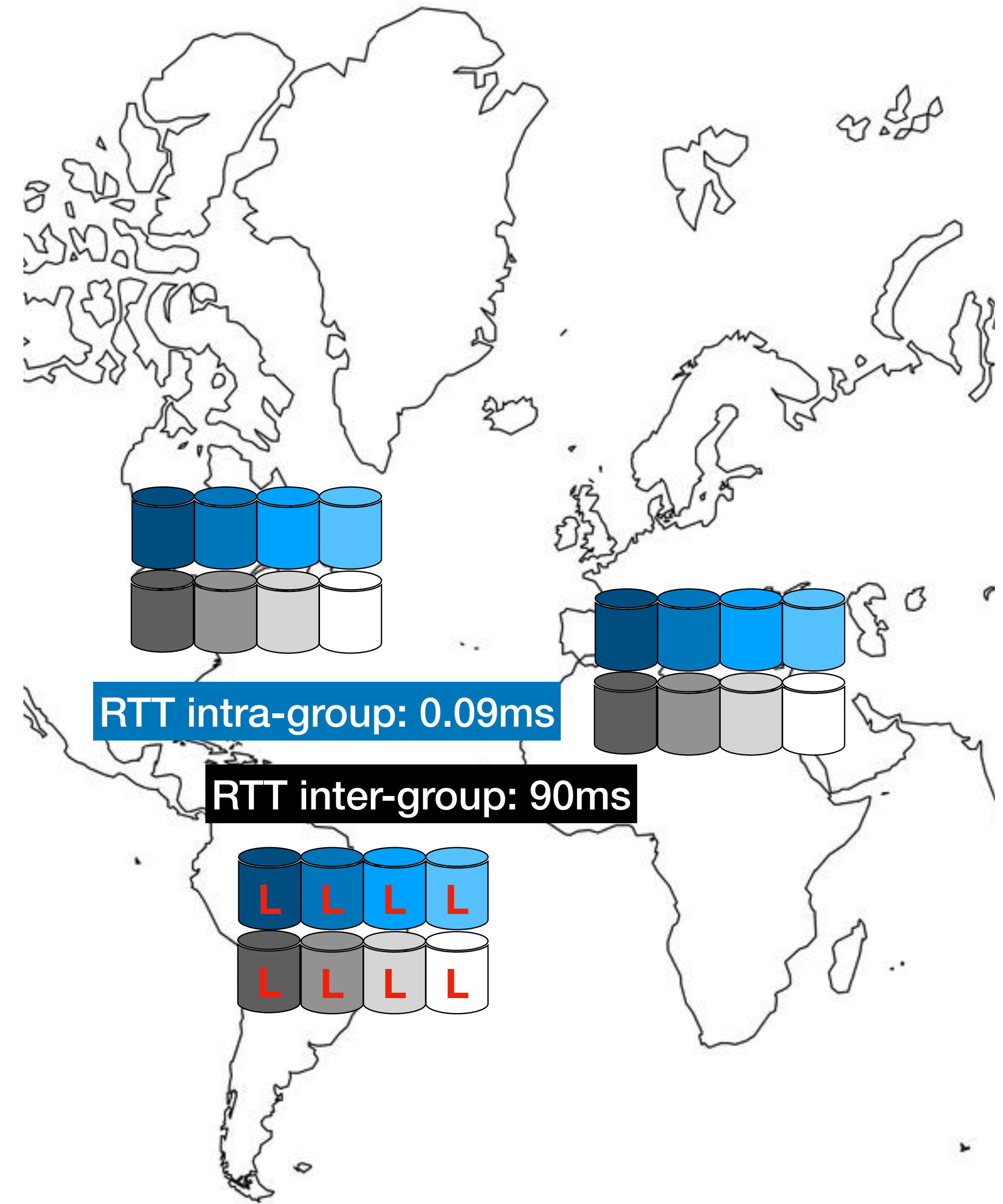
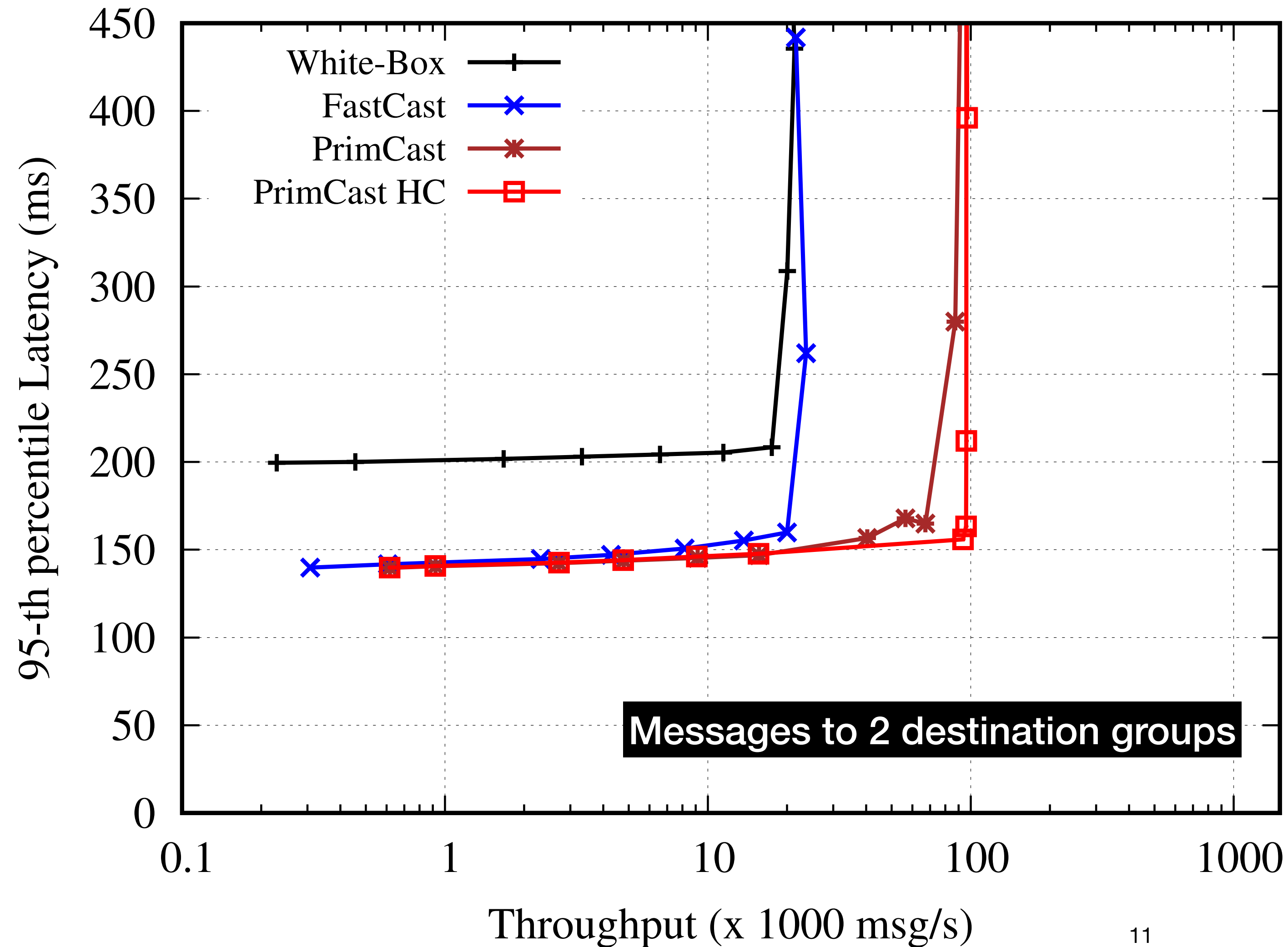
Experimental Evaluation

Setup

- 8 groups of 3 processes
- Comparison of PrimCast, FastCast and WhiteBox
- Scenarios:
 - Scenario 1: WAN - colocated leaders: 3 geographic regions
 - Scenario 2: WAN - distributed leaders: 8 geographic regions
- 1 client / replica with increasing number of outstanding messages

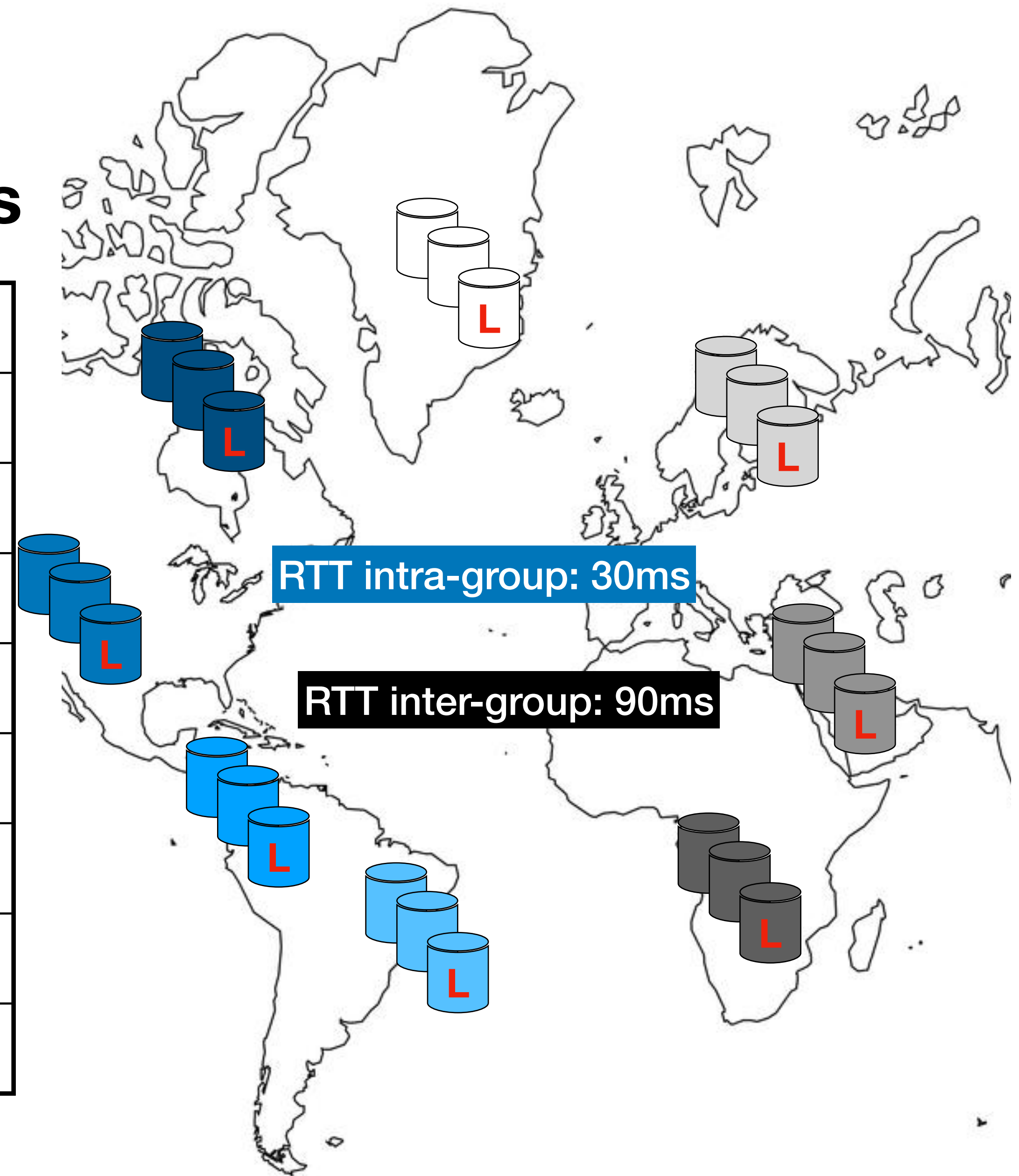
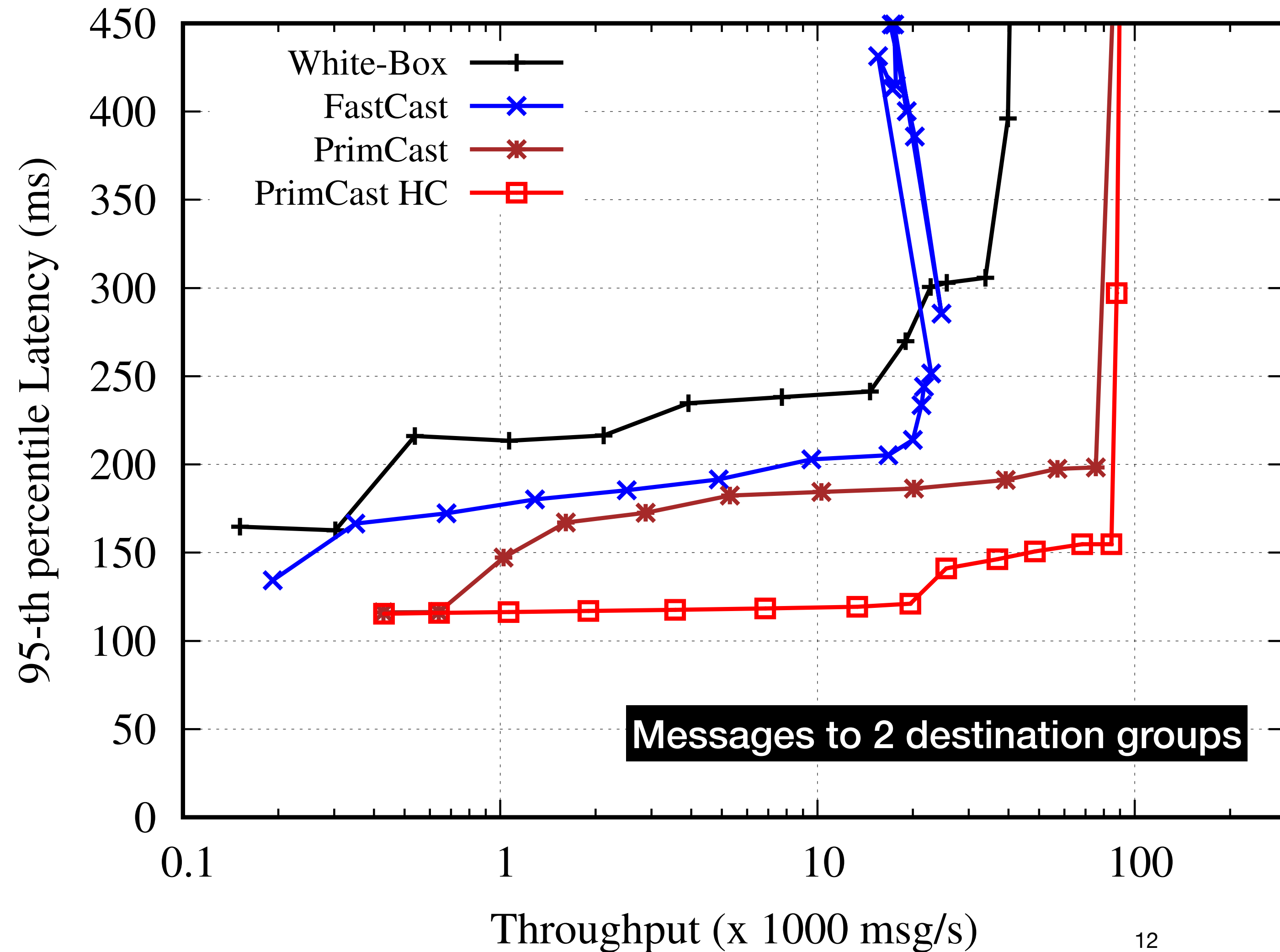
Experimental Evaluation

Scenario 1: WAN - colocated leaders



Experimental Evaluation

Scenario 2: WAN - distributed leaders



Conclusion

- Reducing latency (communication steps) pays off!
- PrimCast is the first genuine atomic multicast protocol to deliver messages in 3 communication delays in every replica
- Hybrid clocks can effectively reduce latency

Thank you!

paulocoelho@ufu.br