The End of Slow Networks: It’s Time for a Redesign

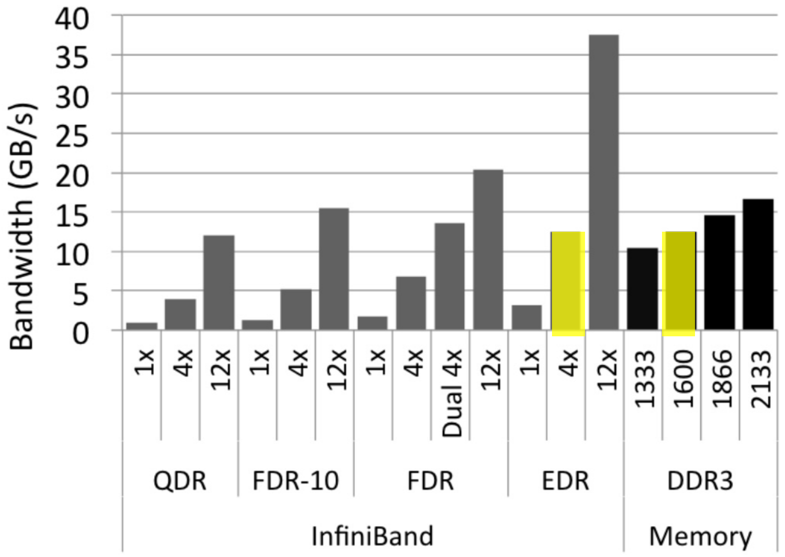
Outline

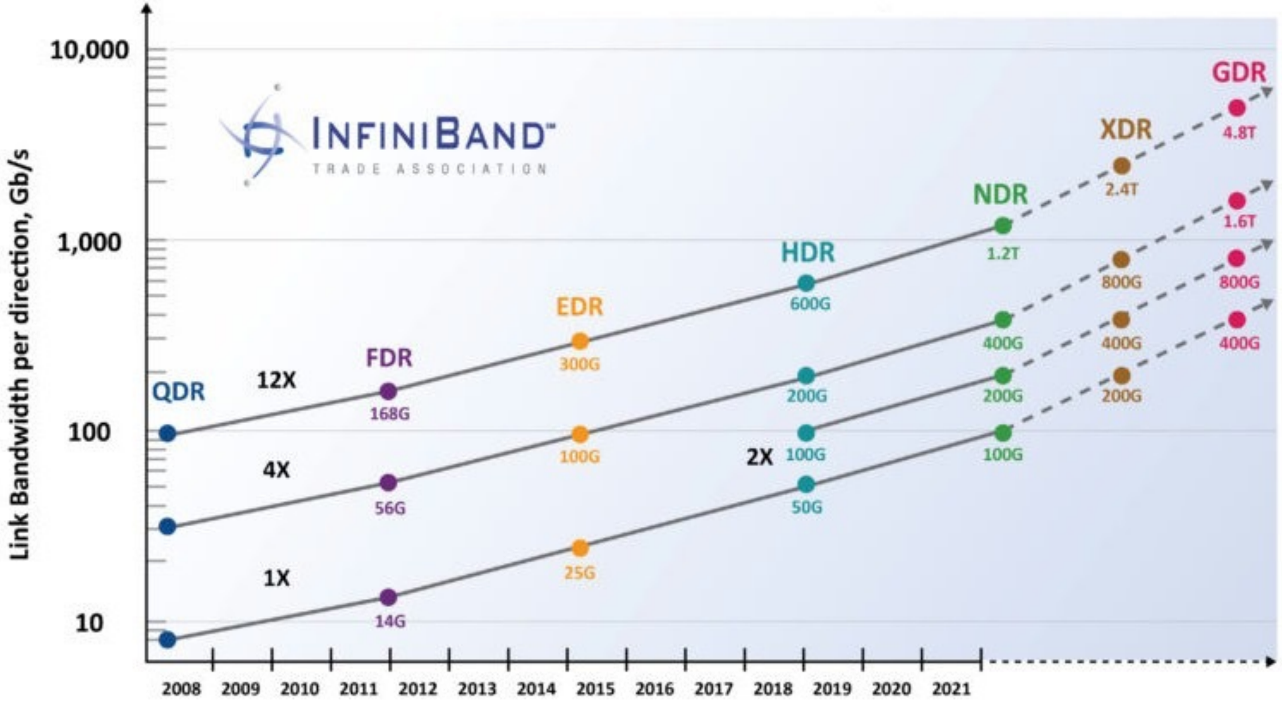
- Remote Direct Memory Access (RDMA)
- InfiniBand
- Performance benchmarks
- New architectures
- Distributed OLTP with RDMA
    - 2PL does scale
- Distributed OLAP with RDMA
    - Joins, aggregations, NAM arch.

- “we argue that it is time for a complete re-design of traditional distributed DBMS architectures to fully leverage the next generation of network technologies.”
- Distributed transactions? More from Geoffrey

InfiniBand

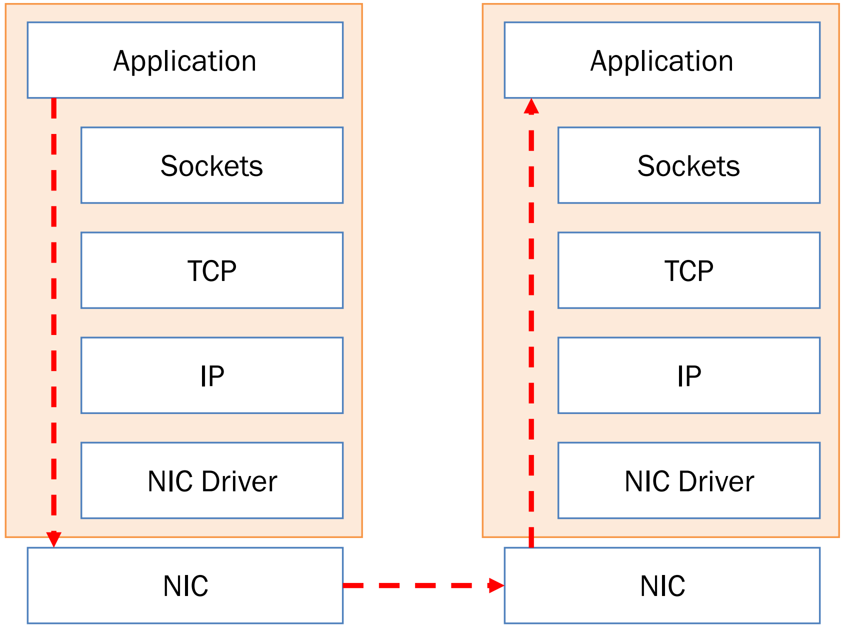
- High-performance network supporting both IP (IPoIB) and RDMA
- Historically very expensive
- Bandwidth of FDR 4x is about the same as a memory channel





Remote Direct Memory Access (RDMA)

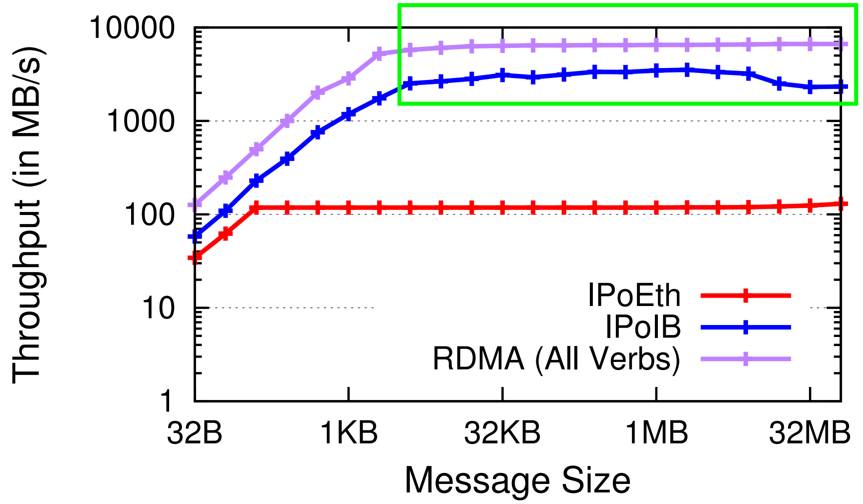
- This paper: RDMA over InfiniBand
  - as opposed to RDMA over Converged Ethernet
- Bypass Kernel TCP/UDP stack
- Supported “verbs”: one-sized atomics, read, write; two-sided send, recv

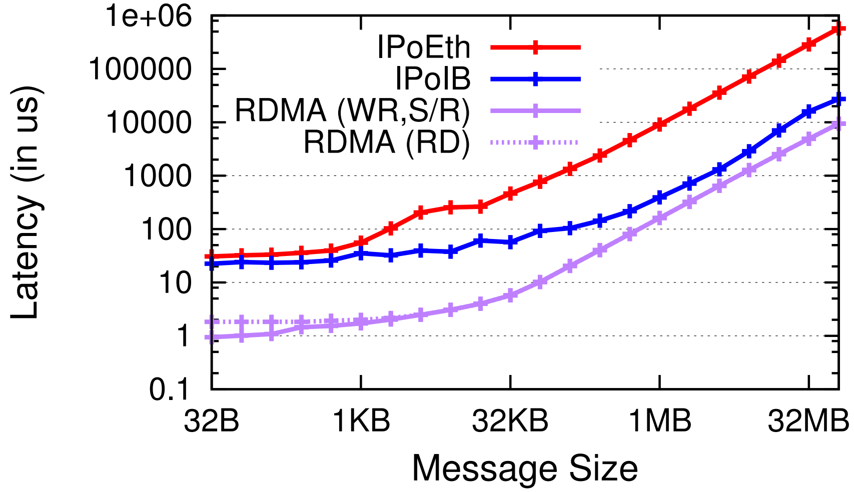


Benchmarks

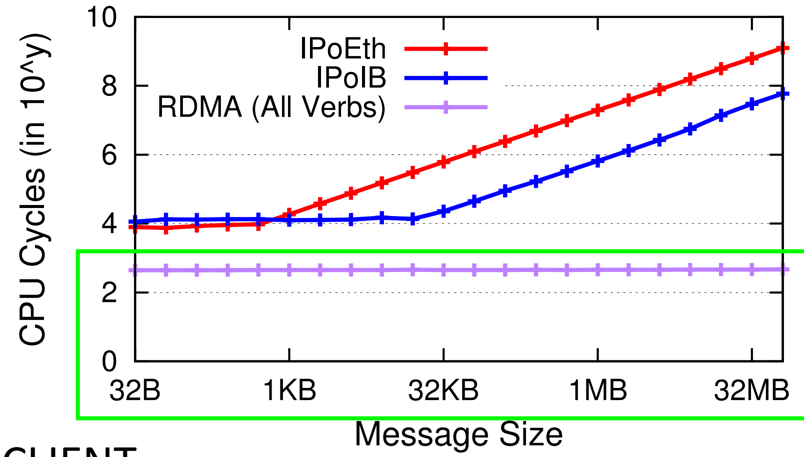
Network

InfiniBand

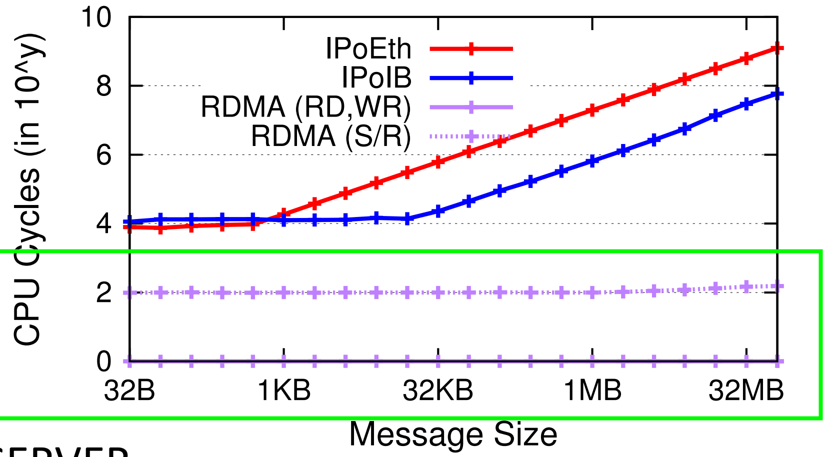




CPU



CLIENT



SERVER

RDMA

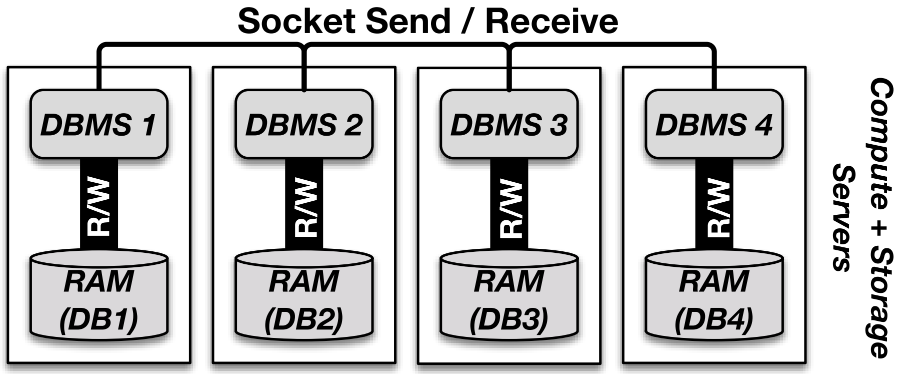
~ constant

Architecture

- Need to solve “distributed control-flow (synchronization)” and “distributed data-flow (data exchange between nodes)”
- Traditional Shared Nothing
- Shared-Nothing for IPoIB
- Distributed Shared-Memory
- Network-Attached Memory

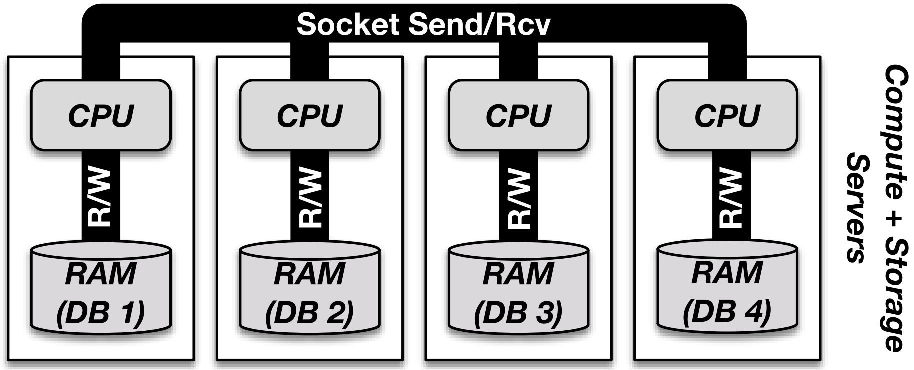
Traditional Shared Nothing

- Ethernet network
- Data transfer is slow, avoid whenever possible
- Choose partitioning carefully to minimize data transfer



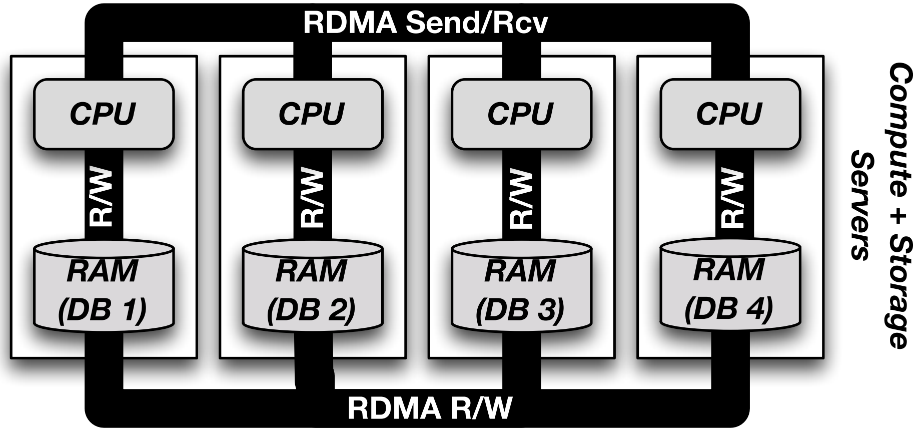
Shared-Nothing for IPoIB

- Same design, but with InfiniBand
- Good for large data movement
- High overhead of IPoIB stack means bad at small messages



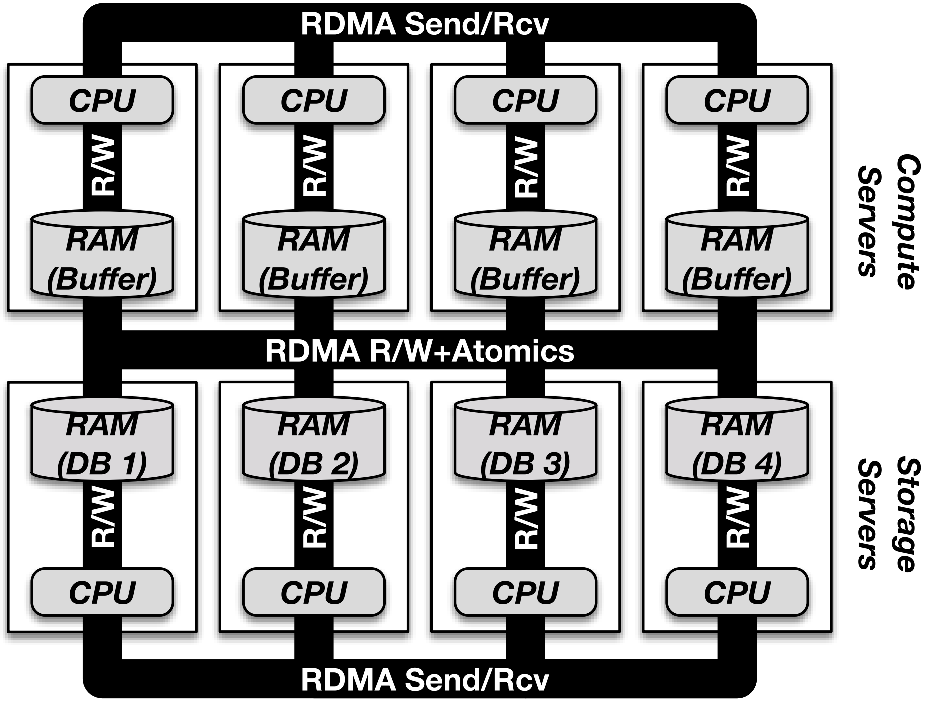
Distributed Shared-Memory

- All InfiniBand
- Create an illusion of one shared memory pool

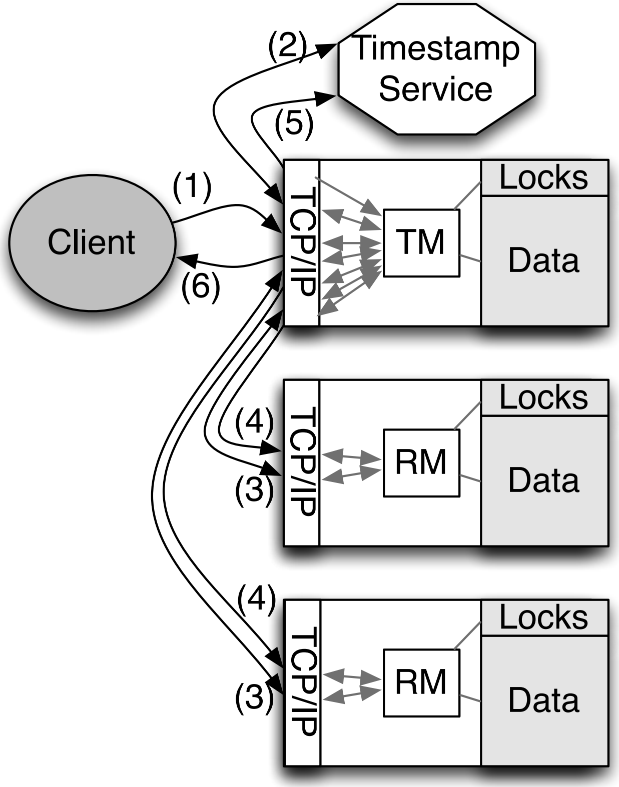


Network-Attached Memory (NAM)

- All InfiniBand
- Storage disaggregation: compute servers handle DB operations, storage servers handle data storage



OLTP: Distributed 2PC does not scale



(a) Traditional SI

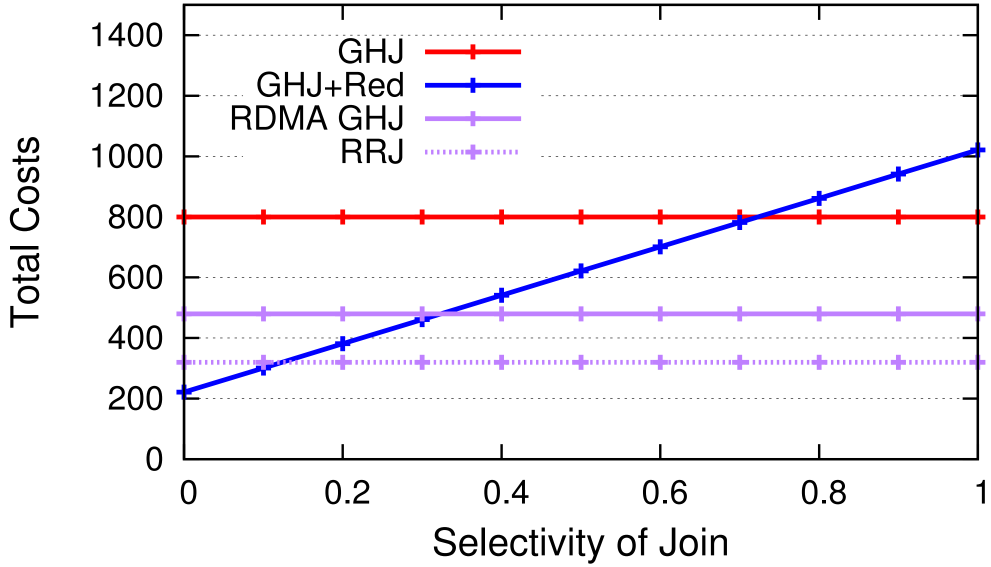
Consider generic 2PC for Snapshot Isolation

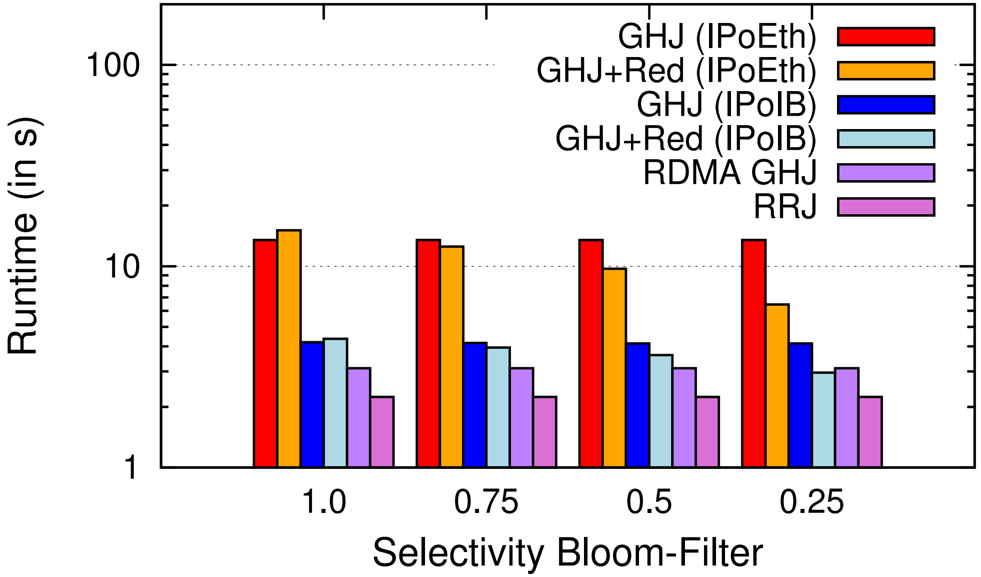
- Normal 2PC incurs 9x message delay, can be decreased to 6x
- Increased latency means more contention, more aborts
- CPU overhead consumes most extra resources gained from adding nodes to cluster

Proposed solution (“RSI”) lets clients process transactions using RDMA compare-and-swap

OLAP: Distributed Hash Join

- Most hash join optimization is decreasing data traffic: semi-join, bloom filters
- Partition across nodes, then join within nodes
- Implemented two proof of concept join algorithms
  - RDMA Grace Hash Join
  - RDMA Radix Join





Sources

- Binnig, C., Crotty, A., Galakatos, A., Kraska, T., & Zamanian, E. (2016). The end of slow networks: It’s time for a redesign. *Proceedings of the VLDB Endowment*, *9*(7), 528–539. https://doi.org/10.14778/2904483.2904485
- Ryan Stutsman, CS6450: Distributed Systems Lecture 18, https://users.cs.utah.edu/~stutsman/cs6450/public/20.pdf