HY-559
Infrastrucure Technologies for Large-Scale Service-Oriented Systems

Kostas Magoutis
magoutis@csd.uoc.gr
http://www.csd.uoc.gr/~hy559
Coordination services

• API for
  – Storing and querying cluster state
    • Live machines, association to services, roles
  – Express interest in conditions, notifications

• High availability and data consistency
  – Replication
  – Order on state updates

• Google Chubby (Paxos), Apache ZooKeeper (ZAB)
Order on state updates
Paxos algorithm

• Way to build fault-tolerant distributed systems
  – Replicated state machines (RSM)

• Consensus via message exchange
  – Asynchronous: no timing guarantees
  – Network can delay, reorder, lose (but not corrupt) packets

• Can guarantee safety
  – Replicas will agree on a single value

• Need additional assumptions to ensure progress
Informally

- Three roles: Proposer, acceptor, learner
- Simplest, but fault-intolerant solution: single acceptor
- With >1 acceptors, agreement by a majority required
- If single value proposed, that value should be chosen
  - Thus, an acceptor must accept the first value proposed to it
- However, this may lead to fragmented electorate
  - Multiple proposals by each proposer should be possible
  - Identify each proposal by a unique integer N
Informally

• After consensus, an acceptor cannot change its mind
  – A value is chosen when single proposal with that value accepted by a majority of the acceptors

• Allow multiple proposals to be chosen, but guarantee that all chosen proposals have the same value
Paxos setup

- Be able to agree in the presence of up to $f$ failures
- $2f+1$ nodes
- Agreement when majority $(f +1)$ agrees on a value