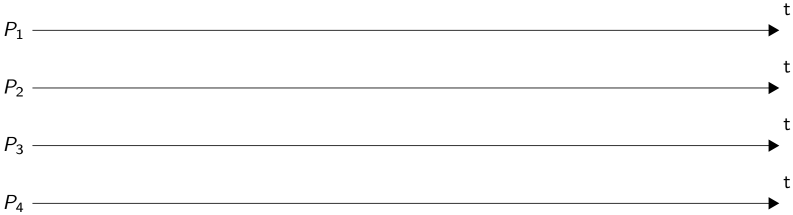
Paxos Algorithm

L. Lamport

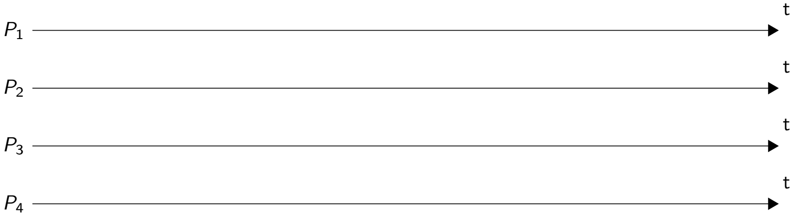
The Part-Time Parliament

2 novembre 2016

Consensus Goal



Consensus Goal



- Replicated state machine (all servers execute the same sequence of commands)
- Ensures proper log replication
- System works as there is a majority of servers up $(2f + 1)$
- Environment : Crash/Stop (not Byzantine), delayed/lost messages

Paxos

Basic Paxos

- Prepare phase
- Accept phase

Multi-Paxos

- Choosing log entries
- Leader election
- Less prepare request
- Full propagation

Requirements

- Safety : only one value,
- Liveness : some proposed value will eventually be chosen

Paxos

Actors

- Proposers
- Acceptors : How many will we need ?

P_1
t
P_2
t
P_3
t
P_4
t
P_5
t

Paxos

Actors

- Proposers
- Acceptors : How many will we need ?

P_1
P_2
P_3
P_4
P_5
t
t
t
t
t

- Acceptor accepts only first value it receives ? Acceptors must sometimes accept multiple (different) values - reject old ones

Prepare

- Each proposal has a unique number
- Block old proposals
- Know about old values - $OK(b, v)$

Accept

- Demand acceptors to accept a new value
- Response - $Voted(b, q)$

Algorithm

1. Proposer p chooses a number b greater than lastTried(p), sets lastTried(p) to b, and sends a Prepare(b) message to acceptors.
2. Upon receipt of a Prepare(b) message from p with b $>$ nextSeq(q), acceptor q sets nextSeq(q) to b and sends a OK(b, v) message to p, where v equals prevVote(q).(A Prepare(b) message is replied *KO* if b $\leq$ nextSeq(q).)
3. After receiving a OK(b, v) message from every acceptor in some majority set $Q = f + 1$, where b $=$ lastTried(p), proposer p initiates a new sequence with number b, and value d, where d is the latest chosen value from the replies or a proposed value from the proposer. He then sends a Accept(b, d) message to every acceptor.

Algorithm

4. Upon receipt of a Accept(b, d) message with b = nextSeq(q), acceptor q casts his vote in sequence number b, sets prevVote(q) to this vote, and sends a Voted(b, q) message to p.(An Accept(b, d) message is ignored if b $\neq$ nextSeq(q).)
5. If p has received a Voted(b, q) message from f+1 acceptors, where b = lastTried(p), then he writes the value d on disk and sends a Success(d) message to every acceptor.
6. Upon receiving a Success(d) message, an acceptor enters value d on disk.

Different values

Suppose that a cluster contains 5 servers and 3 of them have accepted proposal 5.1 with value X. Once this has happened, is it possible that any server in the cluster could accept a different value Y?

P_1
t
P_2
t
P_3
t
P_4
t
P_5
t