Distributed Systems

CS425/ECE428

Instructor: Radhika Mittal

Logistics

- Midterm 1 exam still ongoing – do not discuss anything about it!
- HW2 is due on Wednesday, March 11th.
- HW3 released today.
  - you should be able to tackle first question after today's class.
- MP1 is due on Friday Mar 13th.

Today's agenda

- Wrap up Leader Election
- Consensus

Bully Algorithm

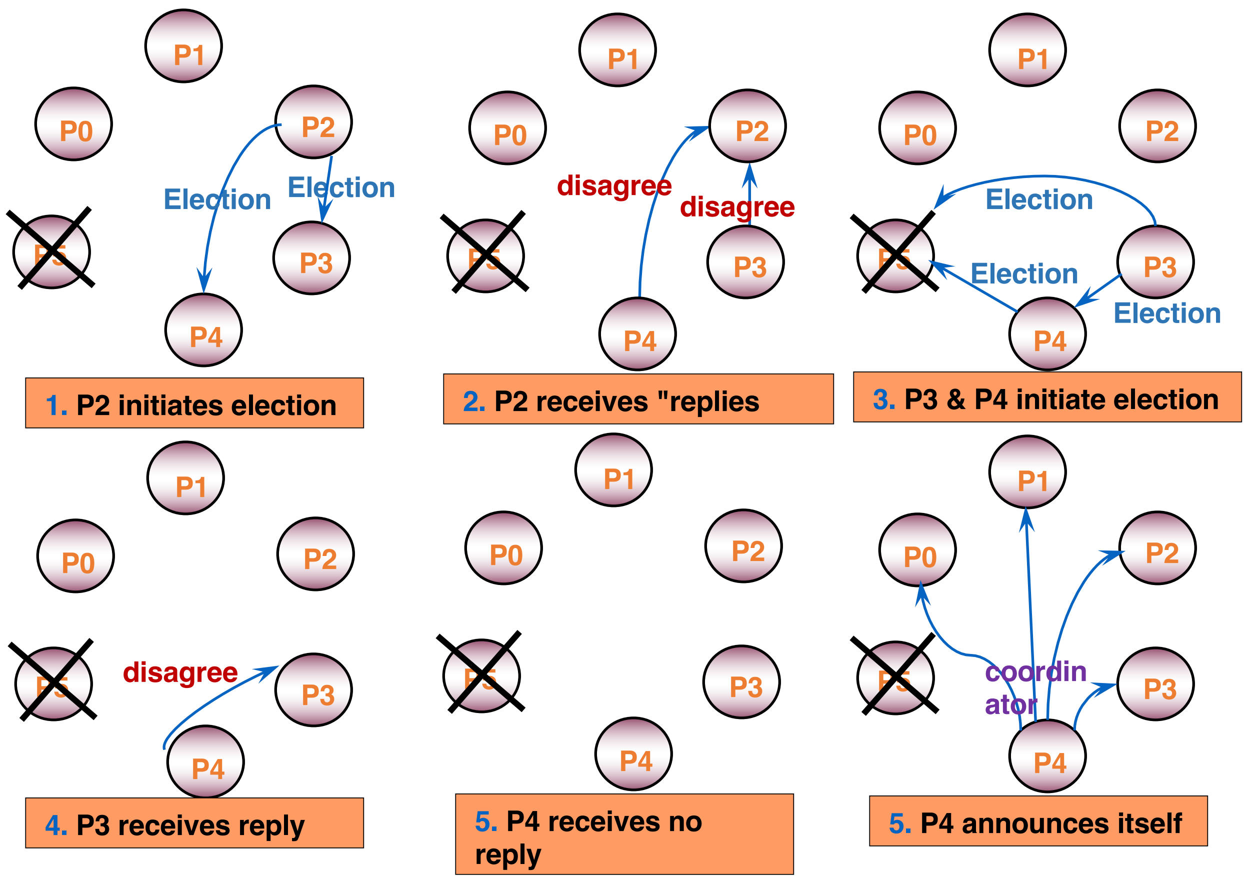
- When a process wants to initiate an election
  - **if** it knows its id is the highest
    - it elects itself as coordinator, then sends a *Coordinator* message to all processes with lower identifiers. Election is completed.
  - **else**
    - it initiates an election by sending an *Election* message
    - (contd.)

Bully Algorithm (2)

- **else** it initiates an election by sending an *Election* message
  - Sends it to only processes that have a *higher id than itself*.
  - **if** receives no answer within timeout, calls itself leader and sends *Coordinator* message to all lower id processes. Election completed.
  - **if** an answer received however, then there is some non-faulty higher process => so, wait for coordinator message. If none received after another timeout, start a new election run.
- A process that receives an *Election* message replies with *disagree* message, and starts its own leader election protocol (unless it has already done so).

Bully Algorithm: Example

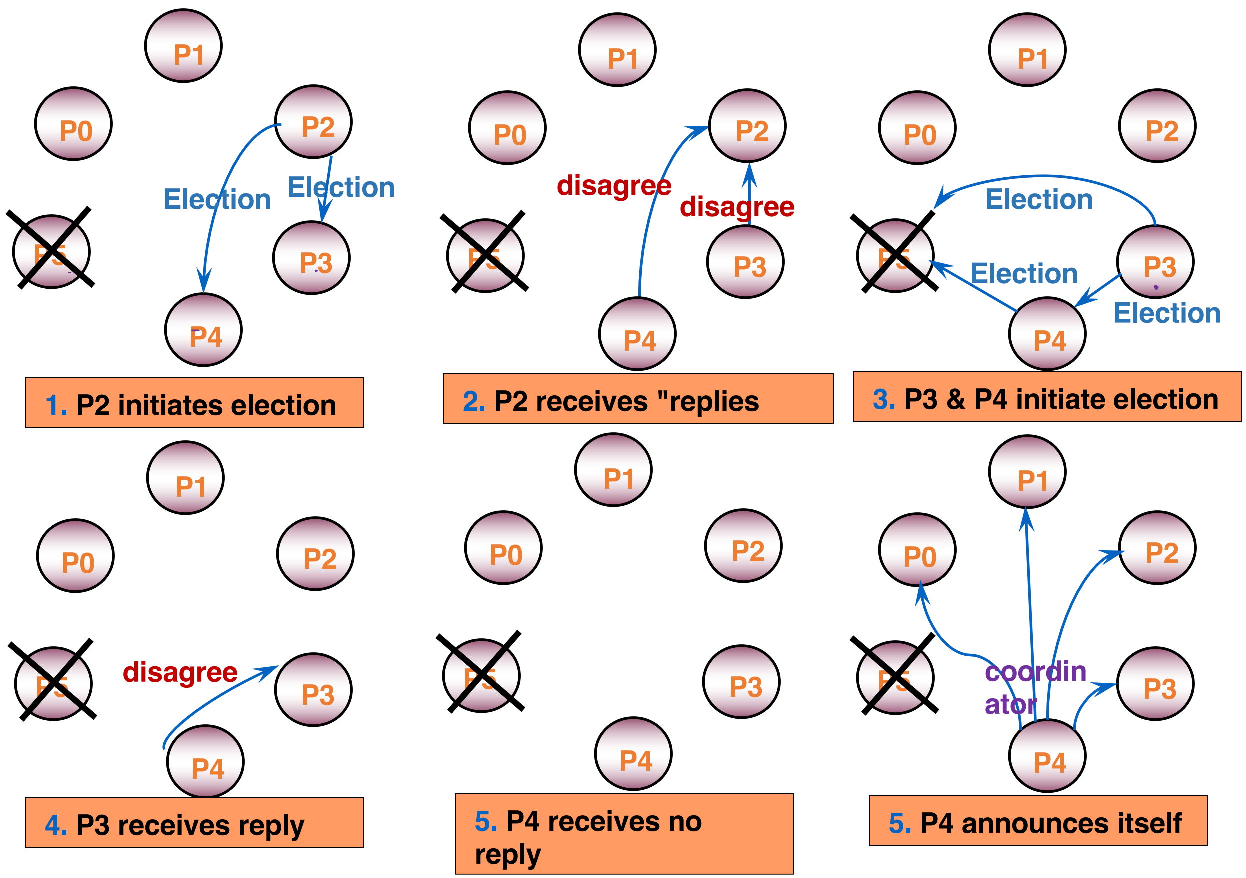
P2 initiates election after detecting P5's failure.



What if P4 fails after step 3?

Bully Algorithm: Example

P2 initiates election after detecting P5's failure.



What if P4 fails after step 4?

Bully Algorithm (2)

- **else** it initiates an election by sending an *Election* message
  - Sends it to only processes that have a *higher id than itself*.
  - **if** receives no answer within **timeout**, calls itself leader and sends *Coordinator* message to all lower id processes. Election completed.
  - **if** an answer received however, then there is some non-faulty higher process => so, wait for coordinator message. If none received after another **timeout**, start a new election run.
- A process that receives an *Election* message replies with *disagree* message, and starts its own leader election protocol (unless it has already done so).

Timeout values

- Assume the one-way message transmission time (T) is known.

- First timeout value (when the process that has initiated election waits for the first response)
  - Must be set as accurately as possible.
    - If it is too small, a lower id process can declare itself to be the coordinator even when a higher id process is alive.
  - What should be the first timeout value be, given the above assumption?
    - 2T + (processing time) ≈ 2T

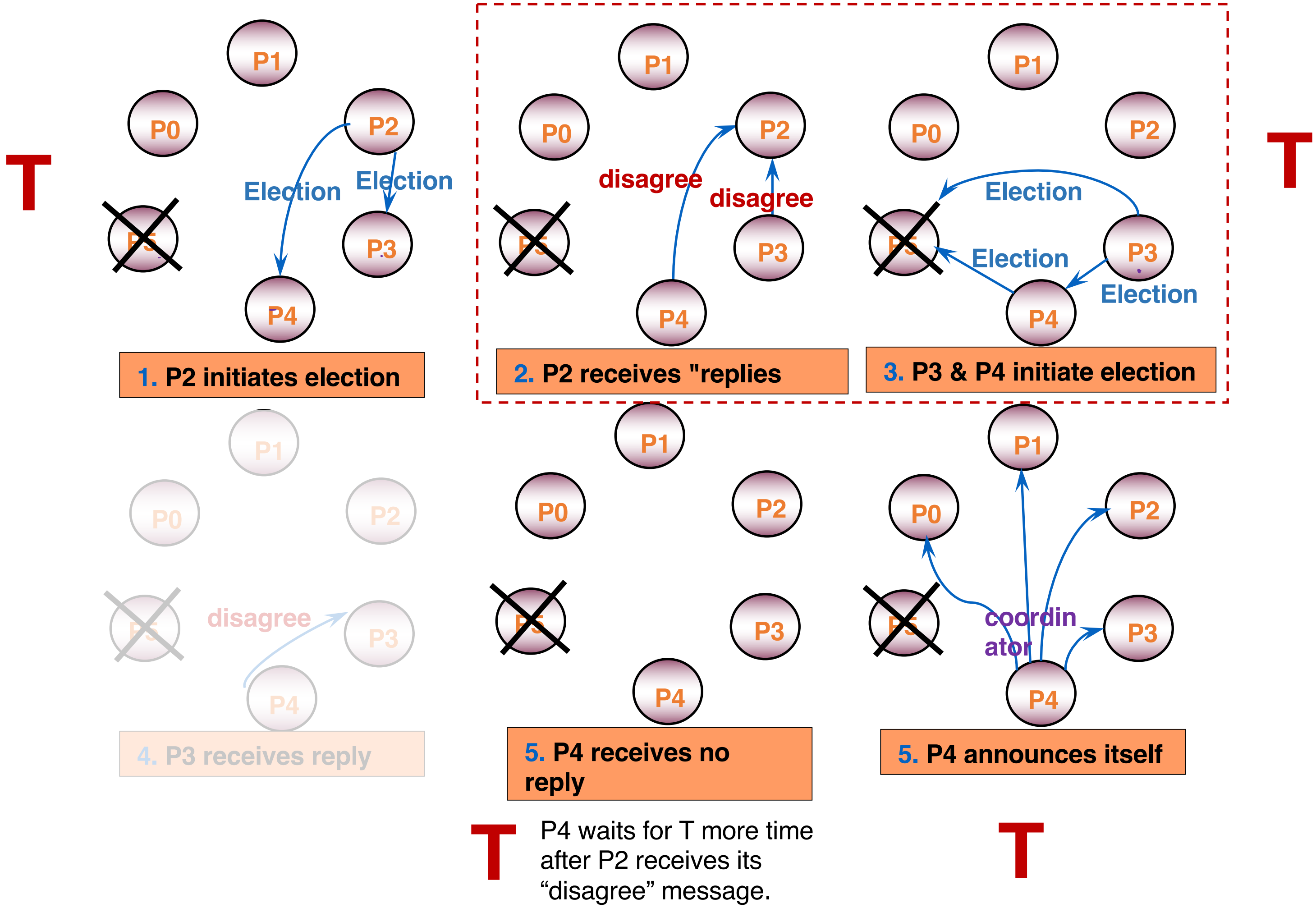
- When the second timeout happens (after 'disagree' message), election is re-started.
  - A very small value will lead to extra "Election" messages.
  - A suitable option is to use the worst-case turnaround time.

Performance Analysis

- Best-case
    - Second-highest id detects leader failure
        - Highest remaining id initiates election.
    - Sends (N-2) Coordinator messages
    - Turnaround time: 1 message transmission time (T)
- Worst-case: For simplicity, assume no failures after a process calls for election.
    - If any lower id process detects failure and starts election.
        - Turnaround time?
            - 4 message transmission times (4T)

Bully Algorithm: Example

P2 initiates election after detecting P5's failure.



Analysis

- Best-case
  - Second-highest id detects leader failure
    - Highest remaining id initiates election.
  - Sends (N-2) Coordinator messages
  - Turnaround time: 1 message transmission time
- Worst-case: For simplicity, assume no failures after a process calls for election.
  - Turnaround time: 4 message transmission times
    - if any lower id process detects failure and starts election.
    - Election + (disagree & Election) + (Timeout – T) + Coordinator
  - When the process with the lowest id in the system detects failure.
    - (N-1) processes altogether begin elections, each sending messages to processes with higher ids.
    - i-th highest id process sends (i-1) election messages
    - Number of Election messages

      $= N\text{-}1 + N\text{-}2 + \ldots + 1 = (N\text{-}1)*N/2 = O(N^2)$

Correctness

- In synchronous system model:
    - Set timeout accurately using known bounds on network delays and processing times.
    - Satisfies safety and liveness.

- In asynchronous system model:
    - Failure detectors cannot be both accurate and complete.
    - Either liveness and safety is violated.

Why is Election so hard?

- Because it is related to the consensus problem!
- If we could solve election, then we could solve consensus!
    - Elect a process, use its id's last bit as the consensus decision.
- But (as we will see next) consensus is impossible in asynchronous systems, so is election!

Summary

- Leader election is an important problem in distributed system.
  - Crucial for implementing any centralized algorithm.

- Two classical algorithms:
  - Ring election algorithm and Bully algorithm

- Hard to guarantee correctness in an asynchronous system with failures.

Today's agenda

- **Consensus**
- Goals:
  - Understand the problem of consensus
  - How to achieve consensus in a synchronous system
  - Difficulty of achieving consensus in an asynchronous system
  - Good-enough consensus algorithms for asynchronous systems

Agenda for the next few weeks

- **Consensus**
  - Consensus in synchronous systems
    - *Chapter 15.4*
  - Impossibility of consensus in asynchronous systems
    - *We will not cover the proof in details*
  - Good enough consensus algorithm for asynchronous systems:
    - *Paxos made simple, Leslie Lamport, 2001*
  - Other forms of consensus algorithm
    - Raft (log-based consensus)
    - Block-chains (distributed consensus)

Agenda for today *(and maybe next class)*

- **Consensus**
  - Consensus in synchronous systems
    - *Chapter 15.4*
  - Impossibility of consensus in asynchronous systems
    - *We will not cover the proof in details*
  - A good enough consensus algorithm for asynchronous systems:
    - *Paxos made simple, Leslie Lamport, 2001*
  - Other forms of consensus
    - Blockchains
    - Raft (log-based consensus)

Consensus

- **Each process proposes a value.**
- **All processes must agree on one of the proposed values.**
- Examples:
  - The generals must agree on the time of attack.
  - An object replicated across multiple servers in a distributed data store.
    - All servers must agree on the current version of the object.
  - Transaction processing on replicated servers
    - Must agree on the order in which updates are applied to an object.
  - …..

Consensus

- **Each process proposes a value.**
- **All processes must agree on one of the proposed values.**
- The final value can be decided based on any criteria:
  - Pick minimum of all proposed values.
  - Pick maximum of all proposed values.
  - Pick the majority (with some deterministic tie-breaking rule).
  - Pick the value proposed by the *leader.*
    - *All processes must agree on who the leader is.*
  - If reliable total-order can be achieved, pick the proposed value that gets delivered first.
    - *All process must agree on the total order.*
  - ……

Consensus Problem

- System of N processes (P_1, P_2, …..., P_n)
- Each process P_i:
  - begins in an *undecided* state.
  - proposes value $\mathbf{v_i}$.
  - at some point during the run of a consensus algorithm, sets a decision variable $\mathbf{d_i}$ and enters the *decided* state.

Required Properties

- **Termination:** Eventually each process sets its decision variable.

- **Agreement:** The decision value of all correct processes is the same.
  - If P_i and P_j are correct and have entered the *decided* state, then $\mathbf{d_i = d_j}$.

- **Integrity:** If the correct processes all proposed the same value, then any correct process in the decided state has chosen that value.
  - *Safeguard against algorithms that decide on a fixed constant value.*

Required Properties

- **Termination:** Eventually each process sets its decision variable.
- **Agreement:** The decision value of all correct processes is the same.
  - If P_i and P_j are correct and have entered the *decided* state, then $\mathbf{d_i = d_j}$.
- **Integrity:** If the correct processes all proposed the same value, then any correct process in the decided state has chosen that value.

Which of these properties is liveness and which is safety?

Required Properties

- **Termination:** Eventually each process sets its decision variable.
  - *Liveness*
- **Agreement:** The decision value of all correct processes is the same.
  - If P_i and P_j are correct and have entered the *decided* state, then $\mathbf{d_i = d_j}$.
  - *Safety*
- **Integrity:** If the correct processes all proposed the same value, then any correct process in the decided state has chosen that value.

How do we agree on a value?

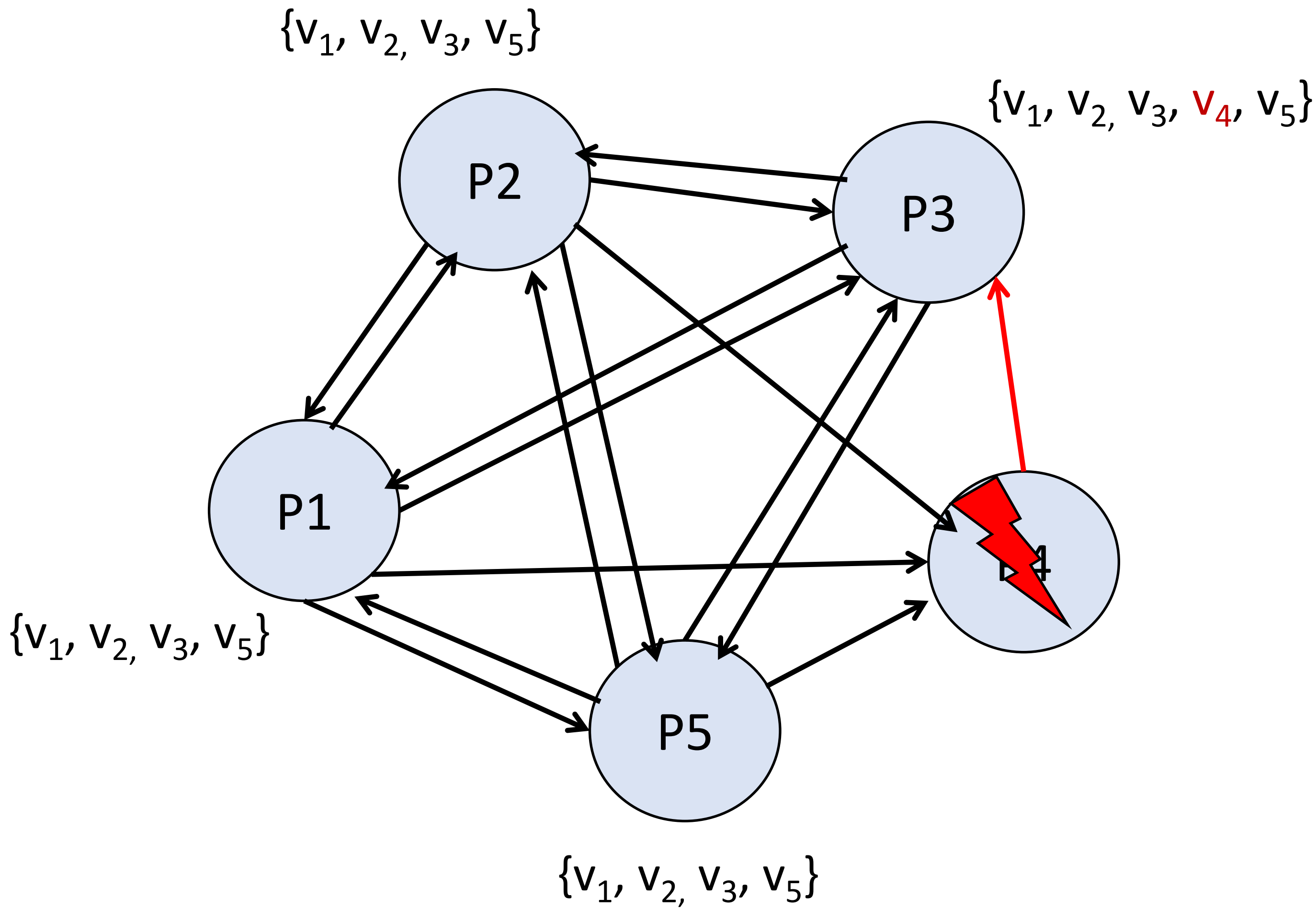
- Ring-based leader election
  - Send proposed value along with *elected* message.
  - Turnaround time: 3NT worst case and 2NT best case (without failures).
    - T is the time taken to transmit a message on a channel.
  - O(NfT) if up to f processes fail during the election run.
  - Can we do better?
- Bully algorithm
  - Send proposed value along with the *coordinator* message.
  - Turnaround time: 4T in the worst case without failures.
  - More than 4fT if up to f processes fail during the election run.

What's the best we can do?

Consider the simplest algorithm

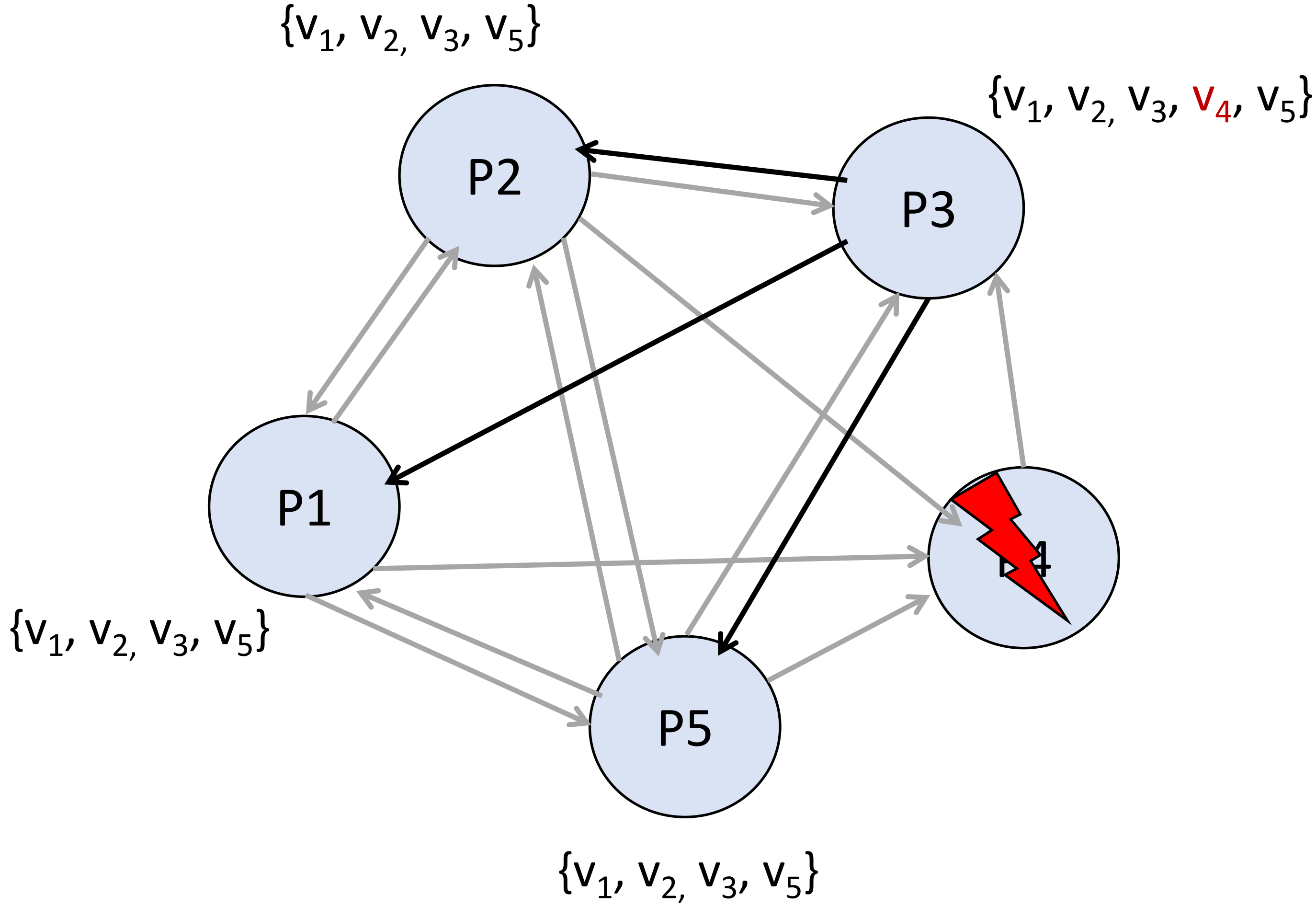
- **Let's assume the system is synchronous.**
- Use a simple B-multicast:
  - All processes B-multicast their proposed value to all other processes.
  - Upon receiving all proposed values, pick the minimum.
- Time taken under no failures?
  - One message transmission time (T)
- What can go wrong?
  - If we consider process failures, is a simple B-multicast enough?

What can go wrong with B-multicast?



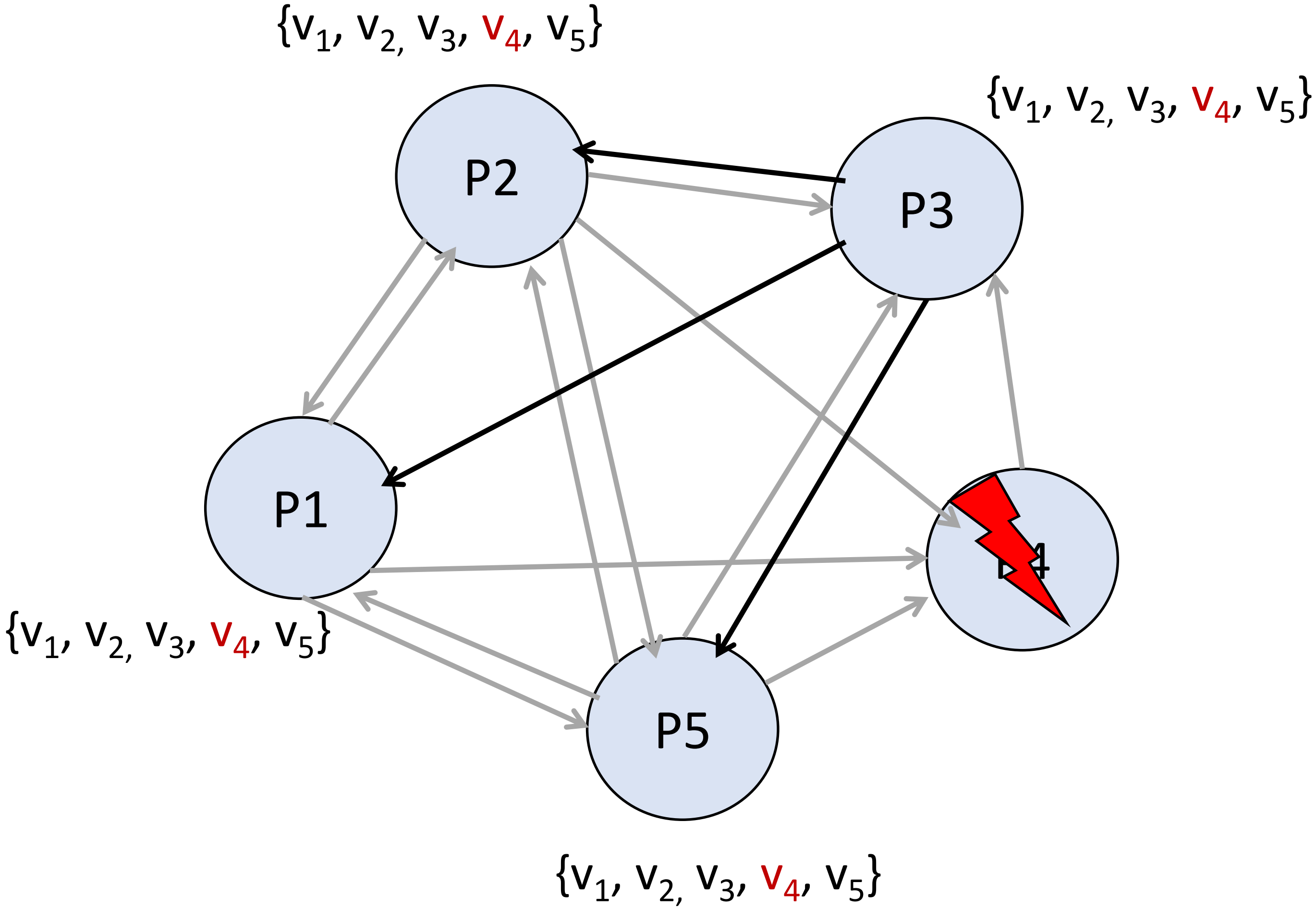
Need R-multicast

B-multicast is not enough for this



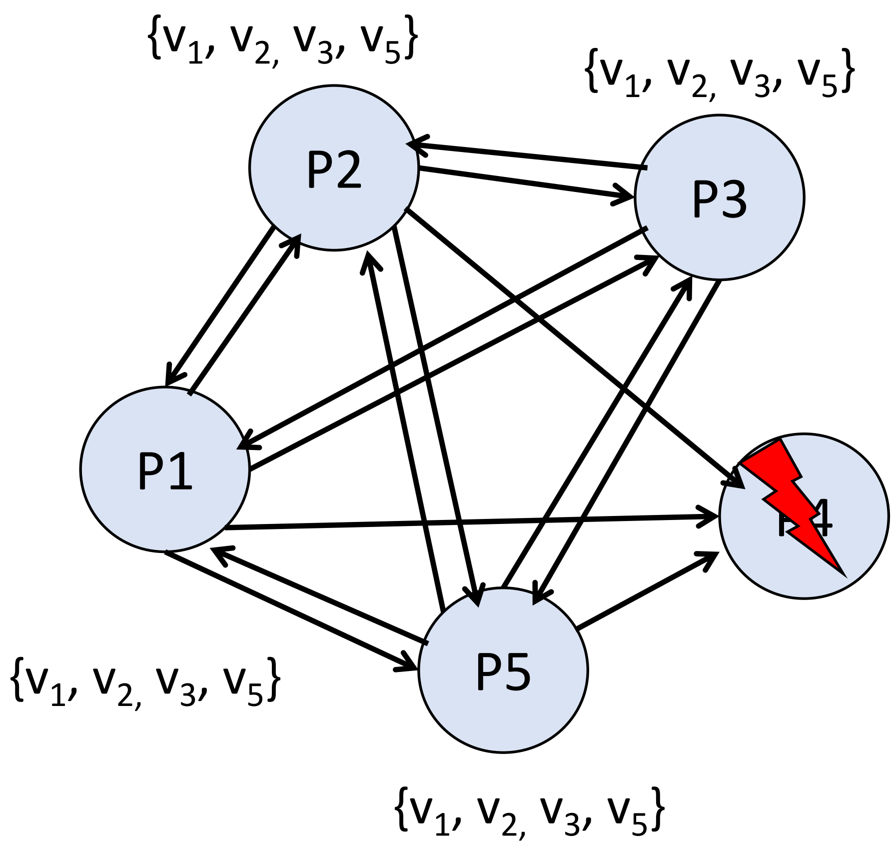
Need R-multicast

B-multicast is not enough for this



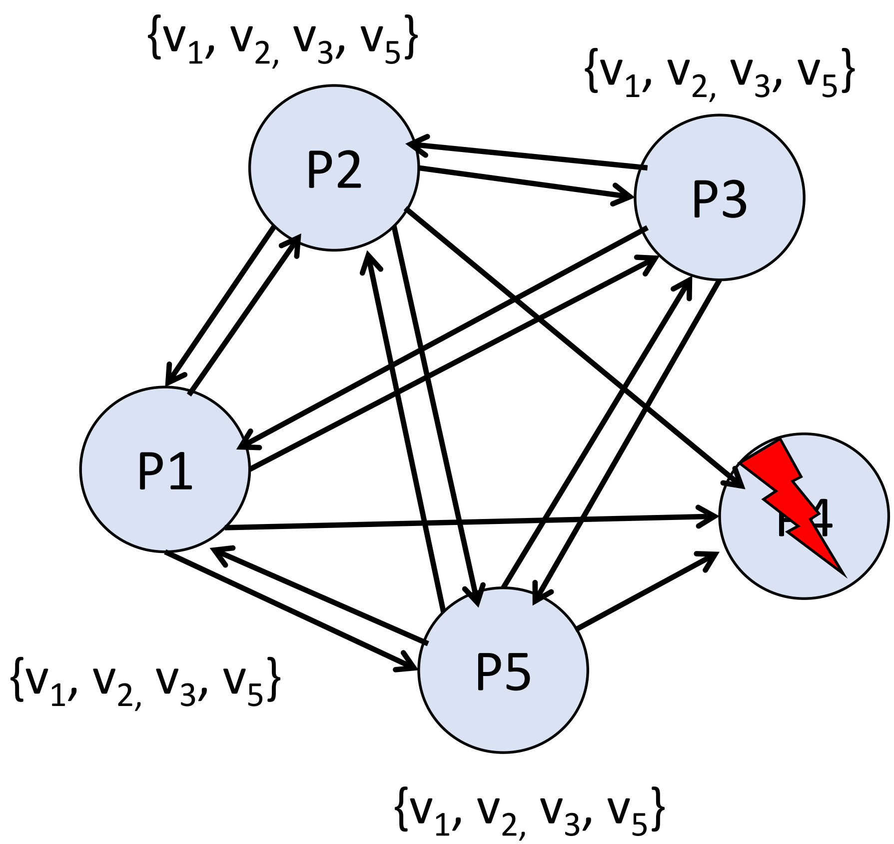
Need R-multicast

Handling failures



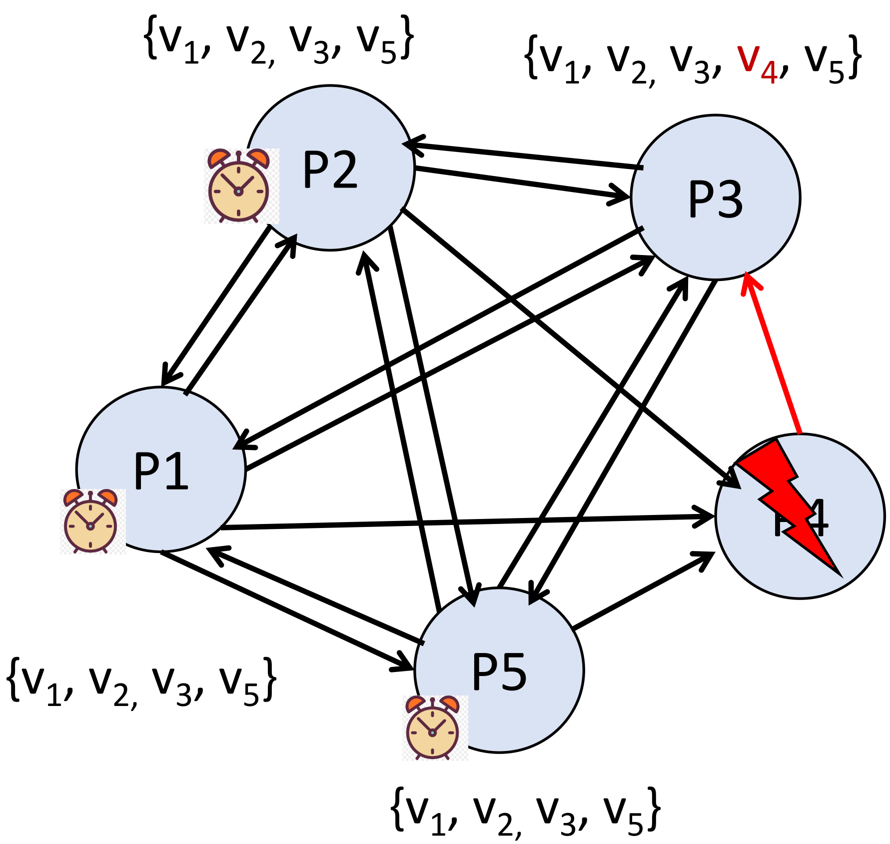
- P4 fails before sending v_4 to anyone.
- What should other processes do?
- Detect failure. *Timeout!*

Handling failures



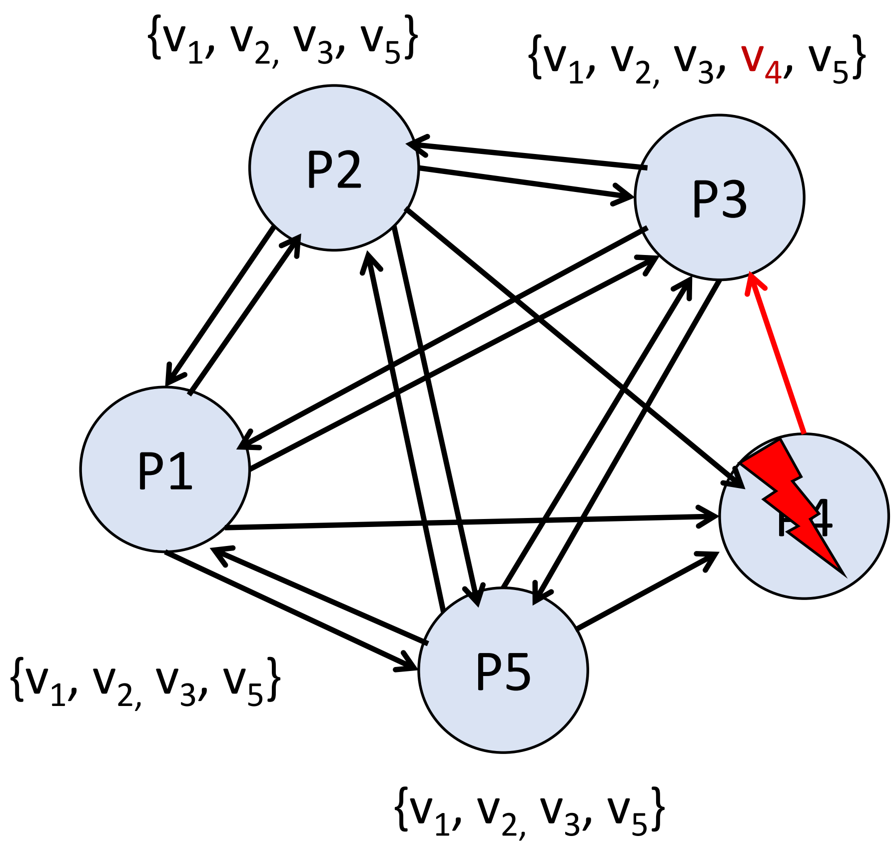
- Assume proposals are sent at time 's'.
- Worst-case skew is ϵ.
- Maximum message transfer time (including local processing) is T.
- What should the timeout value be?
- Option 1: $\epsilon + T$
  - Pi waits for ($\epsilon + T$) time units after sending its proposal at time 's'.
  - Any other process must have sent proposed value before s + ϵ.
  - The proposed value should have reached Pi by (s + $\epsilon + T$).
  - *Will this work?*

Handling failures



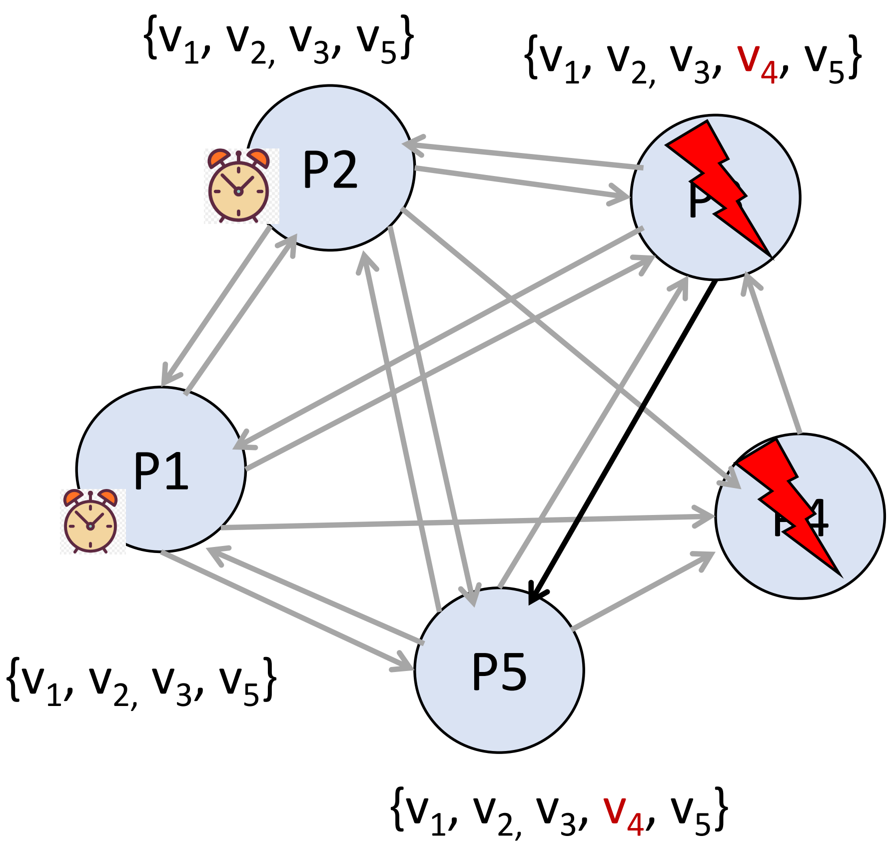
- Assume proposals are sent at time 's'.
- Worst-case skew is ϵ.
- Maximum message transfer time (including local processing) is T.
- What should the timeout value be?
- Option 1: $\epsilon + T$
  - Pi waits for ($\epsilon + T$) time units after sending its proposal at time 's'.
  - Any other process must have sent proposed value before s + ϵ.
  - The proposed value should have reached Pi by ($s + \epsilon + T$).
  - *Will this work?*

Handling failures



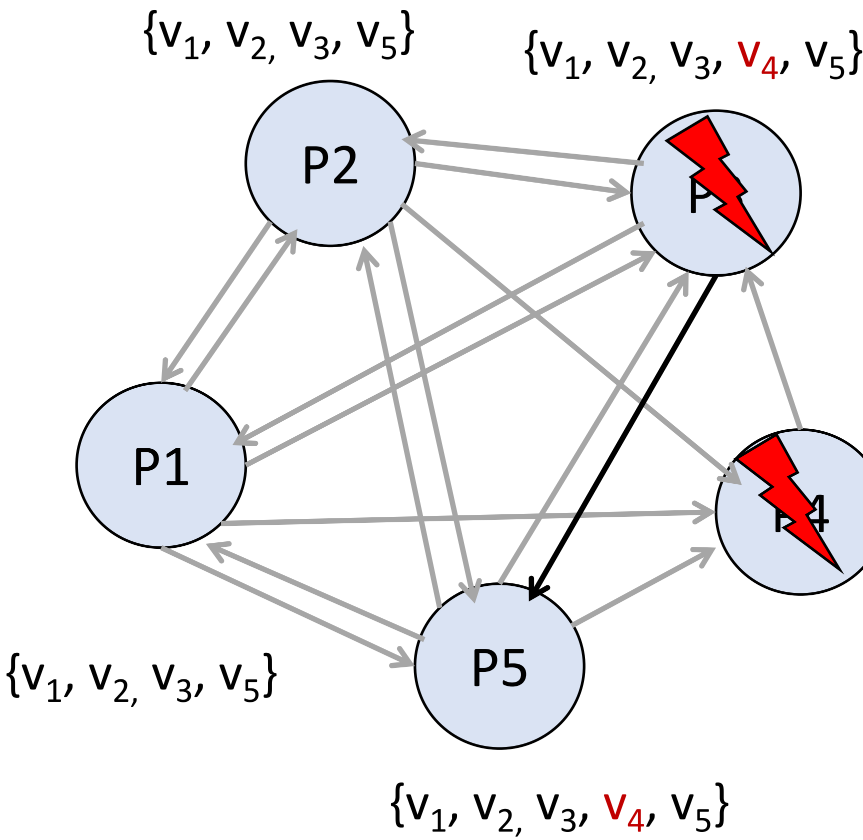
- Assume proposals are sent at time 's'.
- Worst-case skew is ϵ.
- Maximum message transfer time (including local processing) is T.
- What should the timeout value be?
- How about $\epsilon + 2*T$?
  - *Will this work?*

Handling failures



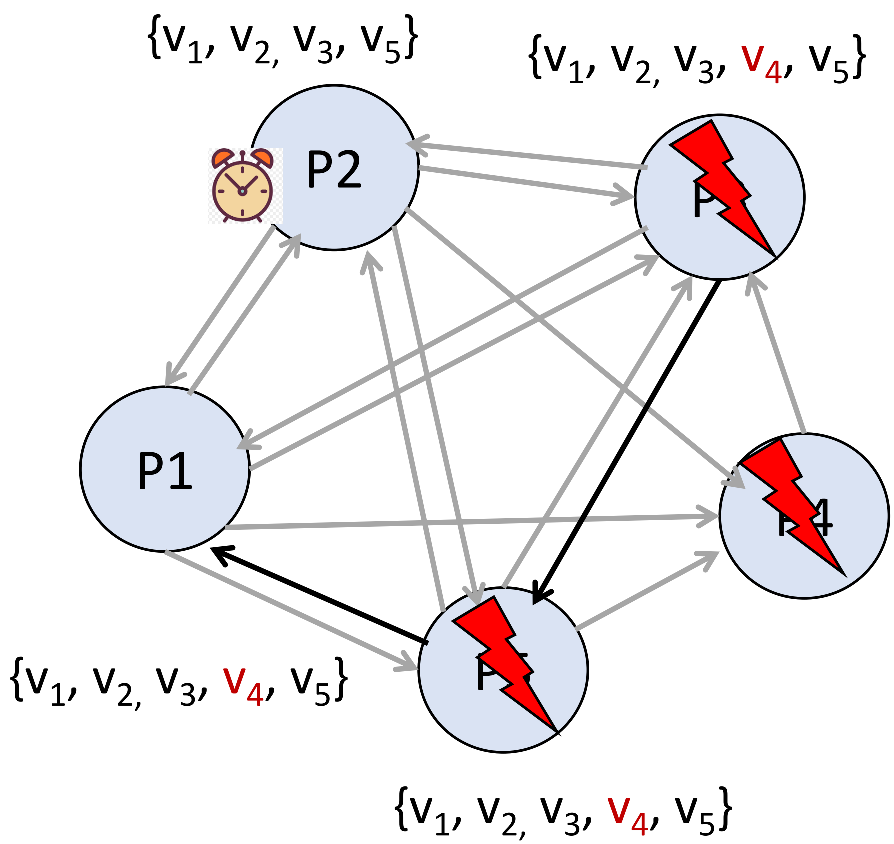
- Assume proposals are sent at time ‘s’.
- Worst-case skew is ϵ.
- Maximum message transfer time (including local processing) is T.
- What should the timeout value be?
- How about $\epsilon + 2*T$?
  - *Will this work?*

Handling failures



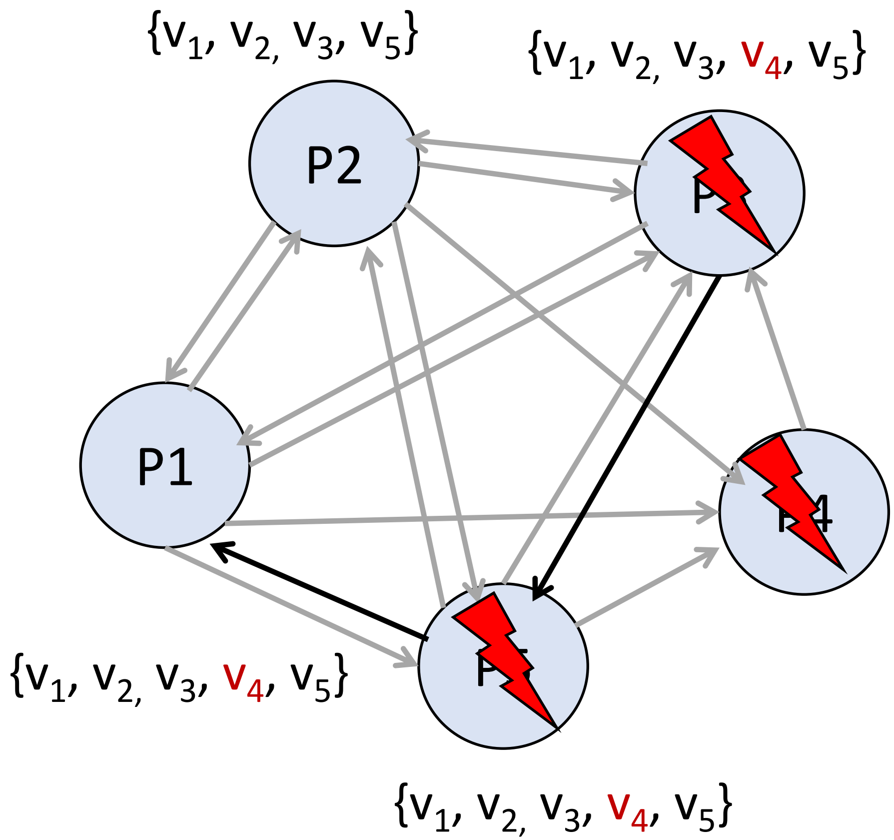
- Assume proposals are sent at time 's'.
- Worst-case skew is ϵ.
- Maximum message transfer time (including local processing) is T.
- What should the timeout value be?
- How about $\epsilon + 3*T$?
  - *Will this work?*

Handling failures



- Assume proposals are sent at time 's'.
- Worst-case skew is ϵ.
- Maximum message transfer time (including local processing) is T.
- What should the timeout value be?
- How about $\epsilon +$ 3*T?
  - *Will this work?*

Handling failures



- Assume proposals are sent at time 's'.
- Worst-case skew is ϵ.
- Maximum message transfer time (including local processing) is T.
- What should the timeout value be?
- Timeout = $\epsilon + (f+1)*T$ for up to f failed process.

Also holds for R-multicast from a single sender.

Round-based algorithm

- For a system with at most f processes crashing
  - All processes are *synchronized* and operate in "rounds" of time.
    - One round of time is equivalent to $\epsilon + T$ units.
    - At each process, the r^{th} round
      - starts at local time $s + (r - 1)*(\epsilon + T)$
      - ends at local time $s + r*(\epsilon + T)$
    - The start or end time of a round in two different processes differs by at most ϵ.
  - The algorithm proceeds in f+1 rounds.
  - Assume communication channels are reliable.

Round-based algorithm

$Values^r_i$: the set of proposed values known to P_i at the beginning of round r.

Initially $Values^1_i = \{v_i\}$

for r = 1 to f+1 do

B-multicast ($Values^r_i - Values^{r-1}_i$)

// iterate through processes, send each a message

$Values^{r+1}_i \leftarrow Values^r_i$

wait until one round of time expires.

for each v_j received in this round

$Values^{r+1}_i = Values^{r+1}_i \cup v_j$

end

end

d_i = minimum($Values^{f+2}_i$)

Why does this work?

- After f+1 rounds, all non-faulty processes would have received the same set of values.
- *Proof by contradiction.*
- Assume that two non-faulty processes, say P_i and P_j , differ in their final set of values (i.e., after f+1 rounds)
- Assume that P_i possesses a value v that P_j does not possess.
  - → P_i must have received v in the very last round, else P_i would have sent v to P_j in that last round
  - → So, in the last round: a third process, P_k, must have sent v to P_i, but then crashed before sending v to P_j.
  - → Similarly, a fourth process sending v in the last-but-one round must have crashed; otherwise, both P_k and P_j should have received v.
  - → Implies at least one (unique) crash in each of the preceding rounds.
  - → This means a total of f+1 crashes, contradicts our assumption of up to f crashes.

Consensus in synchronous systems

Dolev and Strong proved that for a system with up to *f* failures (or faulty processes), at least *f*+*1* rounds of information exchange is required to reach an agreement.

What about asynchronous systems?

- Using time-based "rounds" or timeouts may not work.
- Cannot guarantee both completeness and accuracy for failure detection.
  - Cannot differentiate between an extremely slow process and a failed process.
- Key intuition behind the famous FLP result on the impossibility of consensus in asynchronous systems.
  - *Impossibility of Distributed Consensus with One Faulty Process, Fischer-Lynch-Paterson (FLP), 1985*
  - Stopped many distributed system designers dead in their tracks.
  - A lot of claims of "reliability" vanished overnight.
  - *(Proof is not in your syllabus – optional self-study)*

What about asynchronous systems?

- We cannot "solve" consensus in asynchronous systems.
  - We cannot meet both safety and liveness requirements.
  - Maybe it is ok to guarantee just one requirement.
- Option 1:
  - Let's set super conservative timeout for a terminating algorithm.
  - Safety violated if a process (or the network) is very, very slow.
- Option 2:
  - Let's focus on guaranteeing *safety* under all possible scenarios.
  - If the real situation is not too dire, hopefully the algorithm will terminate.

Paxos Consensus Algorithm

- Paxos algorithm for consensus in asynchronous systems.
    - Most popular consensus-algorithm.
    - A lot of systems use it
        - Zookeeper (Yahoo!), Google Chubby, and many other companies.
    - Not guaranteed to terminate, but never violates safety.

Paxos Consensus Algorithm

- *Guess who invented it?*
  - Leslie Lamport!
- Original paper: The Part-time Parliament.
  - Used analogy of a "part-time parliament" on an ancient Greek island of Paxos.
  - No one understood it.
  - The paper was rejected.
- Published "*Paxos made simple*" 10 years later.
- To be continued in next class……