Distributed Systems
Master of Science in Engineering in Computer Science

AA 2019/2020

LECTURE 5: LEADER ELECTION

Recap on Timing Assumptions

Synchronous

- timing assumptions are <u>explicit</u> either on
  - Bounds on process executions and communication channels, or
  - Existence of a common global clock, or
  - Both

Asynchronous

- there are no timing assumptions

Recap on Timing Assumptions

Partial synchrony requires abstract timing assumptions (after an unknown time t the system becomes synchronous)

Two choices:

1. Put assumption on the system model (including links and processes)
2. Create a separate abstractions that encapsulates those timing assumptions

Note: manipulating time inside a protocol/algorithm is complex and the correctness proof may become very involved and sometimes prone to errors

An alternative

Sometimes, we may be interested in knowing one process that is alive instead of monitoring failures

- E.g., Need of a coordinator

We can use a different oracle (called *leader election* module) that reports a process that is alive

Leader Election Specification

Module 2.7: Interface and properties of leader election

Module:

Name: LeaderElection, **instance** *le*.

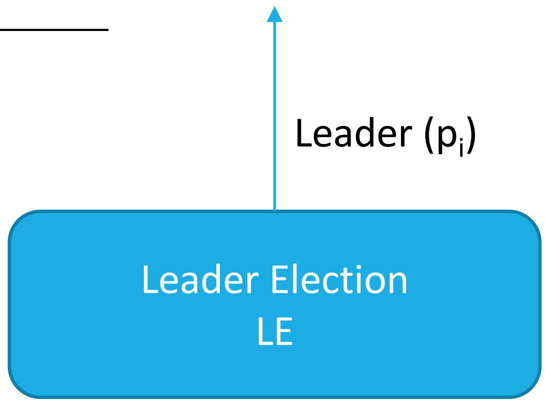
Events:

Indication: $\langle$ *le*, *Leader* $\mid p \rangle$: Indicates that process p is elected as leader.

Properties:

LE1: *Eventual detection:* Either there is no correct process, or some correct process is eventually elected as the leader.

LE2: *Accuracy:* If a process is leader, then all previously elected leaders have crashed.



Leader Election Implementation

Algorithm 2.6: Monarchical Leader Election

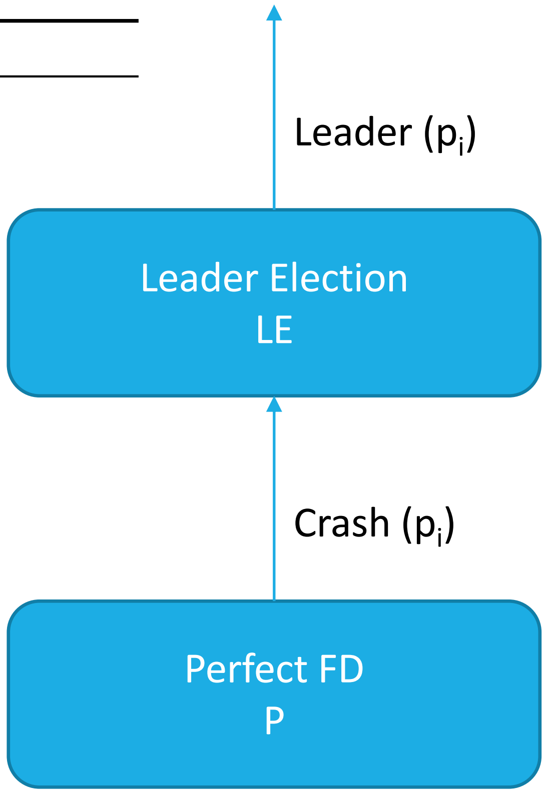
Implements:
LeaderElection, **instance** *le*.

Uses:
PerfectFailureDetector, **instance** $\mathcal{P}$.

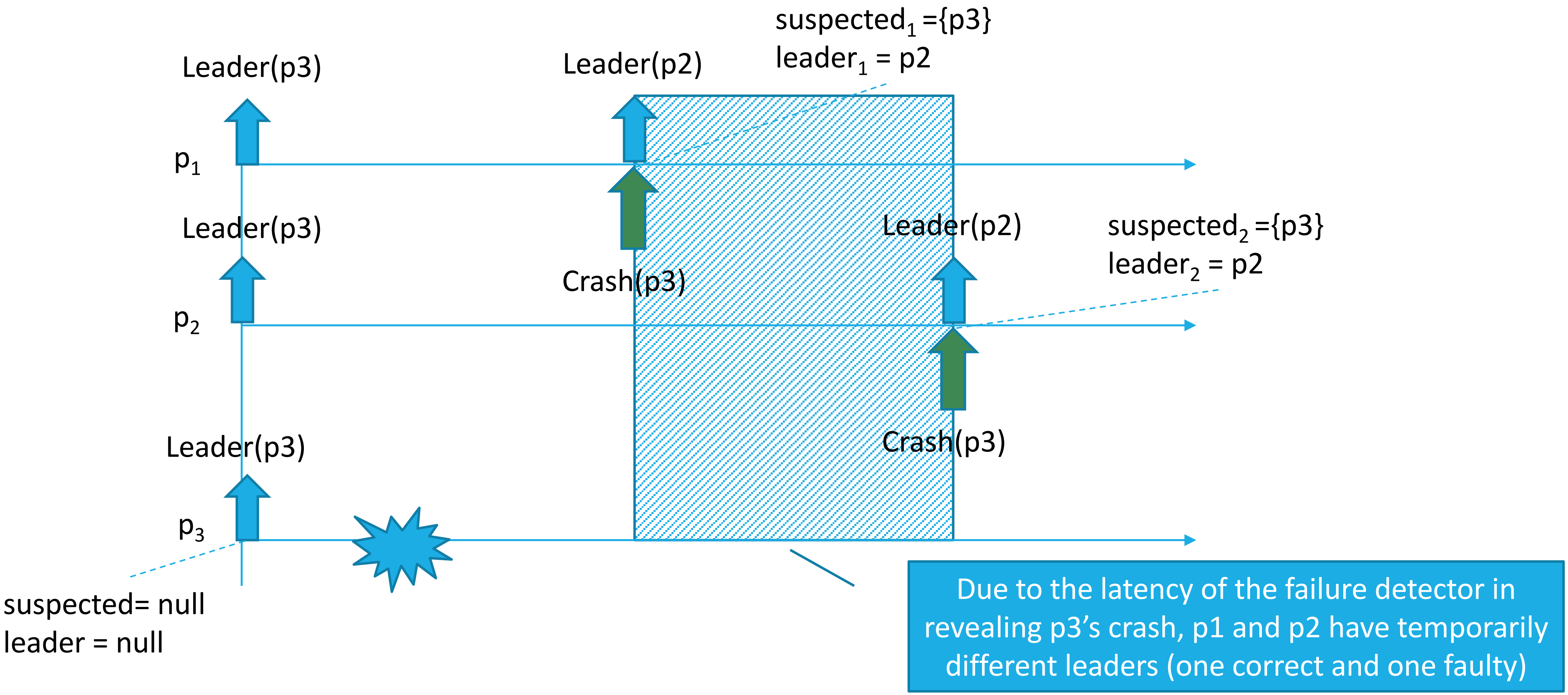
upon event $\langle$ *le*, *Init* $\rangle$ **do**
$suspected := \emptyset$;
$leader := \perp$;

upon event $\langle\, \mathcal{P}, \mathit{Crash} \mid p \,\rangle$ **do**
$suspected := suspected \cup \{p\}$;

upon $leader \neq \mathrm{maxrank}(\Pi \setminus suspected)$ **do**
$leader := \mathrm{maxrank}(\Pi \setminus suspected)$;
trigger $\langle\, le, \mathit{Leader} \mid leader \,\rangle$;

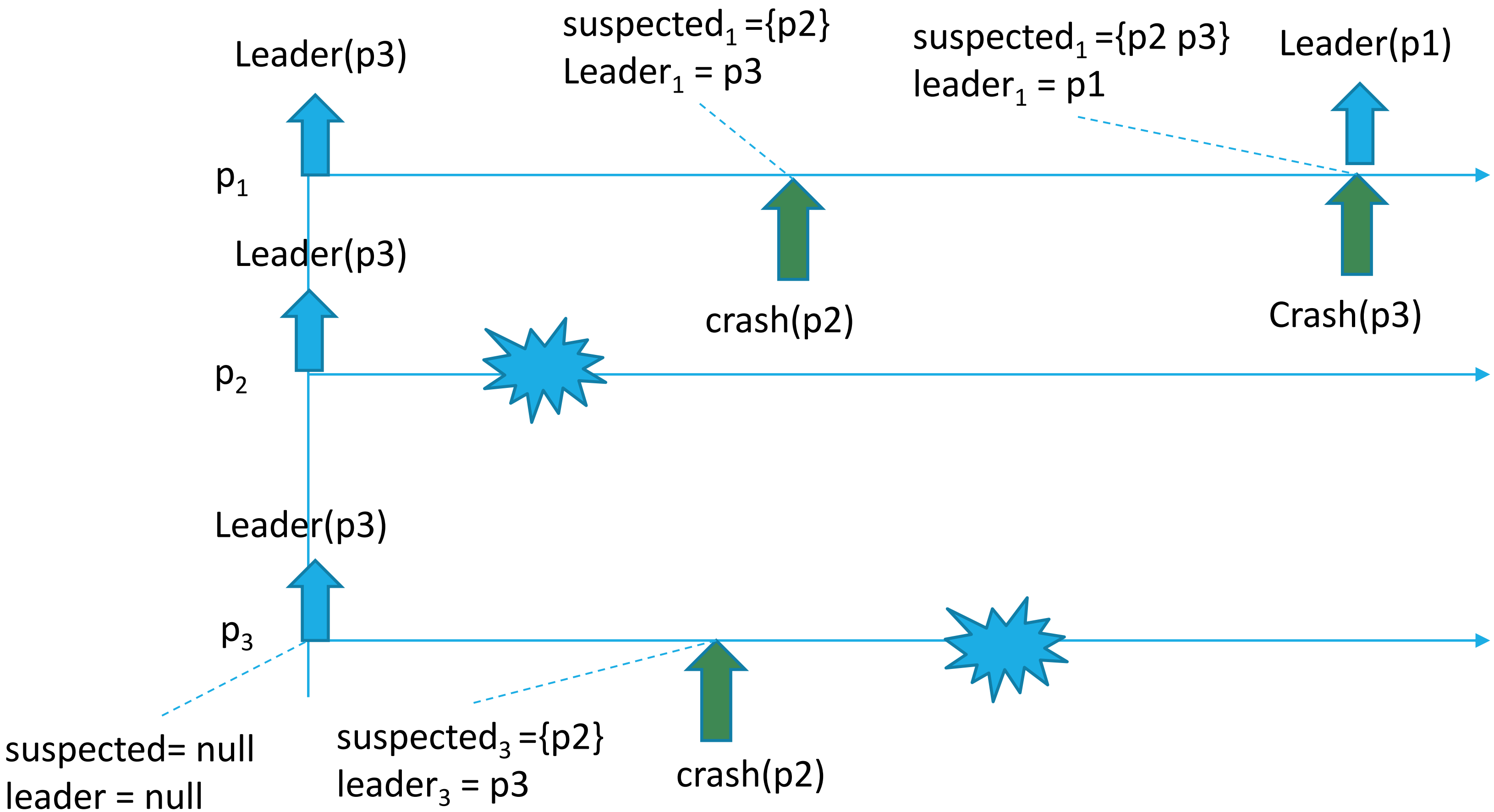


Example



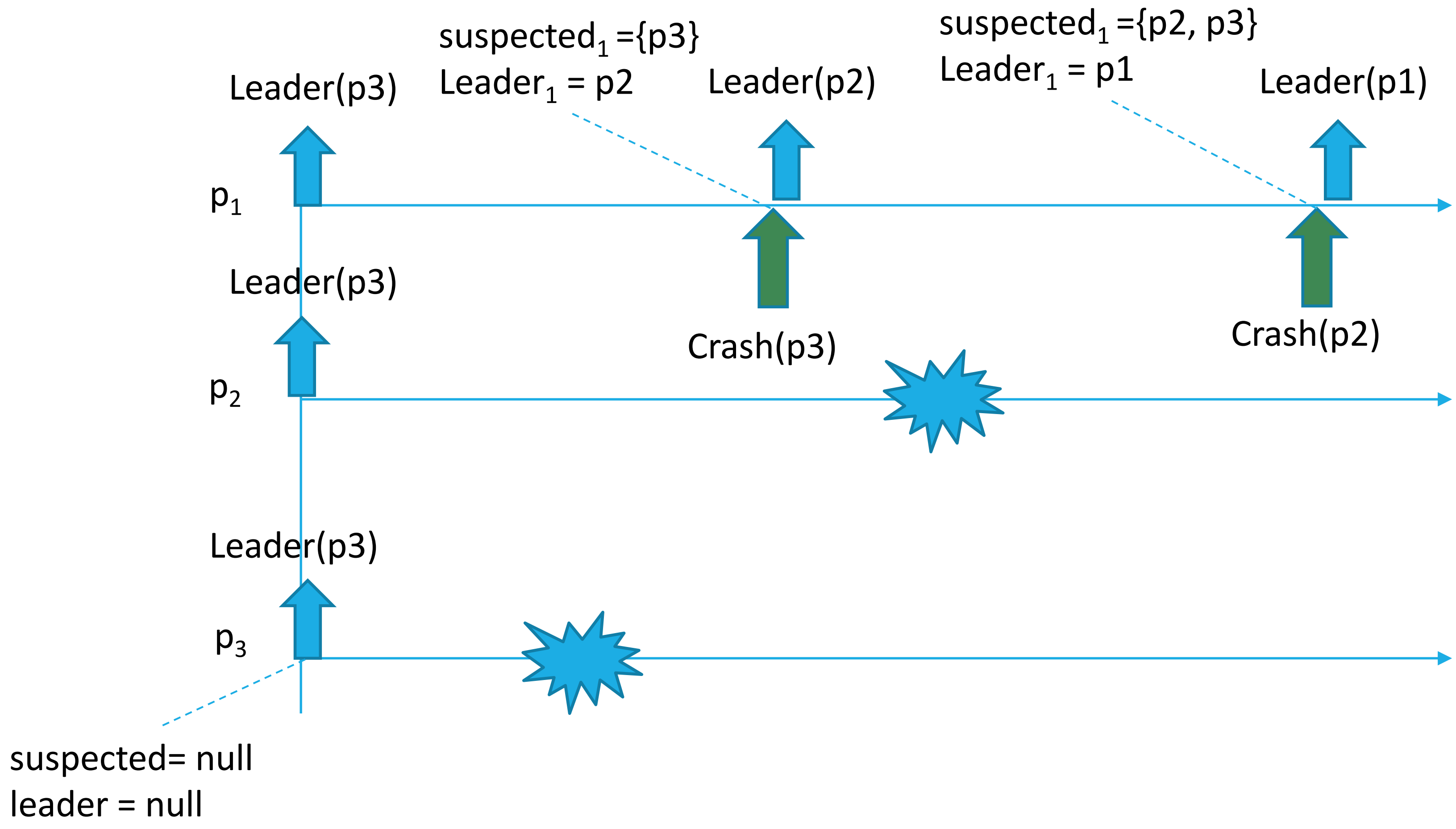
suspected$_1$ ={p3}
leader$_1$ = p2
Leader(p3)
Leader(p2)
p$_1$
Leader(p3)
Leader(p2)
suspected$_2$ ={p3}
leader$_2$ = p2
Crash(p3)
p$_2$
Crash(p3)
Leader(p3)
p$_3$
suspected= null
leader = null
Due to the latency of the failure detector in revealing p3's crash, p1 and p2 have temporarily different leaders (one correct and one faulty)

Example 2



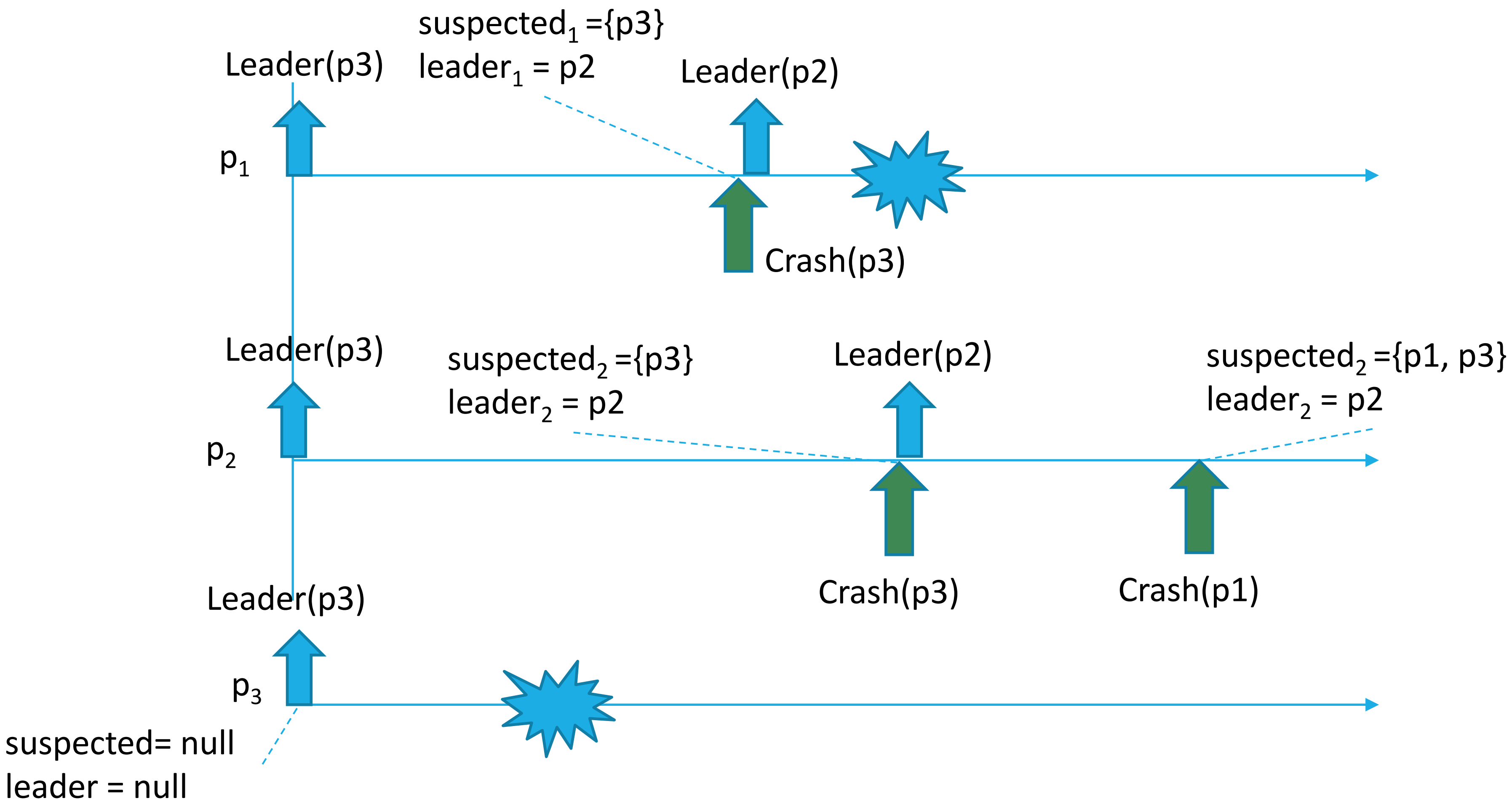
Leader(p3)
suspected$_1$ ={p2}
Leader$_1$ = p3
suspected$_1$ ={p2 p3}
leader$_1$ = p1
Leader(p1)
p$_1$
Leader(p3)
crash(p2)
Crash(p3)
p$_2$
Leader(p3)
p$_3$
suspected= null
leader = null
suspected$_3$ ={p2}
leader$_3$ = p3
crash(p2)

Example 3



suspected$_1$ ={p3}
Leader$_1$ = p2
suspected$_1$ ={p2, p3}
Leader$_1$ = p1
Leader(p3)
Leader(p2)
Leader(p1)
p_1
Leader(p3)
Crash(p3)
Crash(p2)
p_2
Leader(p3)
p_3
suspected= null
leader = null

Example 4



suspected$_1$ ={p3}
leader$_1$ = p2
Leader(p3)
Leader(p2)
p$_1$
Crash(p3)
Leader(p3)
suspected$_2$ ={p3}
leader$_2$ = p2
Leader(p2)
suspected$_2$ ={p1, p3}
leader$_2$ = p2
p$_2$
Crash(p3)
Crash(p1)
Leader(p3)
p$_3$
suspected= null
leader = null

Correctness

What if the Failure detector is not perfect?

Eventual leader election (Ω)

Module 2.9: Interface and properties of the eventual leader detector

Module:

Name: EventualLeaderDetector, **instance** Ω.

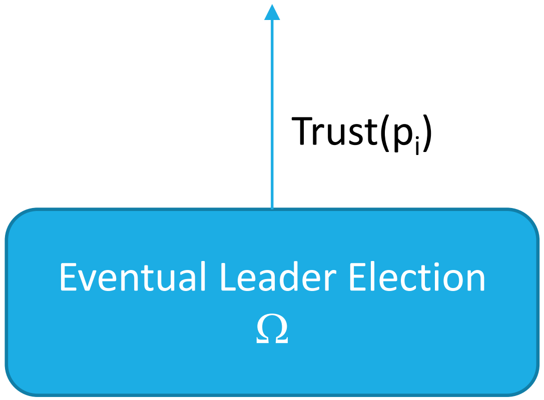
Events:

Indication: $\langle\ \Omega,\ \mathit{Trust} \mid p\ \rangle$: Indicates that process p is trusted to be leader.

Properties:

ELD1: *Eventual accuracy:* There is a time after which every correct process trusts some correct process.

ELD2: *Eventual agreement:* There is a time after which no two correct processes trust different correct processes.



Observation on Ω

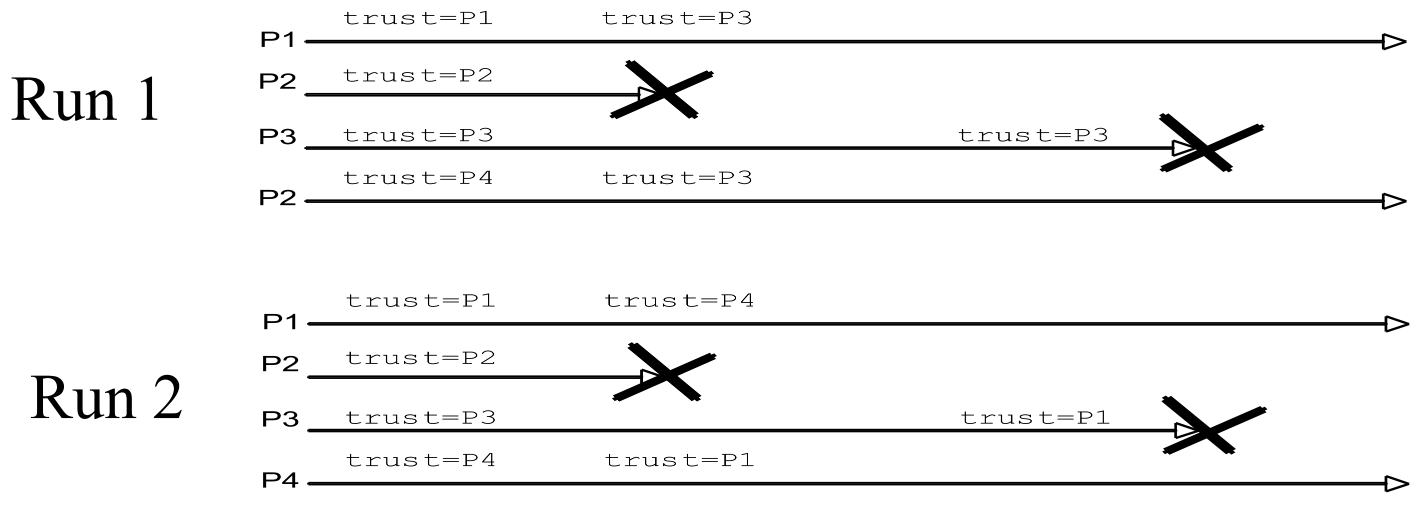
Ω ensures that *eventually* correct processes will elect the same correct process as their leader

Ω does not guarantee that

- Leaders change in an arbitrary manner and for an arbitrary period of time
- many leaders might be elected during the same period of time without having crashed

Once a unique leader is determined, and does not change again, we say that the leader has *stabilized*

Study of Properties



	Run 1	Run 2
Eventual Accuracy	Not verified	Verified
Eventual Agreement	Verified	Not verified

Eventual leader election (Ω)

Using Crash-stop process abstraction

- Obtained directly by <>P by using a deterministic rule on processes that are not suspected by <>P
- trust the process with the highest identifier among all processes that are not suspected by <>P

Assume the existence of a correct process (otherwise Ω cannot be built)

Ω Implementation

Algorithm 2.8: Monarchical Eventual Leader Detection

Implements:
EventualLeaderDetector, **instance** Ω.

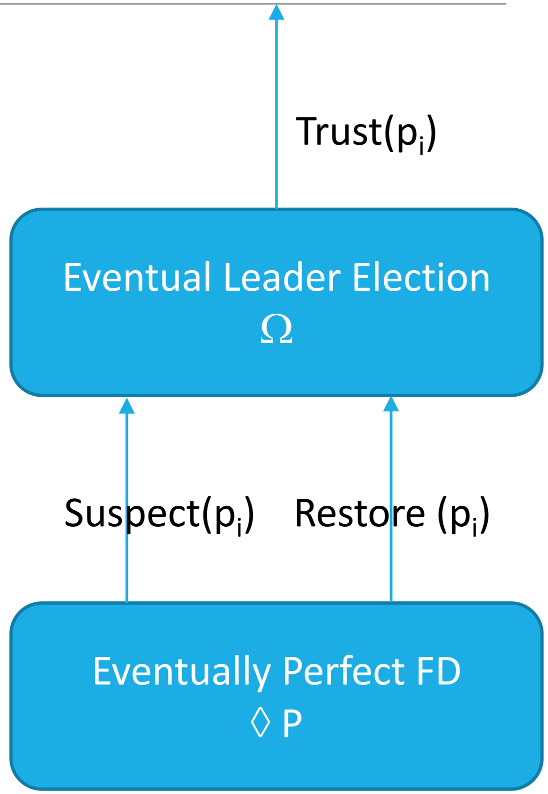
Uses:
EventuallyPerfectFailureDetector, **instance** $\Diamond\mathcal{P}$.

upon event $\langle\ \Omega, \mathit{Init}\ \rangle$ **do**
$\quad$ *suspected* $:= \emptyset$;
$\quad$ *leader* $:= \perp$;

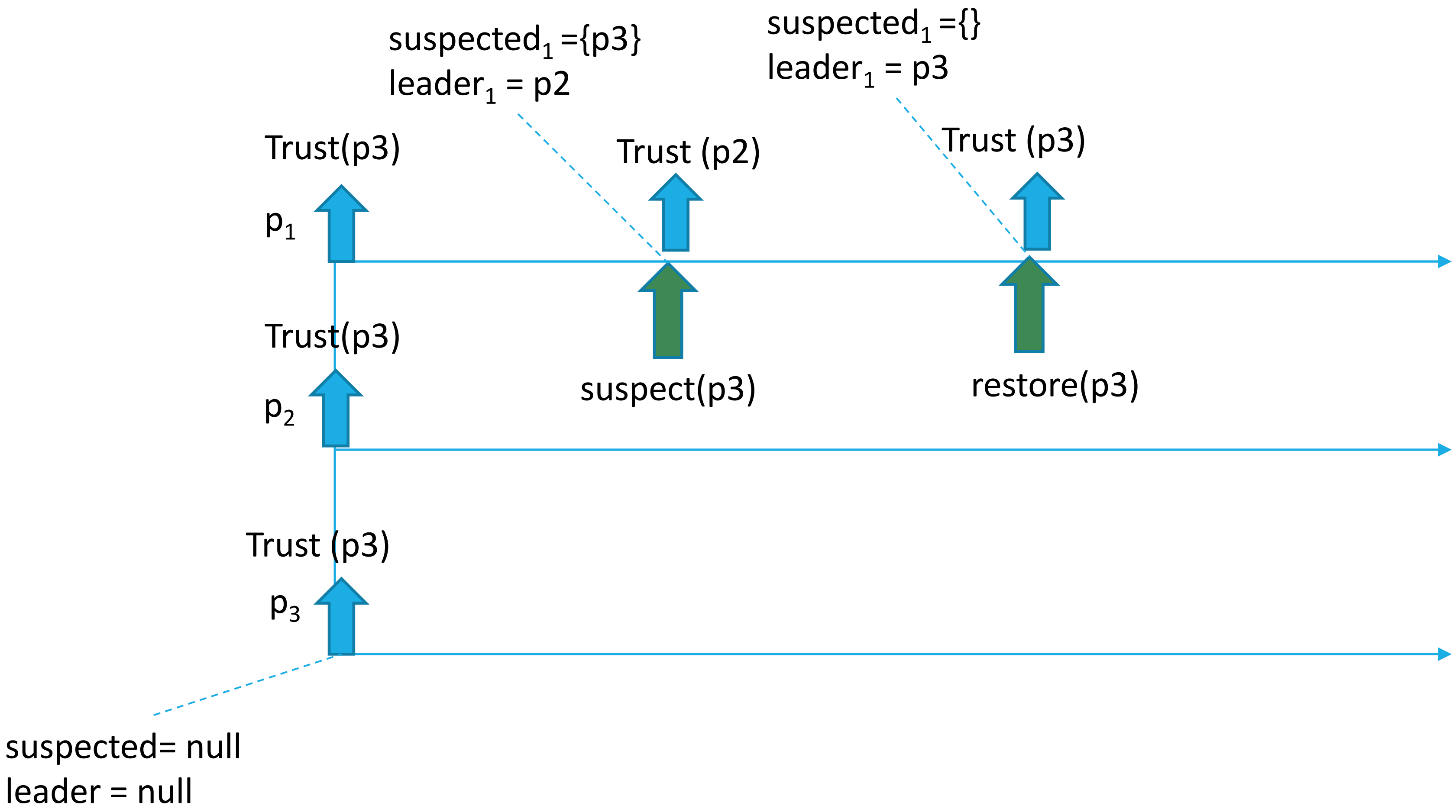
upon event $\langle\ \Diamond\mathcal{P}, \mathit{Suspect} \mid p\ \rangle$ **do**
$\quad$ *suspected* $:=$ *suspected* $\cup \{p\}$;

upon event $\langle\ \Diamond\mathcal{P}, \mathit{Restore} \mid p\ \rangle$ **do**
$\quad$ *suspected* $:=$ *suspected* $\setminus \{p\}$;

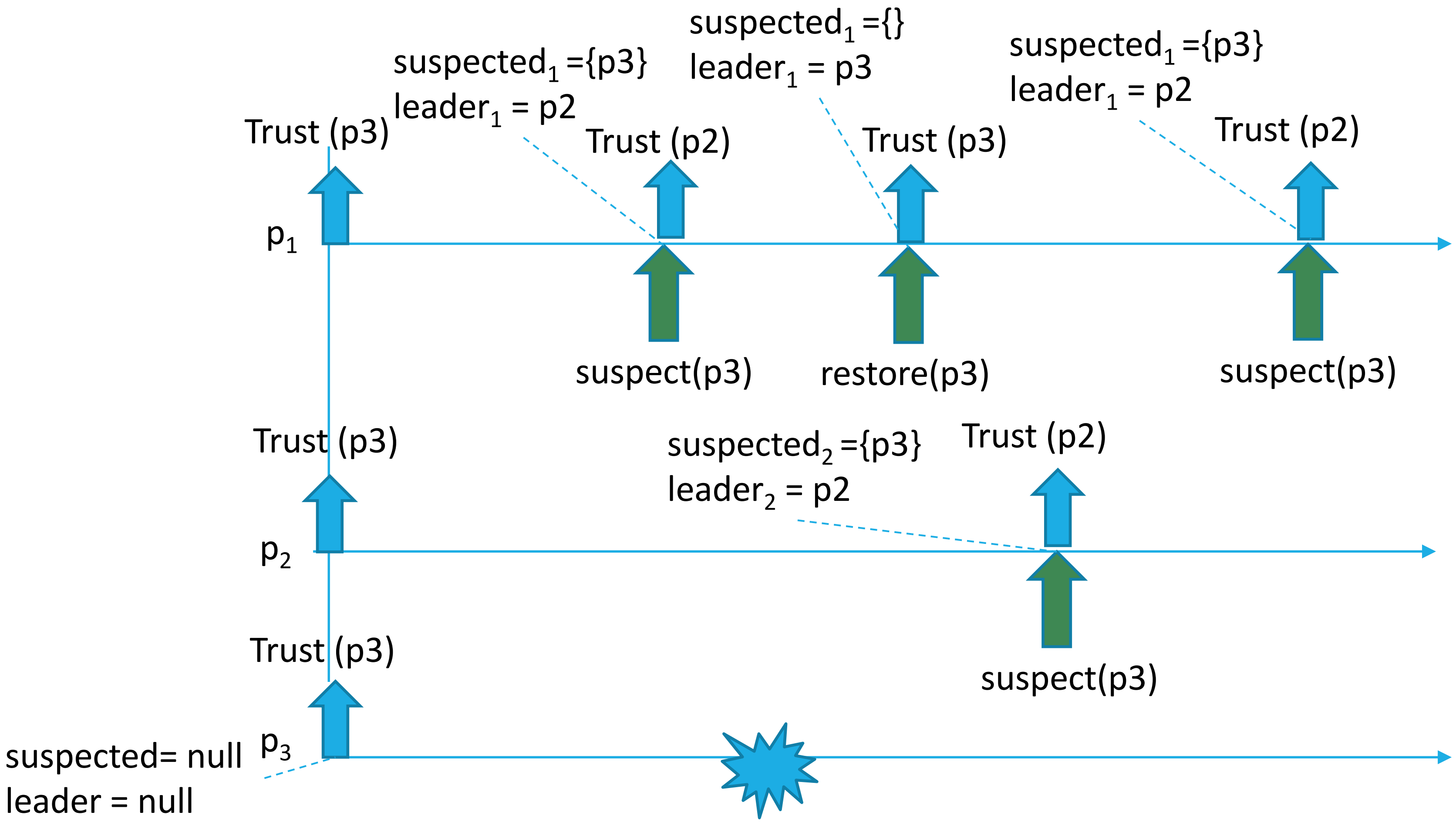
upon $\mathit{leader} \neq \mathrm{maxrank}(\Pi \setminus \mathit{suspected})$ **do**
$\quad$ $\mathit{leader} := \mathrm{maxrank}(\Pi \setminus \mathit{suspected})$;
$\quad$ **trigger** $\langle\ \Omega, \mathit{Trust} \mid \mathit{leader}\ \rangle$;



Example


suspected$_1$ ={p3}
leader$_1$ = p2
suspected$_1$ ={}
leader$_1$ = p3
Trust(p3)
Trust (p2)
Trust (p3)
p$_1$
Trust(p3)
suspect(p3)
restore(p3)
p$_2$
Trust (p3)
p$_3$
suspected= null
leader = null

Example 2



suspected$_1$ ={p3}
leader$_1$ = p2
suspected$_1$ ={}
leader$_1$ = p3
suspected$_1$ ={p3}
leader$_1$ = p2
Trust (p3)
Trust (p2)
Trust (p3)
Trust (p2)
p$_1$
suspect(p3)
restore(p3)
suspect(p3)
Trust (p3)
suspected$_2$ ={p3}
leader$_2$ = p2
Trust (p2)
p$_2$
suspect(p3)
Trust (p3)
p$_3$
suspected= null
leader = null

Eventual leader election (Ω)

System model

- Crash-Recovery
- Partial synchrony

Under this assumption, a correct process means:

1. A process that does not crash or
2. A process that crashes, eventually recovers and never crashes again

Ω With crash-recovery, fair lossy links and timeouts

Algorithm 2.9: Elect Lower Epoch

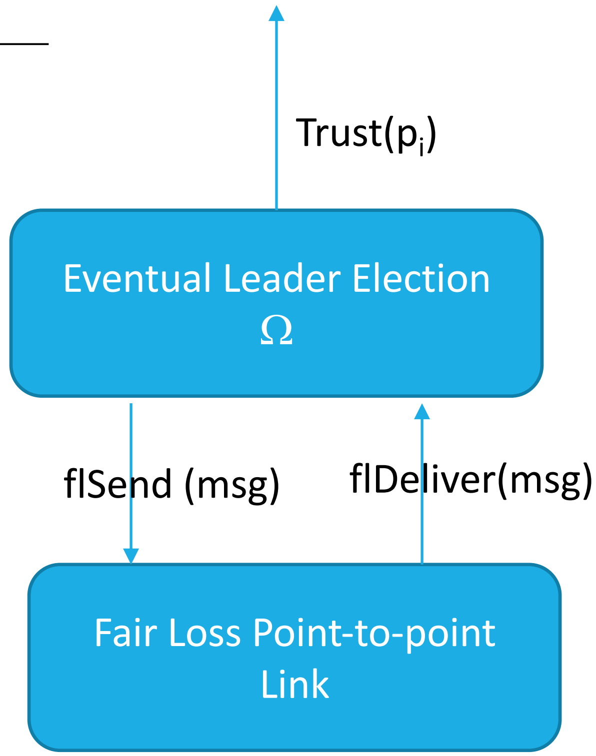
Implements:
  EventualLeaderDetector, **instance** Ω.

Uses:
  FairLossPointToPointLinks, **instance** *fll*.

upon event $\langle$ Ω, *Init* $\rangle$ **do**
  epoch := 0;
  store(*epoch*);
  candidates := $\emptyset$;
  trigger $\langle$ Ω, *Recovery* $\rangle$;

keeps track of how many times the process crashed and recovered

upon event $\langle$ Ω, *Recovery* $\rangle$ **do**
  leader := maxrank(Π);
  trigger $\langle$ Ω, *Trust* | *leader* $\rangle$;
  delay := Δ;
  retrieve(*epoch*);
  epoch := *epoch* + 1;
  store(*epoch*);
  forall $p \in \Pi$ **do**
    trigger $\langle$ *fll*, *Send* | p, [HEARTBEAT, *epoch*] $\rangle$;
  candidates := $\emptyset$;
  starttimer(*delay*);



Ω With crash-recovery, fair lossy links and timeouts

Algorithm 2.9: Elect Lower Epoch

Implements:

EventualLeaderDetector, **instance** Ω.

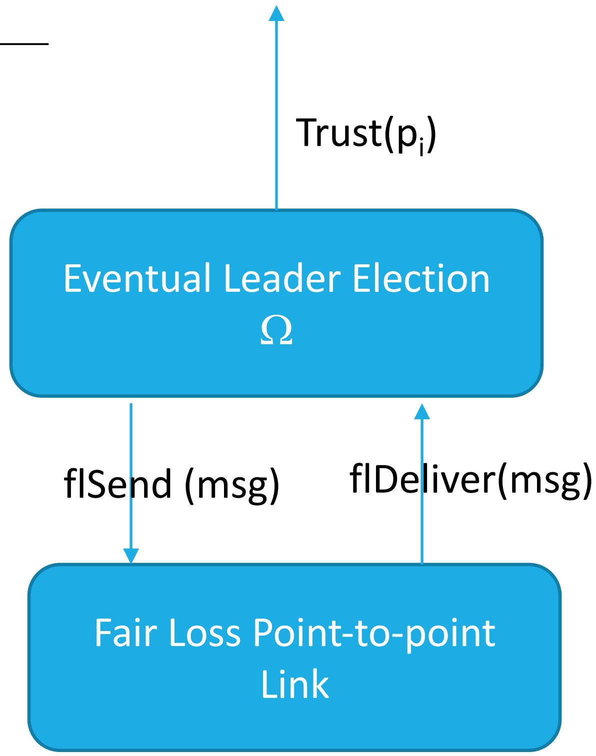
Uses:

FairLossPointToPointLinks, **instance** *fll*.

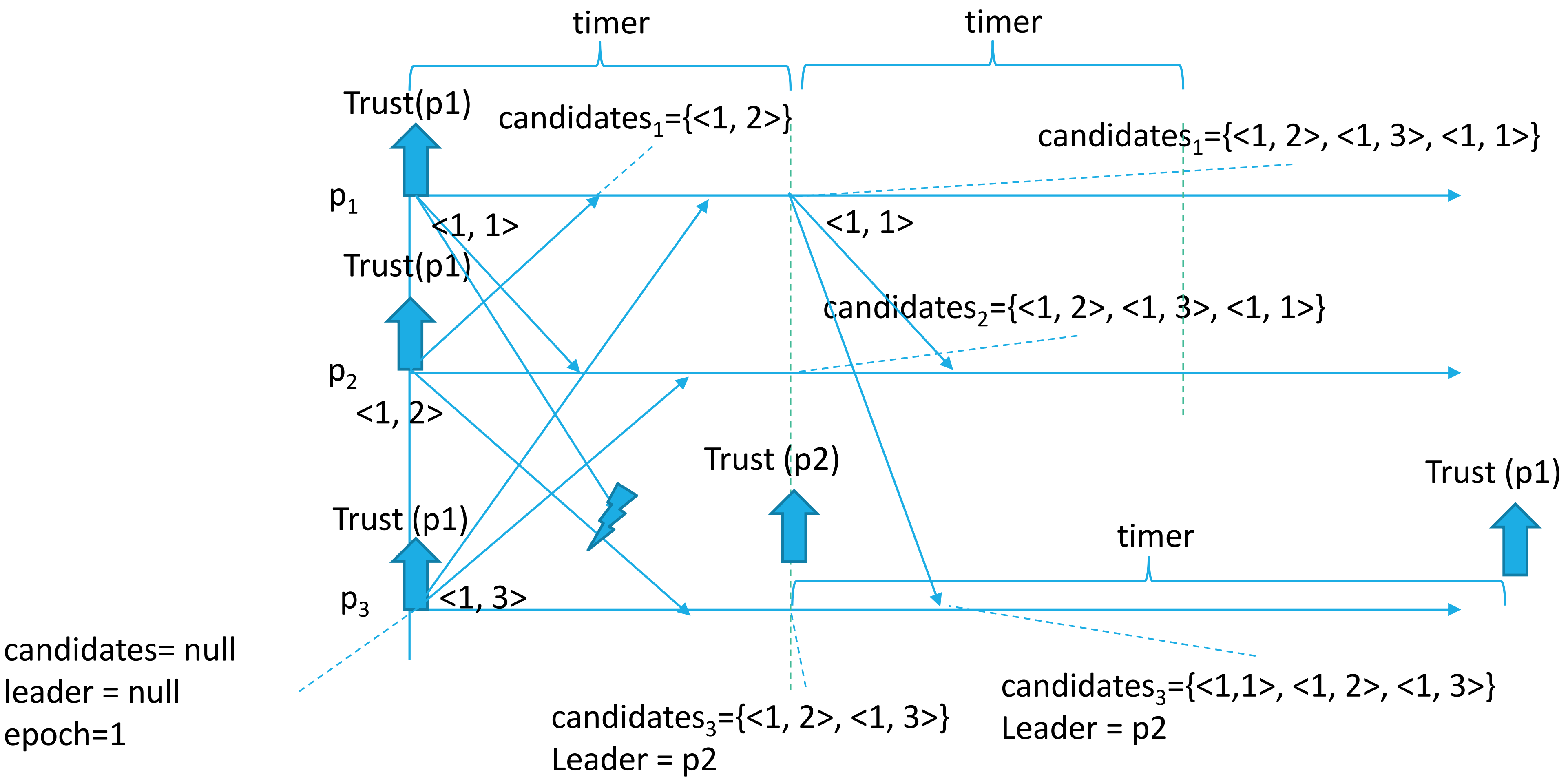
upon event ⟨ *Timeout* ⟩ **do**
  newleader := *select*(*candidates*);
  if *newleader* $\neq$ *leader* **then**
    delay := *delay* + Δ;
    leader := *newleader*;
    trigger ⟨ Ω, *Trust* | *leader* ⟩;
  forall $p \in \Pi$ **do**
    trigger ⟨ *fll*, *Send* | p, [HEARTBEAT, *epoch*] ⟩;
  candidates := $\emptyset$;
  starttimer(*delay*);

upon event ⟨ *fll*, *Deliver* | q, [HEARTBEAT, *ep*] ⟩ **do**
  if exists $(s, e) \in$ *candidates* such that $s = q \wedge e < ep$ **then**
    candidates := *candidates* $\setminus \{(q, e)\}$;
  candidates := *candidates* $\cup$ (q, ep);

deterministic function returning one process among all candidates (i.e., process with the lowest epoch number and among the ones with the same epoch number the one with the lowest identifier)



Example


timer
timer
Trust(p1)
candidates$_1$={<1, 2>}
candidates$_1$={<1, 2>, <1, 3>, <1, 1>}
p$_1$
<1, 1>
<1, 1>
Trust(p1)
candidates$_2$={<1, 2>, <1, 3>, <1, 1>}
p$_2$
<1, 2>
Trust (p2)
Trust (p1)
Trust (p1)
timer
p$_3$
<1, 3>
candidates= null
leader = null
epoch=1
candidates$_3$={<1,1>, <1, 2>, <1, 3>}
Leader = p2
candidates$_3$={<1, 2>, <1, 3>}
Leader = p2

References

C. Cachin, R. Guerraoui and L. Rodrigues. Introduction to Reliable and Secure Distributed Programming, Springer, 2011

- Chapter 2 – from Section 2.6.1 to Section 2.6.5