Distributed Systems

CS425/ECE428

Feb 12 2021

Instructor: Radhika Mittal

Some revision while we wait

- For a process p_i , where events $e_i^0, e_i^1, \ldots$ occur:

  history(p_i) = $h_i = <e_i^0, e_i^1, \ldots >$

  prefix history(p_i^k) = $h_i^k = <e_i^0, e_i^1, \ldots, e_i^k >$

  s_i^k : p_i's state immediately after k^{th} event.

- For a set of processes $<p_1, p_2, p_3, \ldots, p_n>$:

  global history: $H = \cup_i (h_i)$

  a cut $C \subseteq H = h_1^{c_1} \cup h_2^{c_2} \cup \ldots \cup h_n^{c_n}$

  the frontier of $C = \{e_i^{c_i}, i = 1,2, \ldots n\}$

  global state S that corresponds to cut $C = \cup_i (s_i^{c_i})$

- A cut **C** is consistent if and only if $\forall$**e** $\in$ **C** (if **f** $\rightarrow$ **e** then **f** $\in$ **C**)
  - A global state **S** is consistent if and only if it corresponds to a consistent cut.

Logistics Related

- MP0 is due today 11:59pm.
- No class next Wednesday (Feb 17th) – non-instructional day.
  - I have moved my office hours to Thursday 10-11am for next week.
- HW1 is due on Thursday (Feb 18th) at 11:59pm.

Today's agenda

- **Global State**
  - Chapter 14.5
  - Goal: reason about how to capture the state across all processes of a distributed system without requiring time synchronization.

Recap: How to capture global state?

- State of each process (and each channel) in the system *at a given instant of time.*
  - Difficult to capture -- requires precisely synchronized time.

- Relax the problem
  - For a system with n processes <p_1, p_2, p_3, …., p_n>, capture the state of the system after the c_i th event at process p_i.
    - State corresponding to the cut defined by frontier events {$e_i^{c_i}$, for i = 1,2, … n}.
  - We want the state to be consistent.
    - Must correspond to a consistent cut.
      - If an event e belongs to the cut, all events that "happened before" e must also belong to the cut.

Recap: Chandy-Lamport Algorithm

- *Goal: Record consistent state by identifying a consistent cut.*
- *System model and assumptions:*
  - System of **n** processes: <p_1, p_2, p_3, …., p_n>.
  - There are two uni-directional communication channels between each ordered process pair : p_j to p_i and p_i to p_j.
  - Communication channels are FIFO-ordered (first in first out).
  - All messages arrive intact, and are not duplicated.
  - No failures: neither channel nor processes fail.
- *Requirements:*
  - Snapshot should not interfere with normal application actions, and it should not require application to stop sending messages.
  - Any process may initiate algorithm.

Chandy-Lamport Algorithm Intuition

- First, initiator p_i:
  - records its own state.
  - creates a special **marker** message.
  - sends the **marker** to all other process.
  - start recording messages received on other channels.
    - until a marker is received on a channel.
- When a process receives a **marker**.
  - If marker is received for the first time.
    - records its own state.
    - sends marker on all other channels.
    - start recording messages received on other channels.
      - until a marker is received on a channel.

Chandy-Lamport Algorithm

- First, initiator **p_i**:
  - records its own state.
  - creates a special **marker** message.
  - for *j=1 to n* except *i*
    - **p_i** sends a **marker** message on outgoing channel **c_{ij}**
    - starts recording the incoming messages on each of the incoming channels at **p_i** : **c_{ji}** (for *j=1 to n* except *i*).

Chandy-Lamport Algorithm

Whenever a process p_i receives a **marker** message from p_k on incoming channel c_{ki}

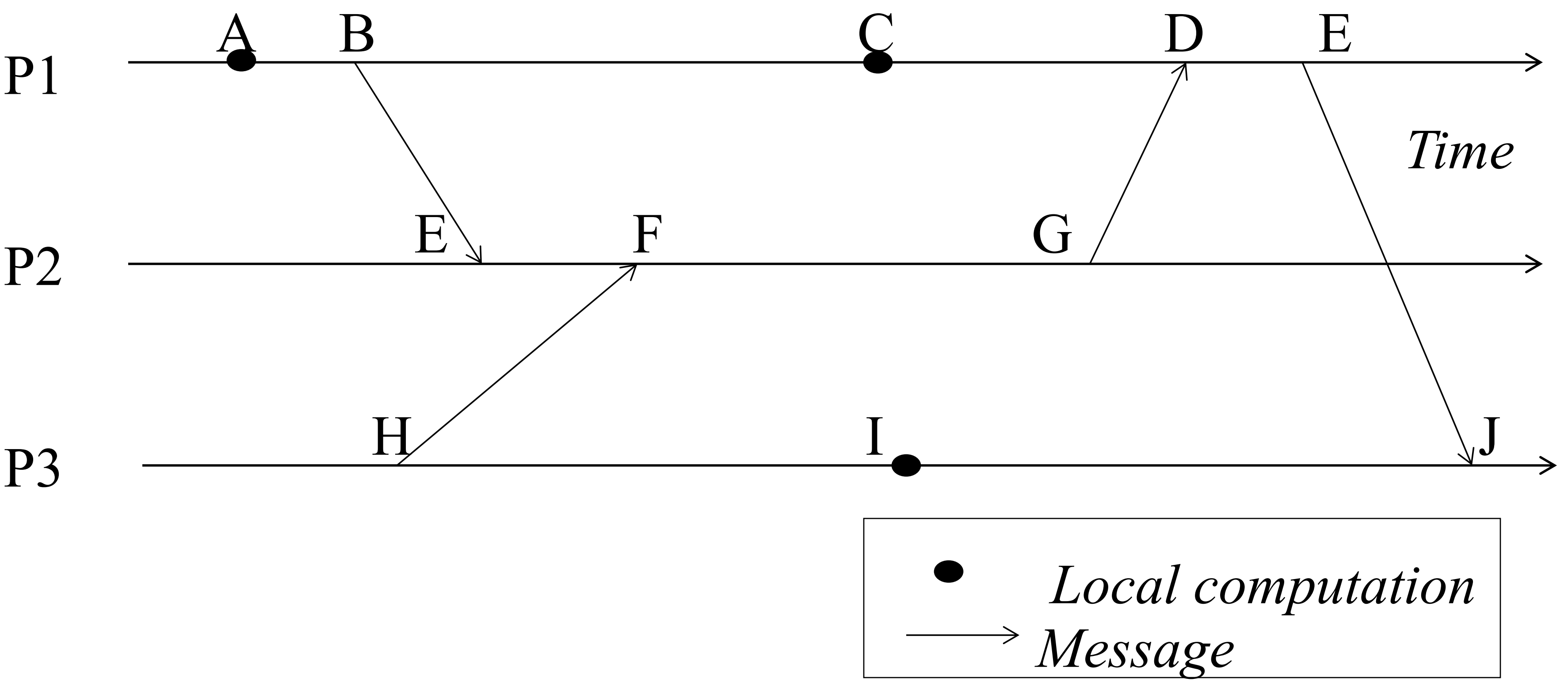
- if (this is the first **marker** p_i is seeing)
  - p_i records its own state first
  - marks the state of channel c_{ki} as "empty"
  - for j=1 to n except i
    - p_i sends out a **marker** message on outgoing channel c_{ij}
  - starts recording the incoming messages on each of the incoming channels at p_i : c_{ji} (for *j=1 to n* except *i and k*).
- else // already seen a **marker** message
  - mark the state of channel c_{ki} as all the messages that have arrived on it since recording was turned on for c_{ki}

Chandy-Lamport Algorithm

The algorithm terminates when

- All processes have received a **marker**
  - To record their own state
- All processes have received a **marker** on all the (n-1) incoming channels
  - To record the state of all channels

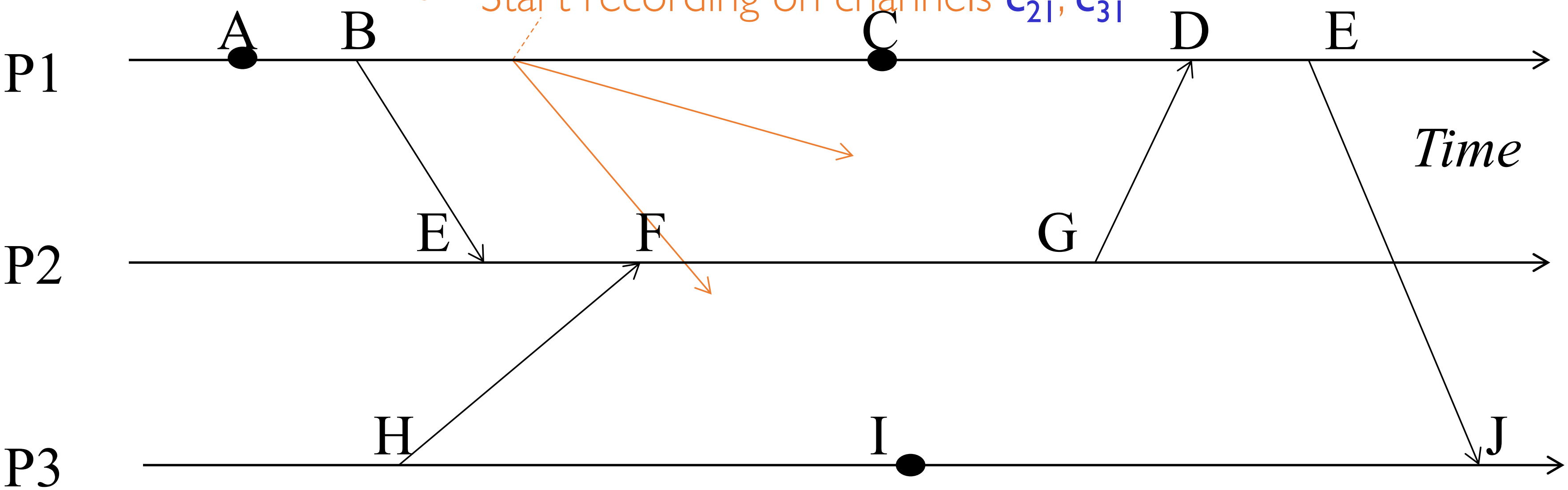
Example



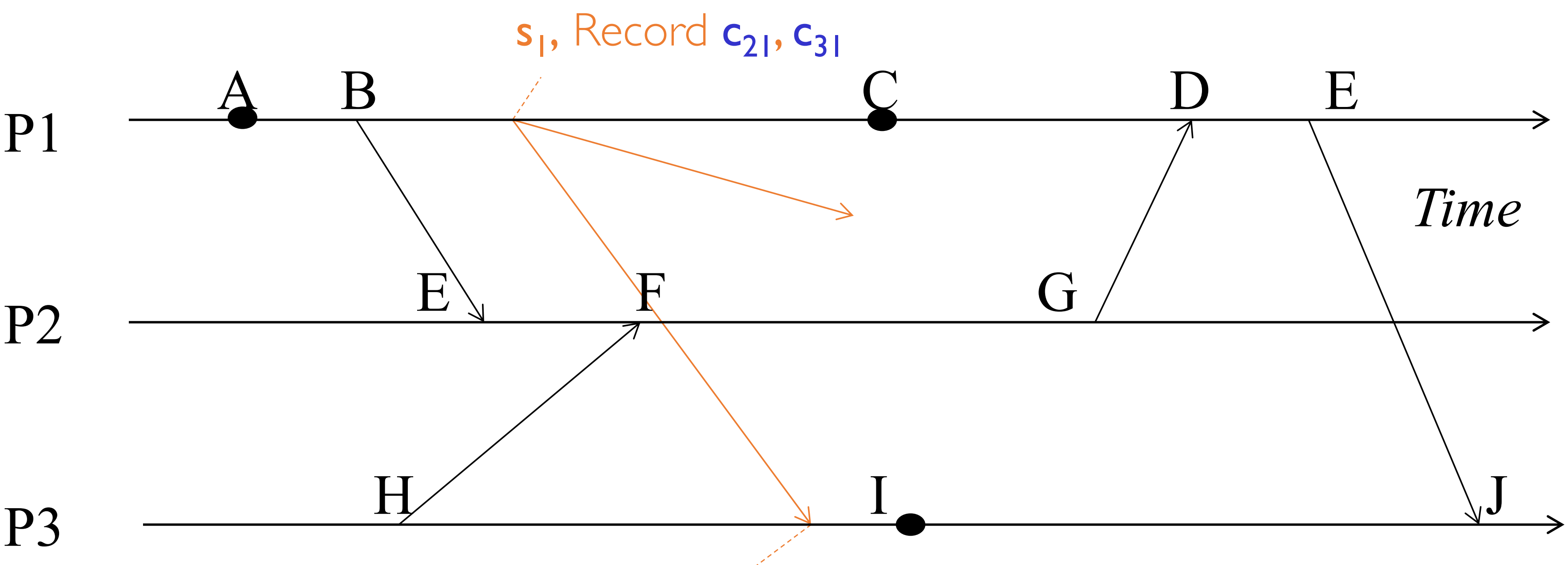
Example

p_1 is initiator:

- Record local state s_1,
- Send out markers
- Start recording on channels c_{21}, c_{31}

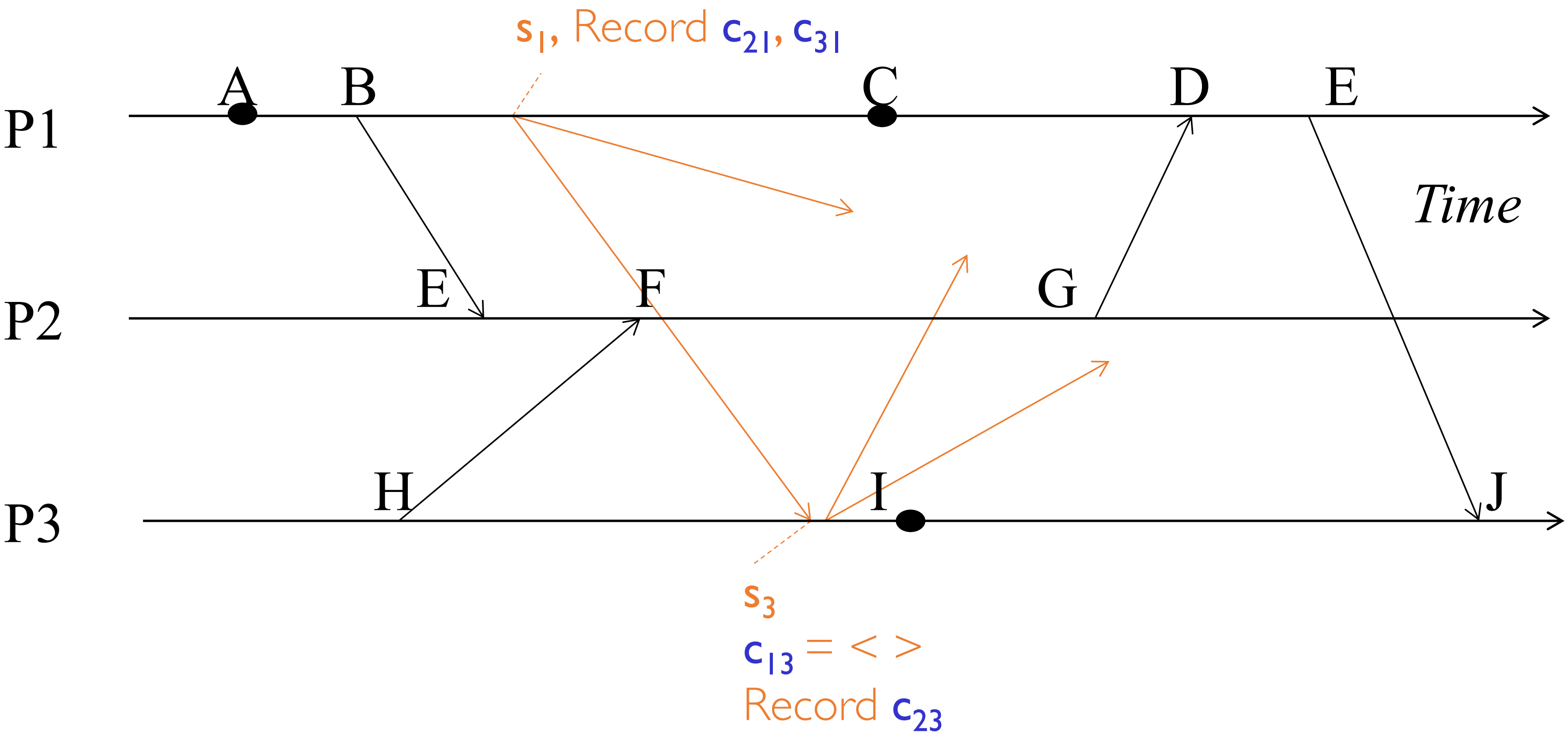


Example

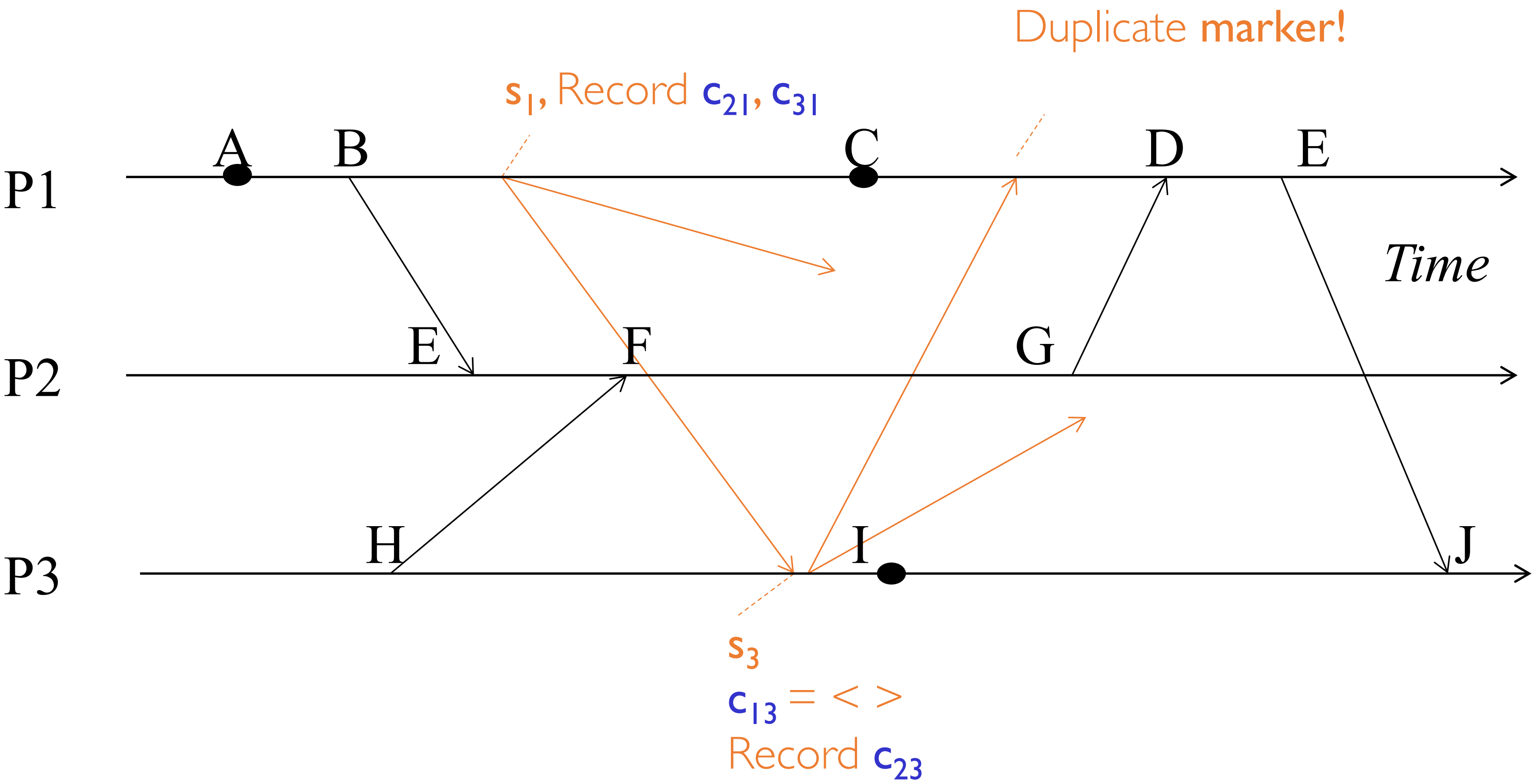


- First **marker**!
- Record own state as $\mathbf{s_3}$
- Mark $\mathbf{c_{13}}$ state as empty
- Start recording on other incoming $\mathbf{c_{23}}$
- Send out markers

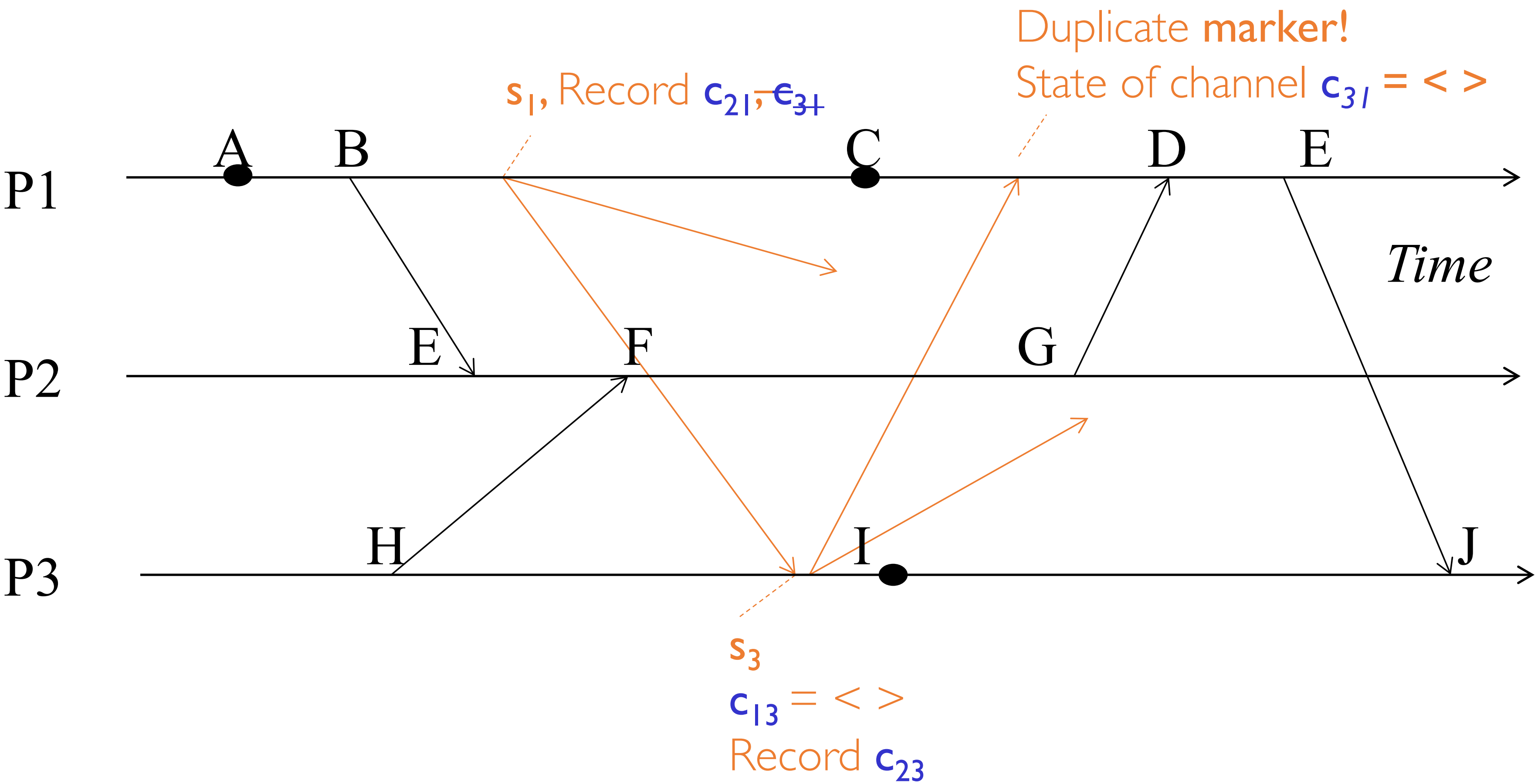
Example



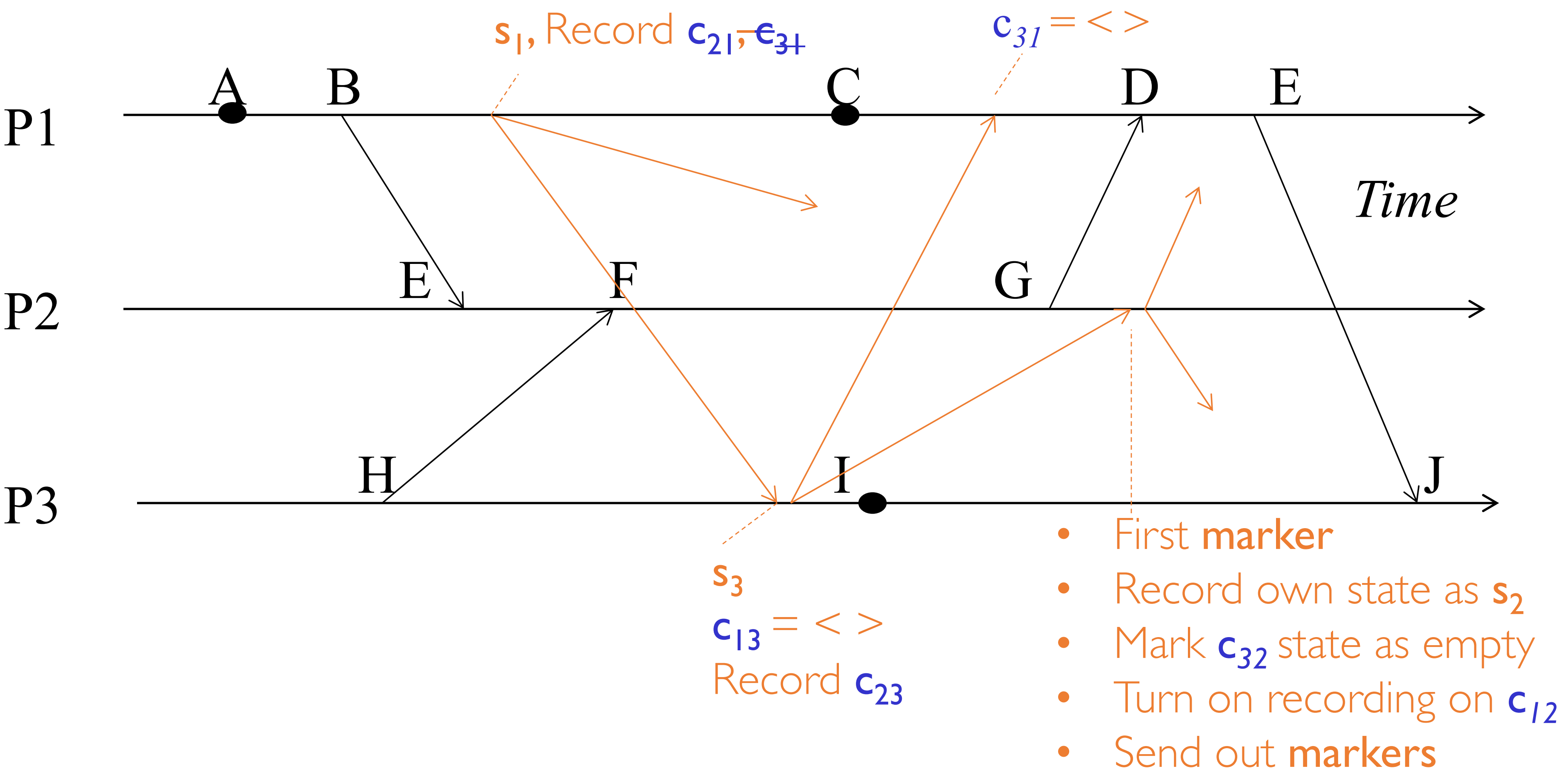
Example


Duplicate marker!
s1, Record c21, c31
A
B
C
D
E
P1
Time
E
F
G
P2
H
I
J
P3
s3
c13 = < >
Record c23

Example

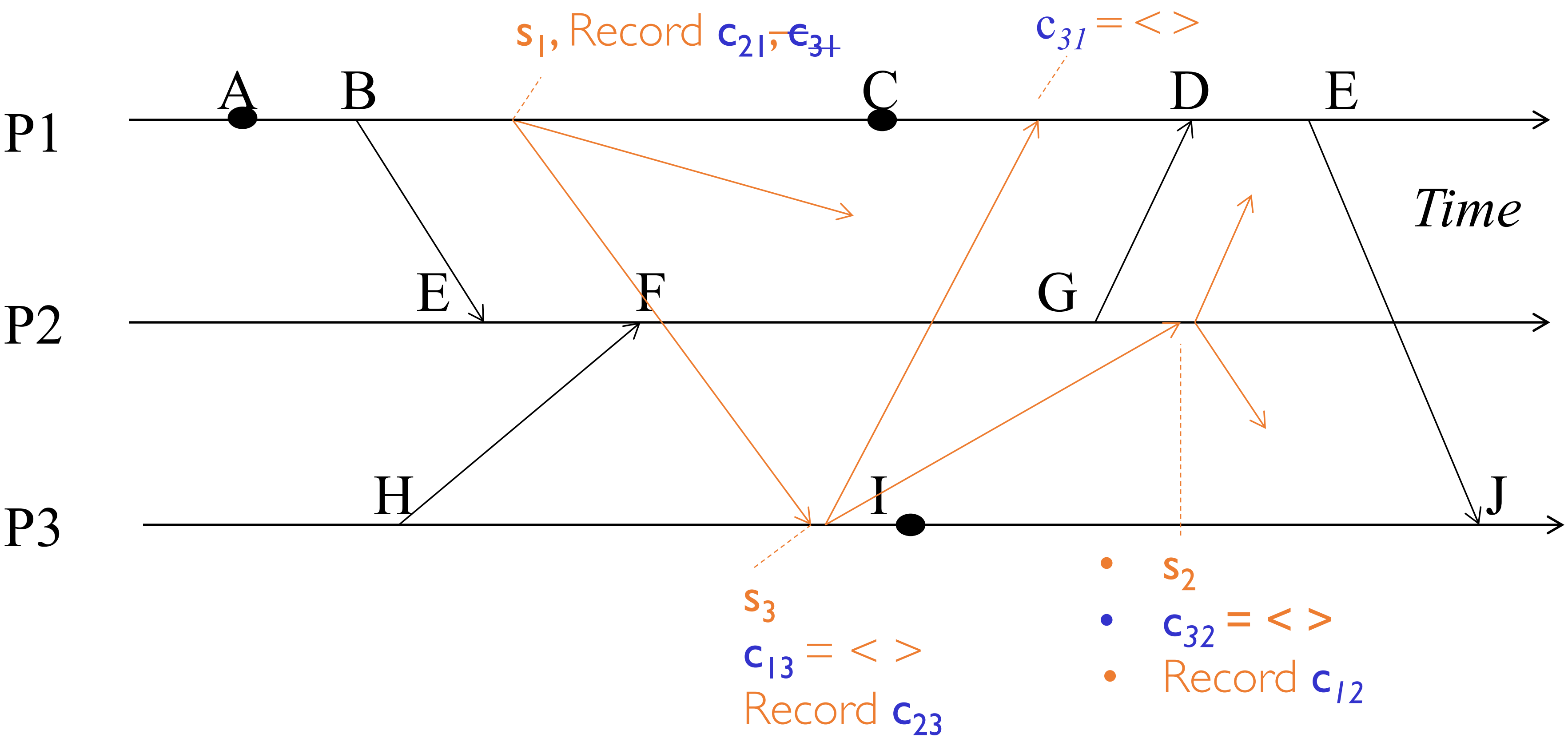

Duplicate marker!
State of channel c31 = < >
s1, Record c21, ~~c31~~
A
B
C
D
E
P1
Time
E
F
G
P2
H
I
J
P3
s3
c13 = < >
Record c23

Example



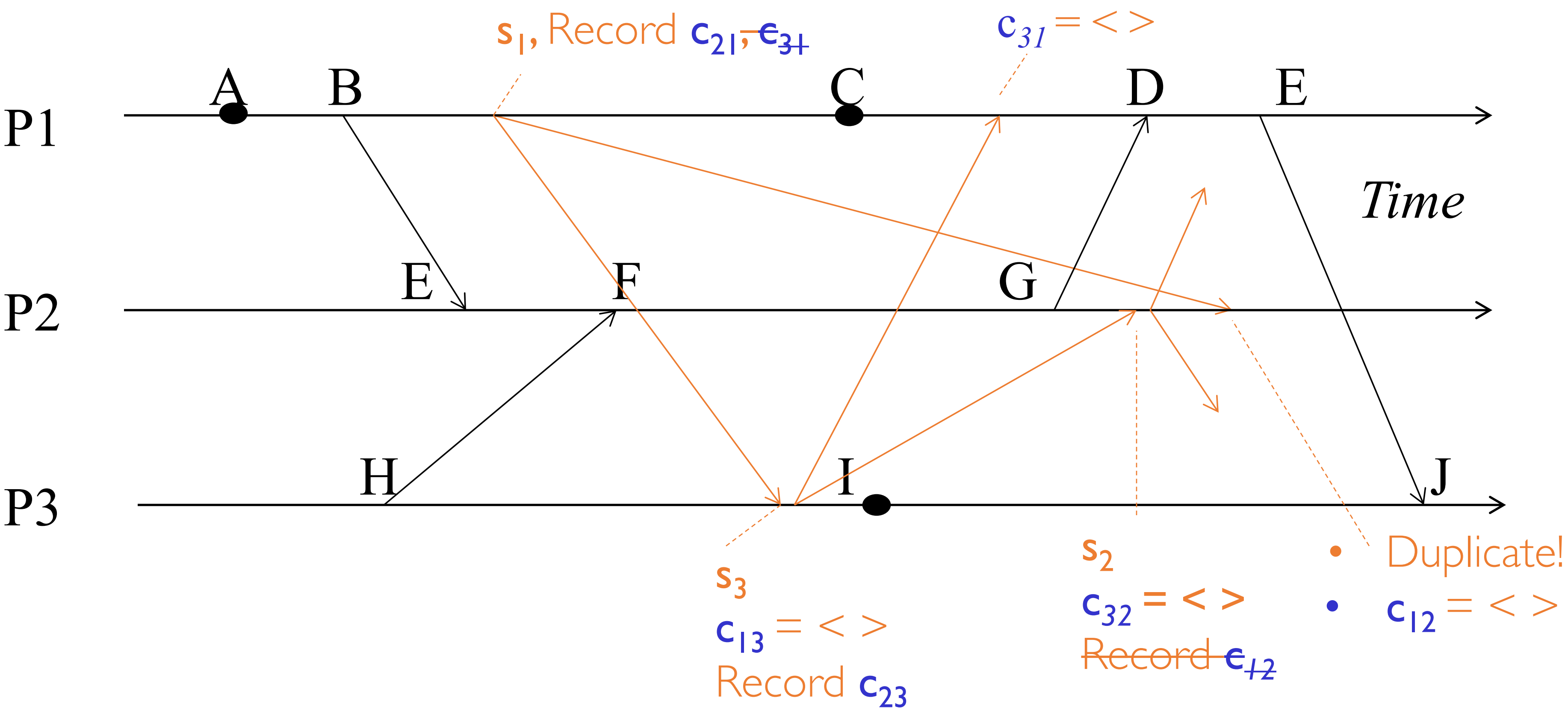
s1, Record c21, ~~c31~~
c31 = < >
A
B
C
D
E
P1
Time
E
F
G
P2
H
I
J
P3
s3
c13 = < >
Record c23
• First marker
• Record own state as s2
• Mark c32 state as empty
• Turn on recording on c12
• Send out markers

Example



s1, Record c21, c31
c31 = < >
A
B
C
D
E
P1
Time
E
F
G
P2
H
I
J
P3
s3
c13 = < >
Record c23
s2
c32 = < >
Record c12

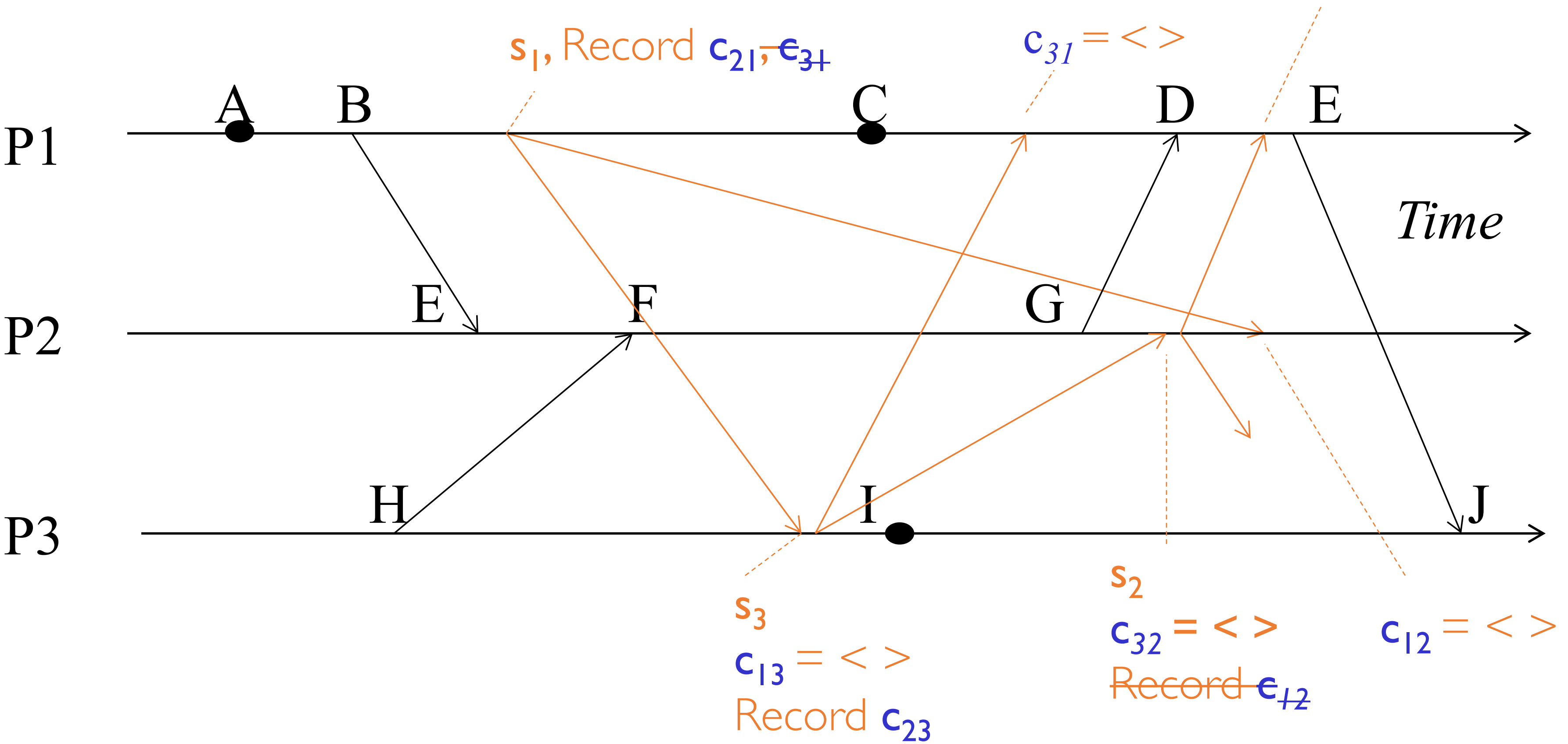
Example



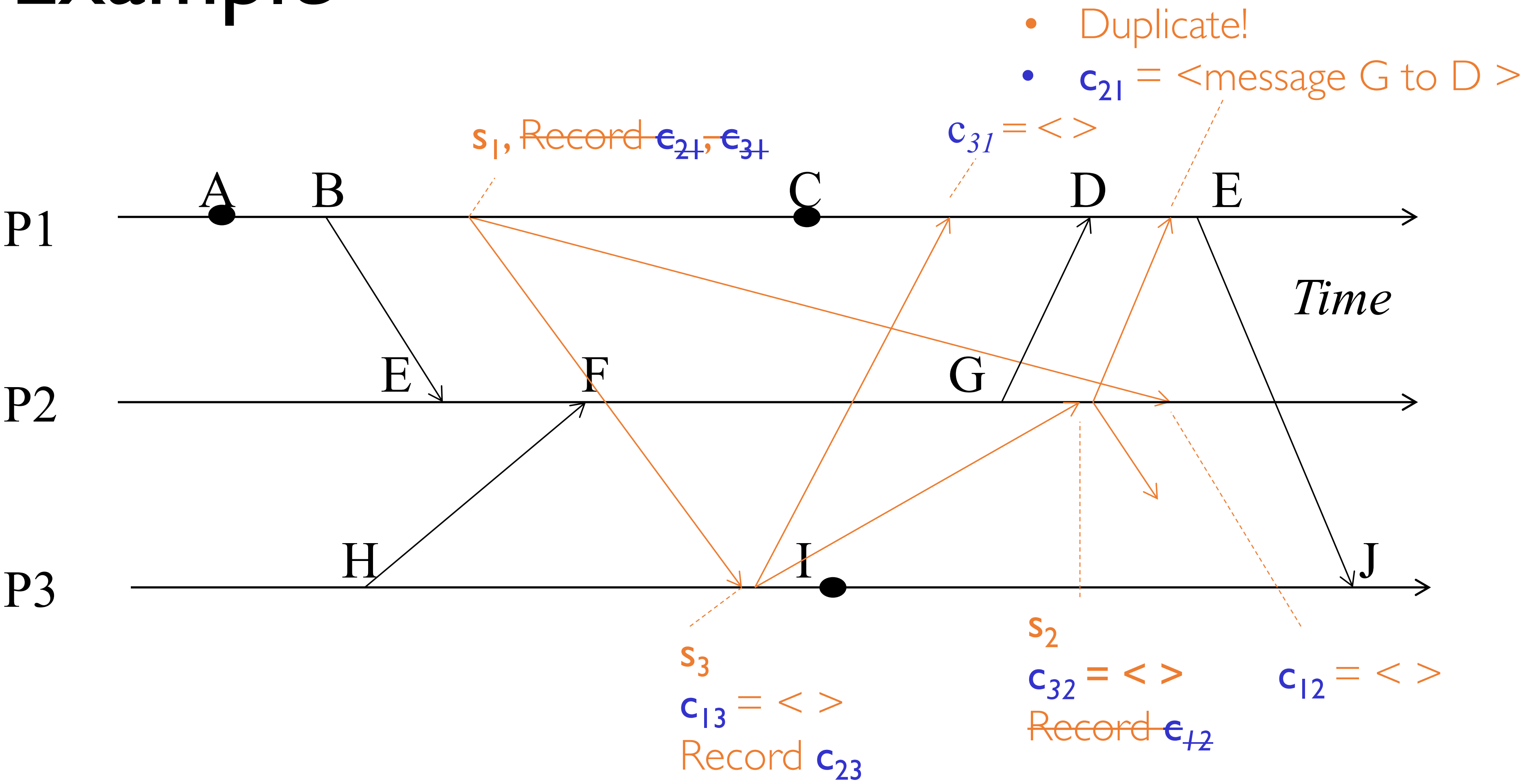
s1, Record c21, ~~c31~~
c31 = < >
P1
P2
P3
A
B
C
D
E
E
F
G
H
I
J
Time
s3
c13 = < >
Record c23
s2
c32 = < >
~~Record c12~~
• Duplicate!
• c12 = < >

Example

- Duplicate!

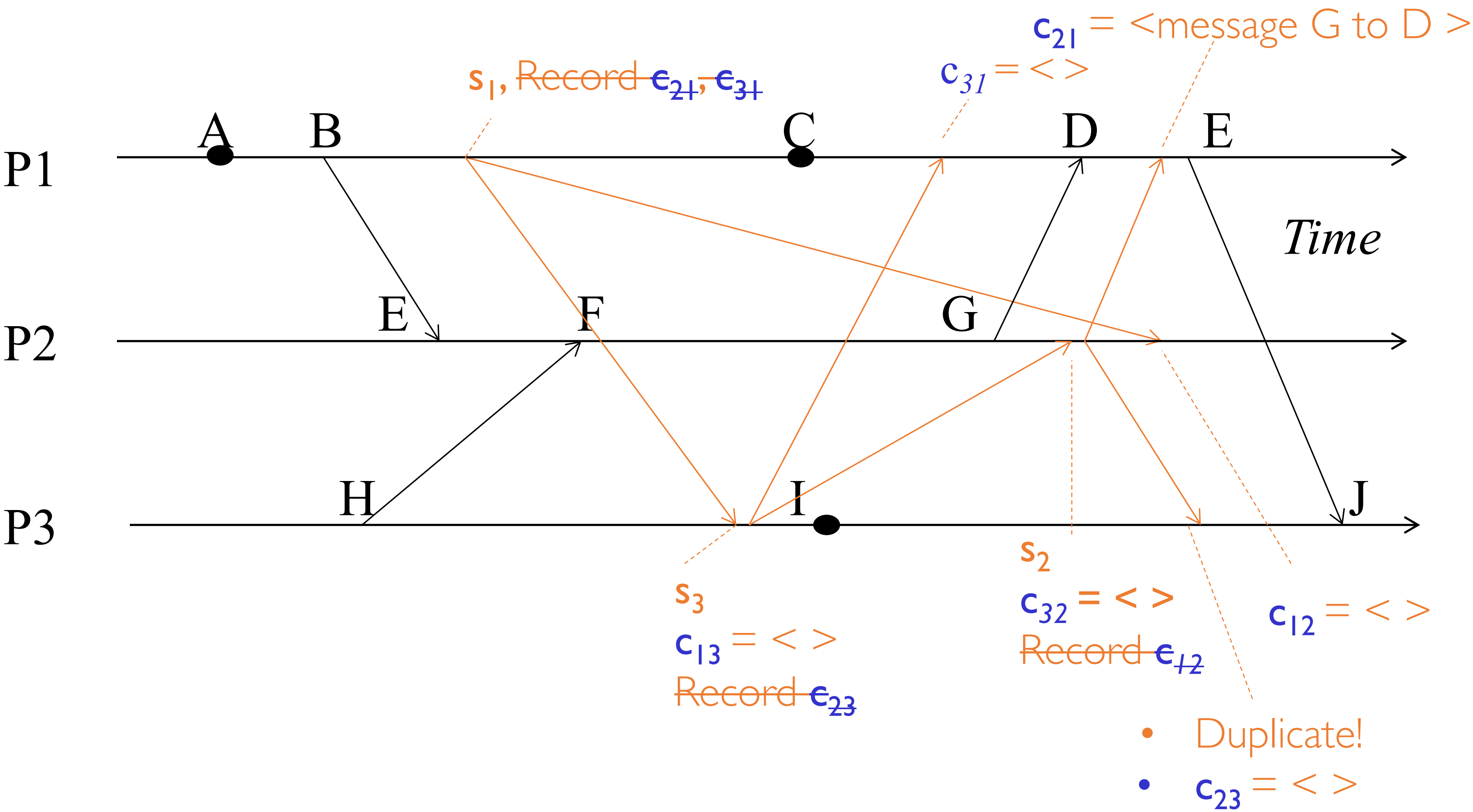


Example



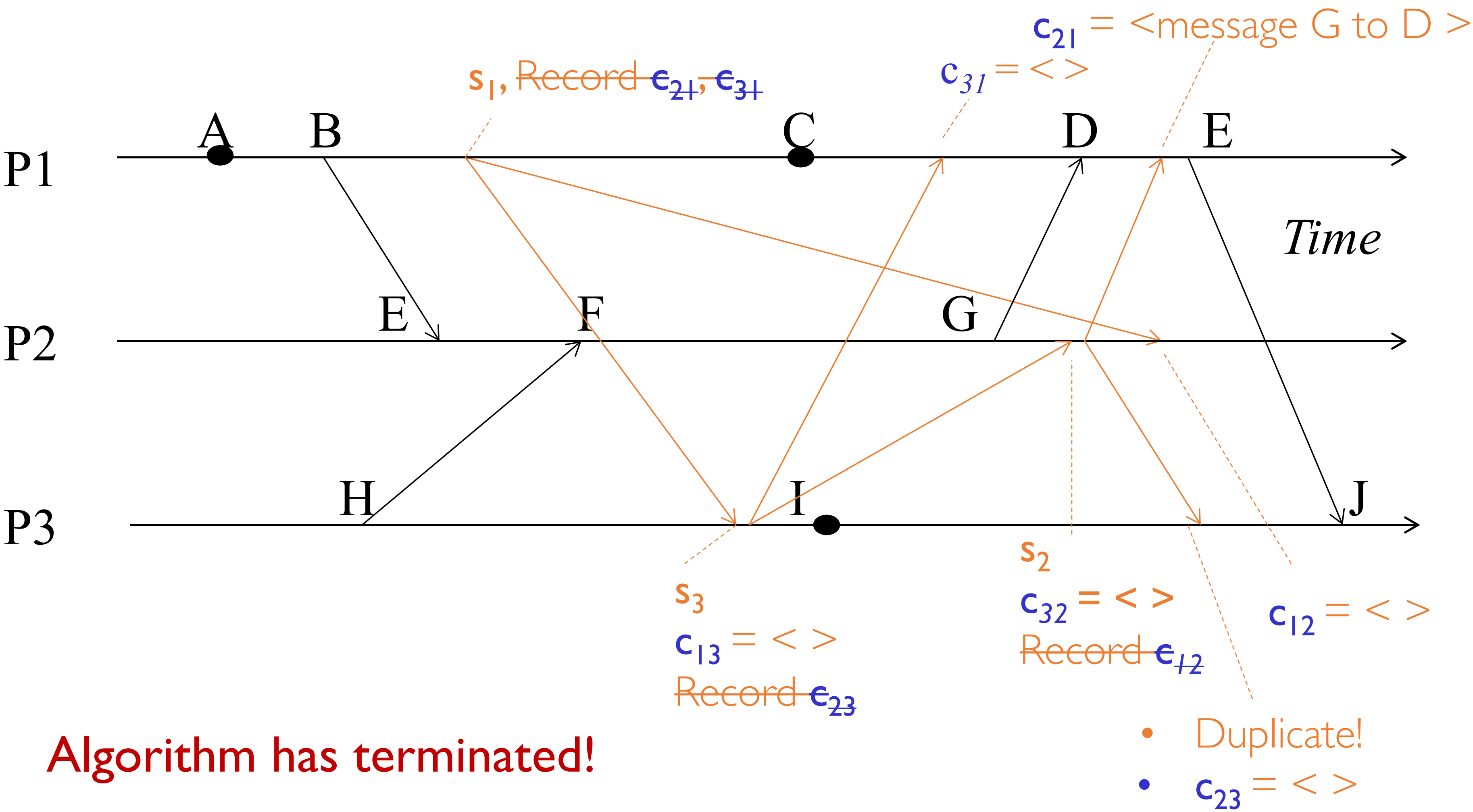
- Duplicate!
- c_{21} = <message G to D >
s_1, ~~Record c_{21}, c_{31}~~
c_{31} = < >
A
B
C
D
E
P1
Time
E
F
G
P2
H
I
J
P3
s_3
c_{13} = < >
Record c_{23}
s_2
c_{32} = < >
~~Record c_{12}~~
c_{12} = < >

Example



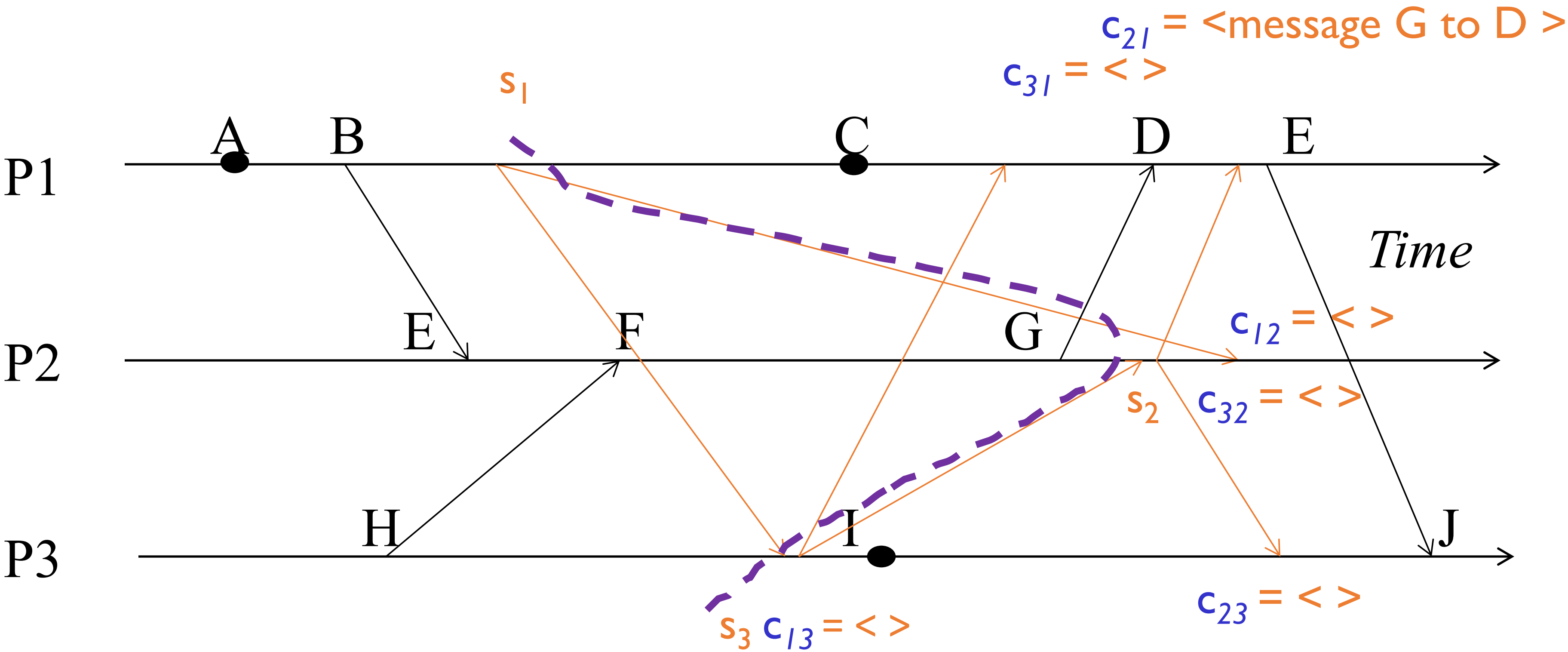
P1
P2
P3
A
B
C
D
E
E
F
G
H
I
J
Time
s1, Record c21, c31
c31 = < >
c21 = <message G to D >
s3
c13 = < >
Record c23
s2
c32 = < >
Record c12
c12 = < >
• Duplicate!
• c23 = < >

Example



Algorithm has terminated!

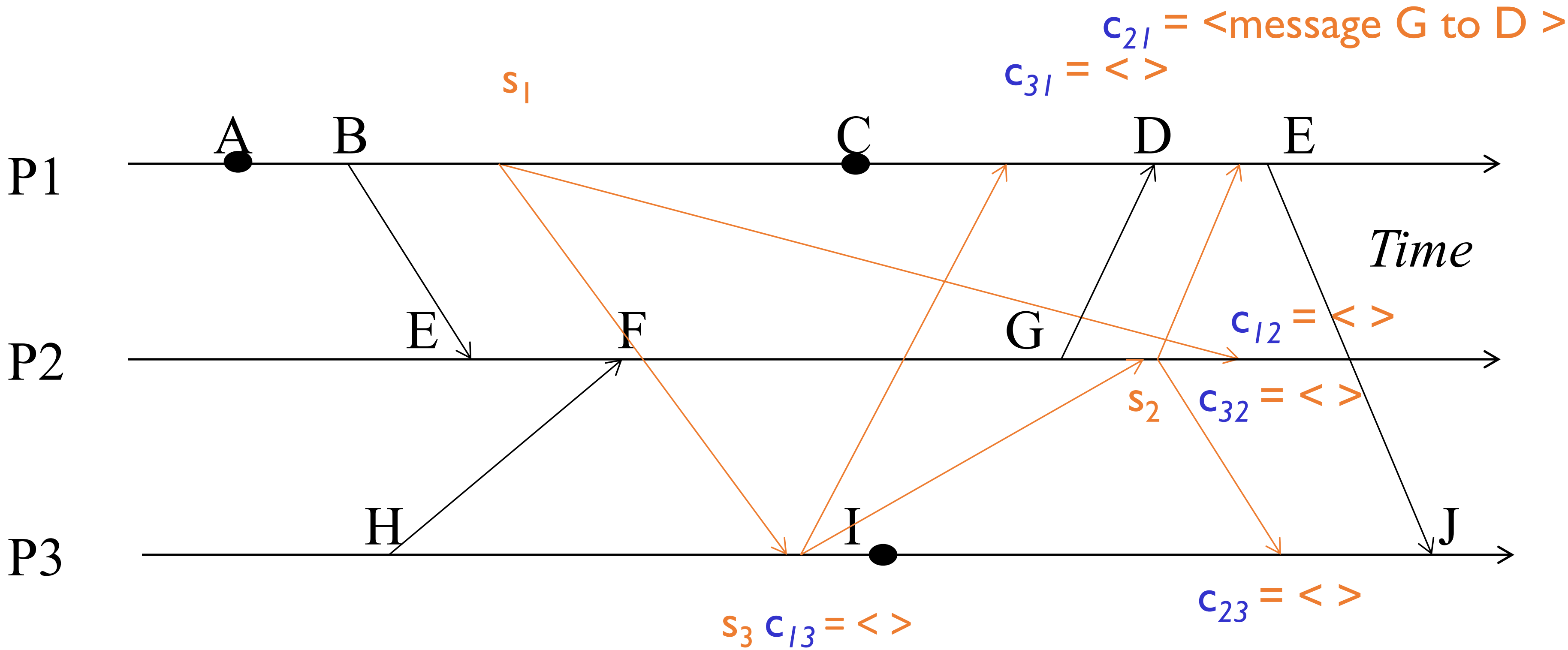
Example



Frontier for the resulting cut:
{B, G, H}

Channel state for the cut:
Only c_{21} has a pending message.

Example



Global snapshots pieces can be collected at a central location.

Chandy-Lamport Algorithm: Properties

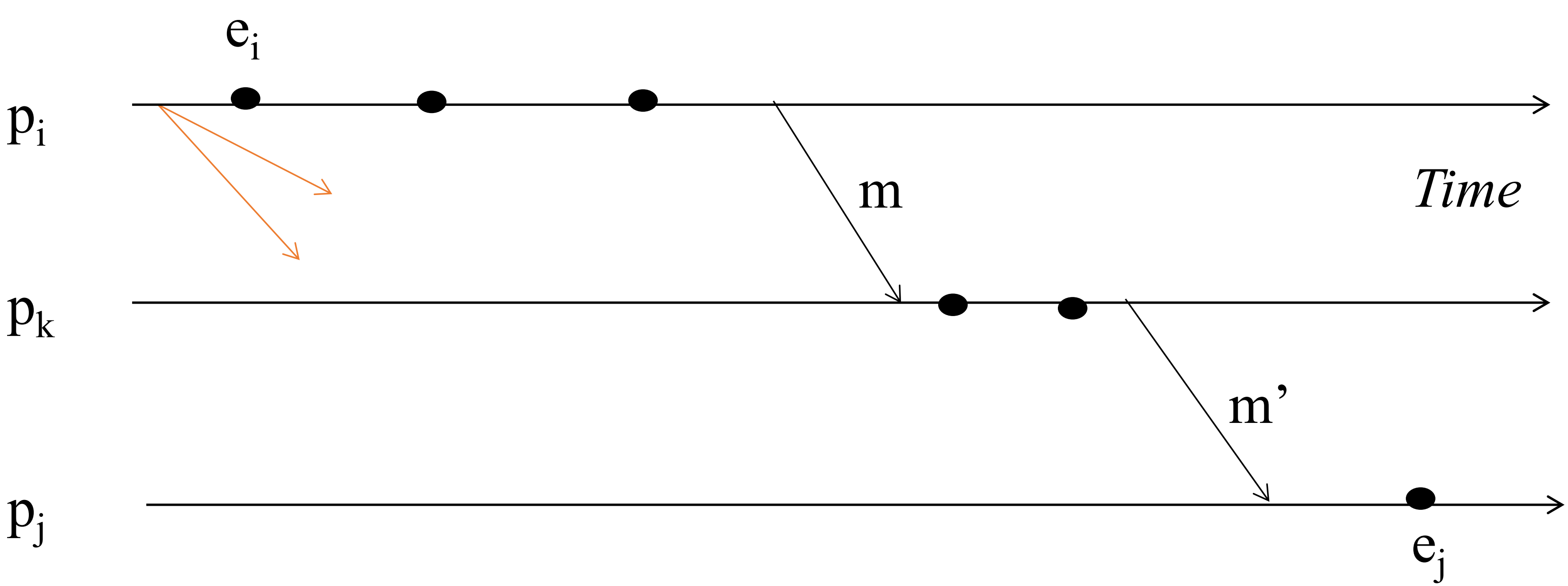
- Any run of the Chandy-Lamport Global Snapshot algorithm creates a consistent cut.
- Let $\mathbf{e}_i$ and $\mathbf{e}_j$ be events occurring at $\mathbf{p}_i$ and $\mathbf{p}_j$, respectively such that
  - $\mathbf{e}_i \rightarrow \mathbf{e}_j$ ($\mathbf{e}_i$ happens before $\mathbf{e}_j$)
- The snapshot algorithm ensures that

  if $\mathbf{e}_j$ is in the cut then $\mathbf{e}_i$ is also in the cut.

- That is: if $\mathbf{e}_j \rightarrow$ < $\mathbf{p}_j$ records its state>, then it must be true that $\mathbf{e}_i \rightarrow$ <$\mathbf{p}_i$ records its state>.

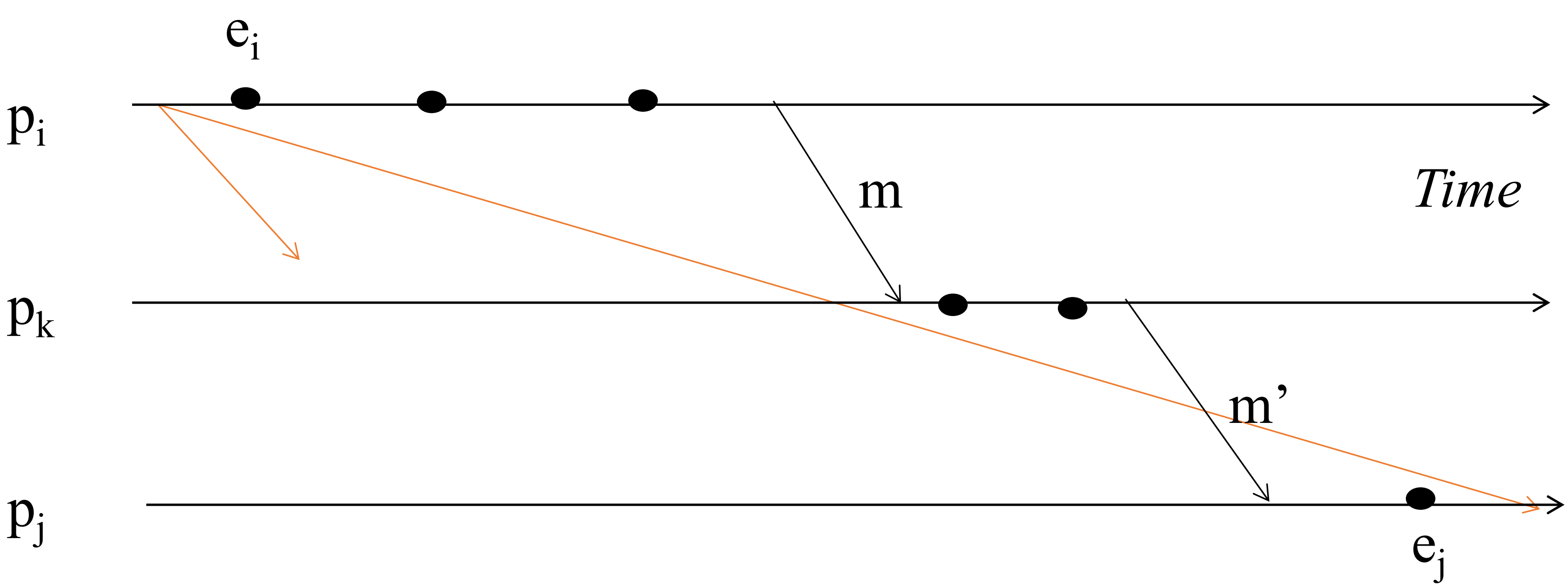
Chandy-Lamport Algorithm: Properties

- Given **e_i** → **e_j**. If **e_j** → < **p_j** records its state>, then it must be true that **e_i** → <**p_i** records its state>.
- By contradiction, *suppose* **e_j** → *<* ***p_j*** *records its state>, and <****p_i*** *records its state>* → **e_i**.



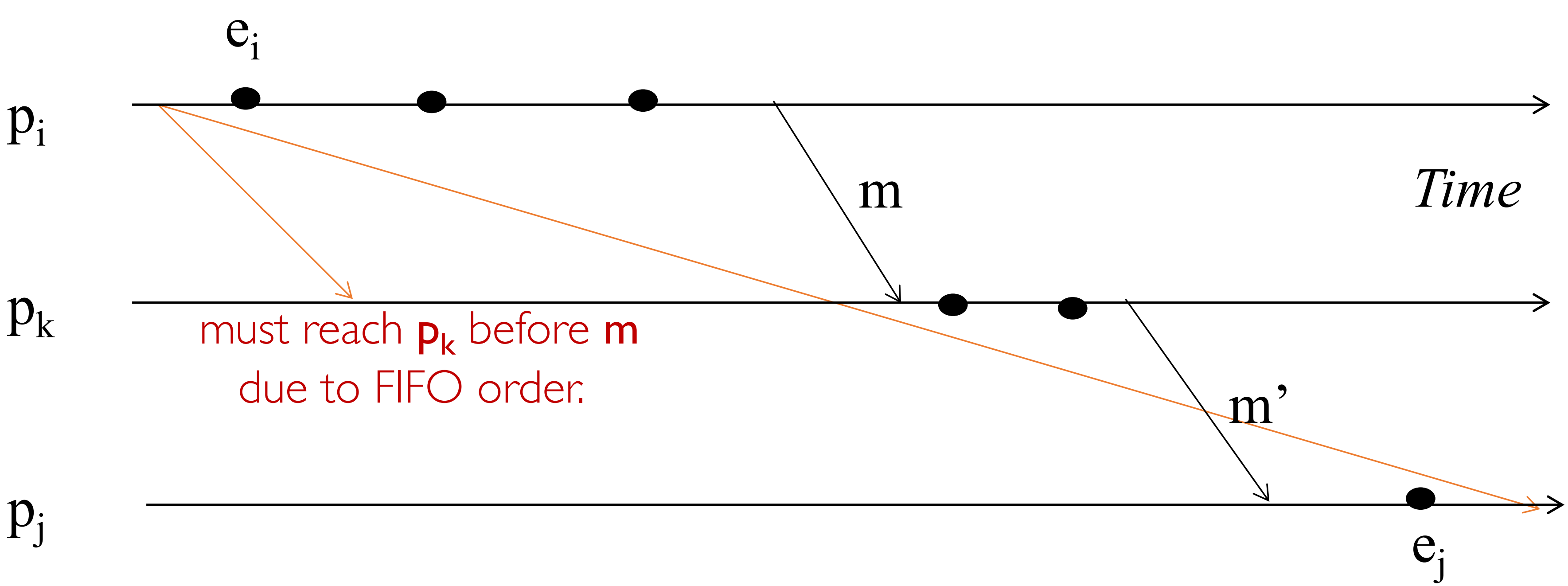
Chandy-Lamport Algorithm: Properties

- Given $\mathbf{e_i} \rightarrow \mathbf{e_j}$. If $\mathbf{e_j} \rightarrow$ < $\mathbf{p_j}$ records its state>, then it must be true that $\mathbf{e_i} \rightarrow$ <$\mathbf{p_i}$ records its state>.
- By contradiction, *suppose* $\mathbf{e_j} \rightarrow$ *< p_j records its state>, and <p_i records its state>* $\rightarrow \mathbf{e_i}$.



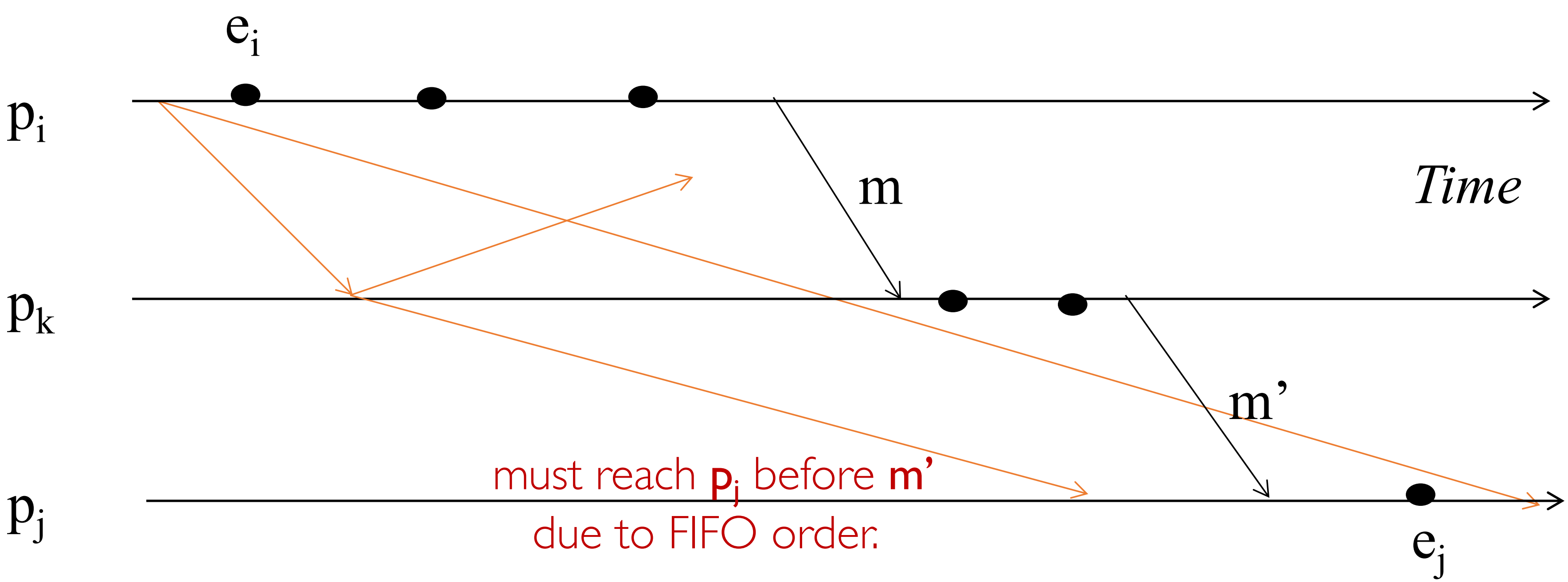
Chandy-Lamport Algorithm: Properties

- Given $\mathbf{e_i} \rightarrow \mathbf{e_j}$. If $\mathbf{e_j} \rightarrow$ < $\mathbf{p_j}$ records its state>, then it must be true that $\mathbf{e_i} \rightarrow$ <$\mathbf{p_i}$ records its state>.
- By contradiction, *suppose* $\mathbf{e_j} \rightarrow$ *< $\mathbf{p_j}$ records its state>, and <$\mathbf{p_i}$ records its state>* $\rightarrow \mathbf{e_i}$.



Chandy-Lamport Algorithm: Properties

- Given $\mathbf{e_i} \rightarrow \mathbf{e_j}$. If $\mathbf{e_j} \rightarrow$ < $\mathbf{p_j}$ records its state>, then it must be true that $\mathbf{e_i} \rightarrow$ <$\mathbf{p_i}$ records its state>.
- By contradiction, *suppose* $\mathbf{e_j} \rightarrow$ *< $\mathbf{p_j}$ records its state>, and <$\mathbf{p_i}$ records its state>* $\rightarrow \mathbf{e_i}$.



Chandy-Lamport Algorithm: Properties

- Given $\mathbf{e}_i \rightarrow \mathbf{e}_j$. If $\mathbf{e}_j \rightarrow$ < $\mathbf{p}_j$ records its state>, then it must be true that $\mathbf{e}_i \rightarrow$ <$\mathbf{p}_i$ records its state>.
- By contradiction, *suppose* $\mathbf{e}_j \rightarrow$ *< p_j records its state>, and <p_i records its state>* $\rightarrow \mathbf{e}_i$.
- Consider the path of app messages (through other processes) that go from $\mathbf{e}_i$ to $\mathbf{e}_j$.
- Due to FIFO ordering, markers on each link in above path will precede regular app messages.
- Thus, since <$\mathbf{p}_i$ records its state> $\rightarrow \mathbf{e}_i$, it must be true that $\mathbf{p}_j$ received a marker before $\mathbf{e}_j$.
- Thus $\mathbf{e}_j$ is not in the cut => contradiction.

Summary

- The ability to calculate global snapshots in a distributed system is very important.
- But don't want to interrupt running distributed application.
- Chandy-Lamport algorithm calculates global snapshot.
- Obeys causality (creates a consistent cut).
- Can be used to detect global properties.
  - Safety vs. Liveness.

Chandy-Lamport Algorithm: Usefulness

- Consistent global snapshots are useful for detecting global system properties:
  - Safety
  - Liveness

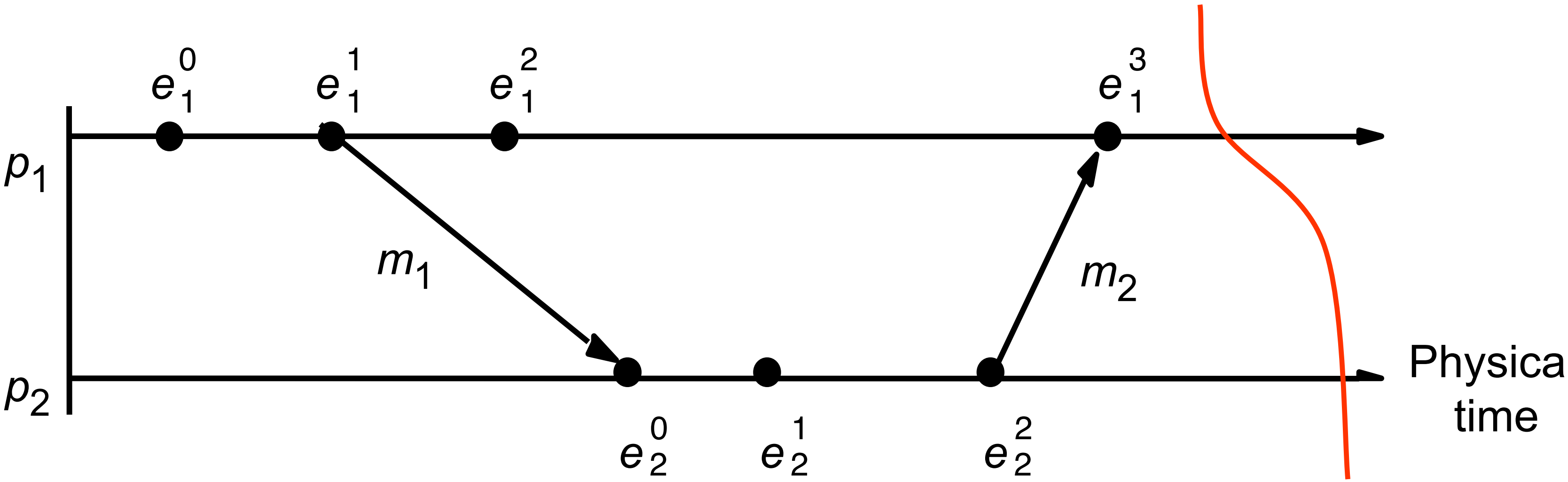
More notations and definitions

- **history(p_i)** = $h_i = <e_i^0, e_i^1, \ldots >$
- **global history**: $H = \cup_i (h_i)$

- A **run** is a total ordering of events in H that is consistent with each **h_i**'s ordering.

- A **linearization** is a run consistent with happens-before ($\rightarrow$) relation in H.

Example



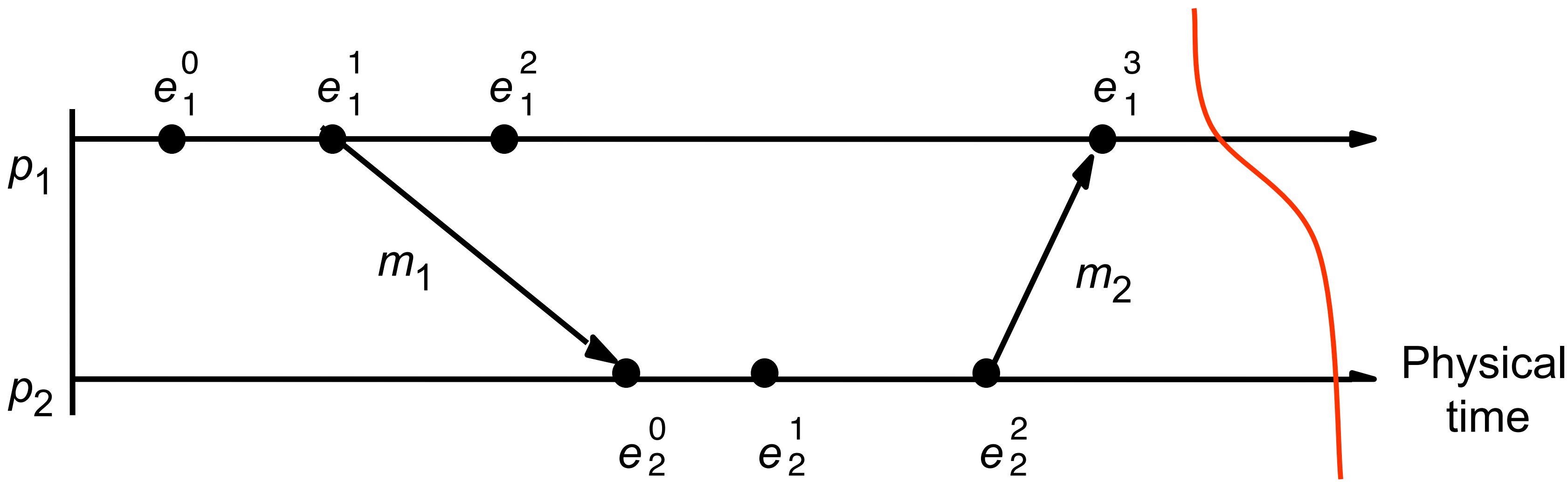
Order at p_1: < e_1^0, e_1^1, e_1^2, e_1^3 >   Order at p_2: < e_2^0, e_2^1,e_2^2>

Causal order across p_1 and p_2: < e_1^0, e_1^1, e_2^0, e_2^1, e_2^2 , e_1^3 >

Run: < e_1^0, e_1^1, e_1^2, e_1^3 , e_2^0, e_2^1 e_2^2 >

Linearization: < e_1^0, e_1^1, e_1^2, e_2^0, e_2^1 e_2^2 , e_1^3 >

Example



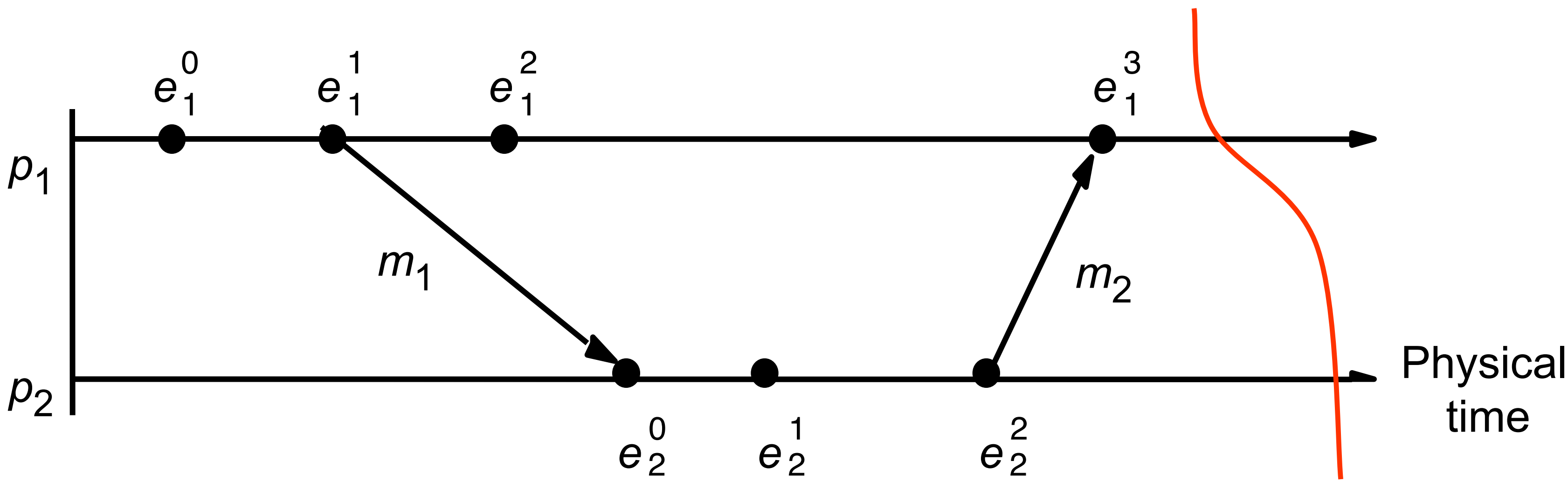
Order at p_1: $< e_1^0, e_1^1, e_1^2, e_1^3 >$ Order at p_2: $< e_2^0, e_2^1, e_2^2 >$

Causal order across p_1 and p_2: $< e_1^0, e_1^1, e_2^0, e_2^1, e_2^2, e_1^3 >$

Run: $< e_1^0, e_1^1, e_1^2, e_1^3, e_2^0, e_2^1\ e_2^2 >$

Linearization: $< e_1^0, e_1^1, e_1^2, e_2^0, e_2^1\ e_2^2, e_1^3 >$

Example

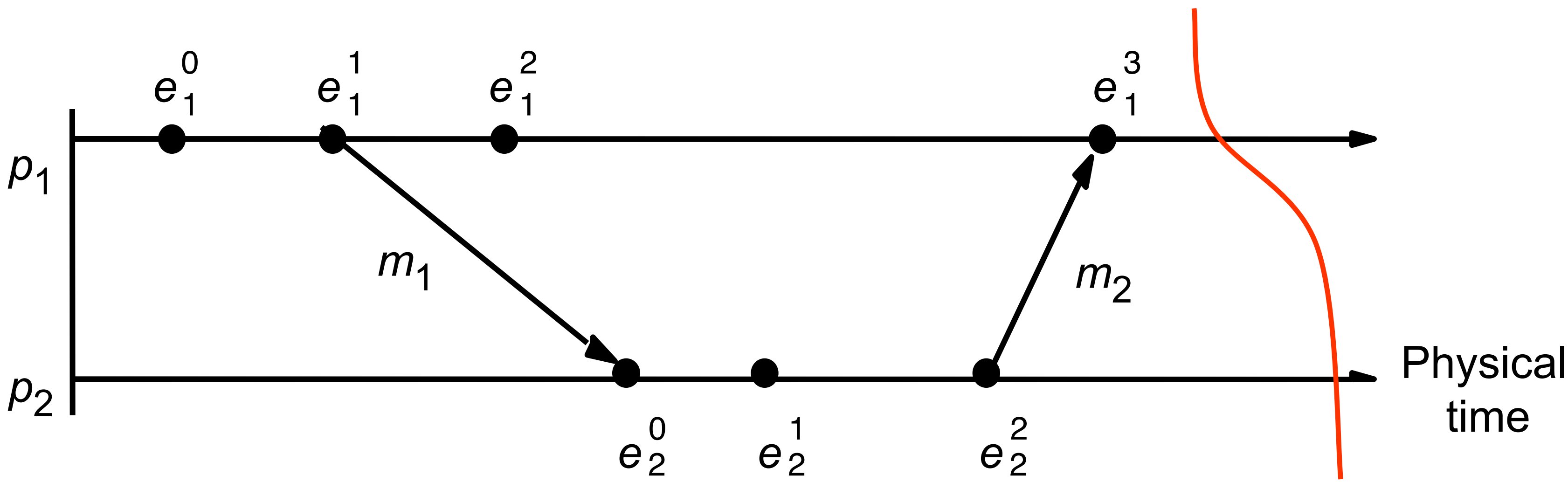


Order at p_1: $< e_1^0, e_1^1, e_1^2, e_1^3 >$ Order at p_2: $< e_2^0, e_2^1, e_2^2 >$

Causal order across p_1 and p_2: $< e_1^0, e_1^1, e_2^0, e_2^1, e_2^2, e_1^3 >$

$\mathbf{< e_1^0, e_1^1, e_2^0, e_2^1, e_1^2, e_2^2, e_1^3 >}$

Example



Order at p_1: $< e_1^0, e_1^1, e_1^2, e_1^3 >$ Order at p_2: $< e_2^0, e_2^1, e_2^2 >$

Causal order across p_1 and p_2: $< e_1^0, e_1^1, e_2^0, e_2^1, e_2^2, e_1^3 >$

$< e_1^0, e_1^1, e_2^0, e_2^1, e_1^2, e_2^2, e_1^3 >$: Linearization

$< e_1^0, e_2^1, e_2^0, e_1^1, e_1^2, e_2^2, e_1^3 >$: Not even a run

More notations and definitions

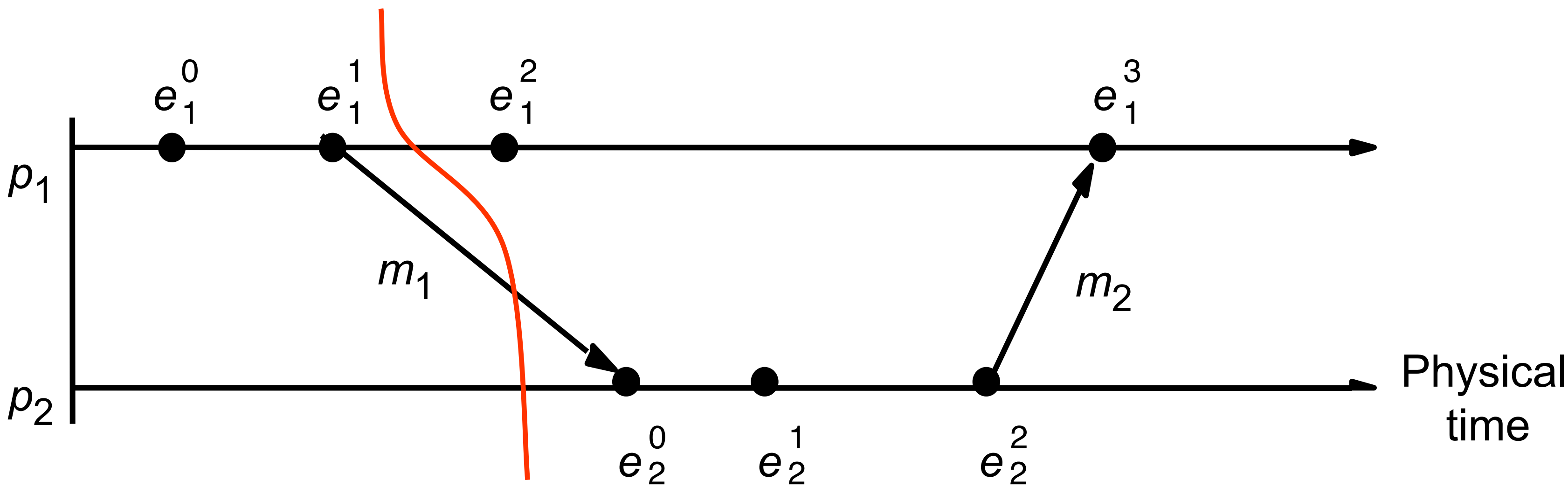
- **history(p_i)** = $h_i = <e_i^0, e_i^1, \ldots>$
- **global history**: $H = \cup_i (h_i)$

- A **run** is a total ordering of events in H that is consistent with each $\mathbf{h_i}$'s ordering.

- A **linearization** is a run consistent with happens-before ($\rightarrow$) relation in H.

- Linearizations pass through consistent global states.

Example

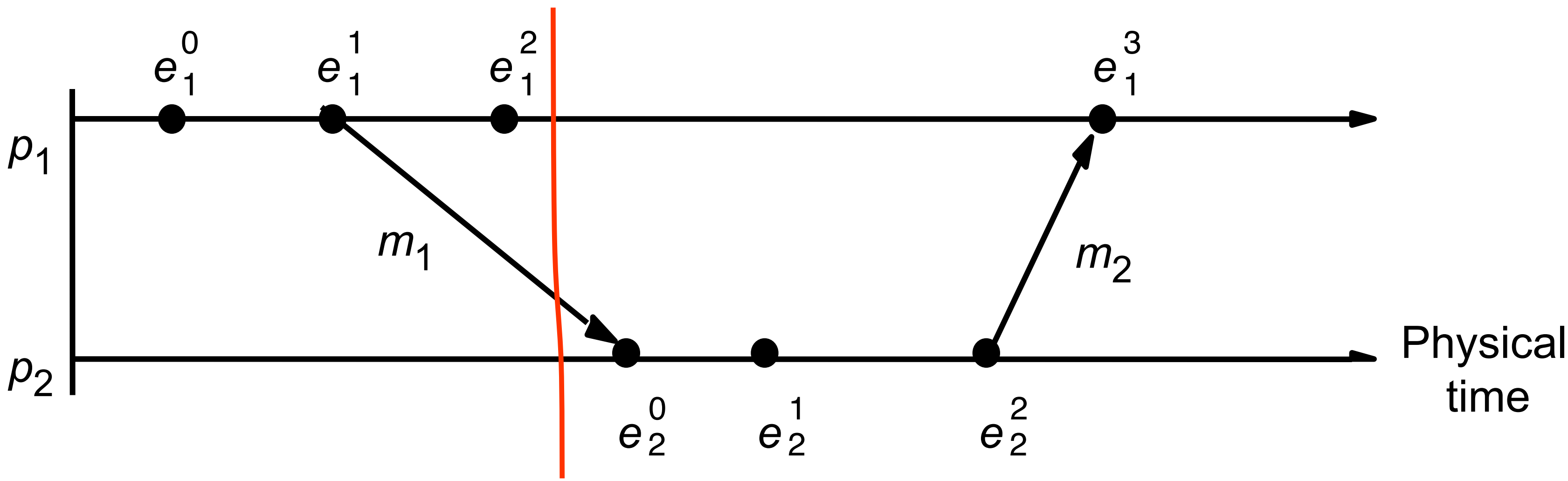


Order at p_1: < e_1^0, e_1^1, e_1^2, e_1^3 > Order at p_2: < e_2^0, e_2^1, e_2^2 >

Causal order across p_1 and p_2: < e_1^0, e_1^1, e_2^0, e_2^1, e_2^2 , e_1^3 >

Linearization: < e_1^0, e_1^1, | e_1^2, e_2^0, e_2^1 e_2^2 , e_1^3 >

Example

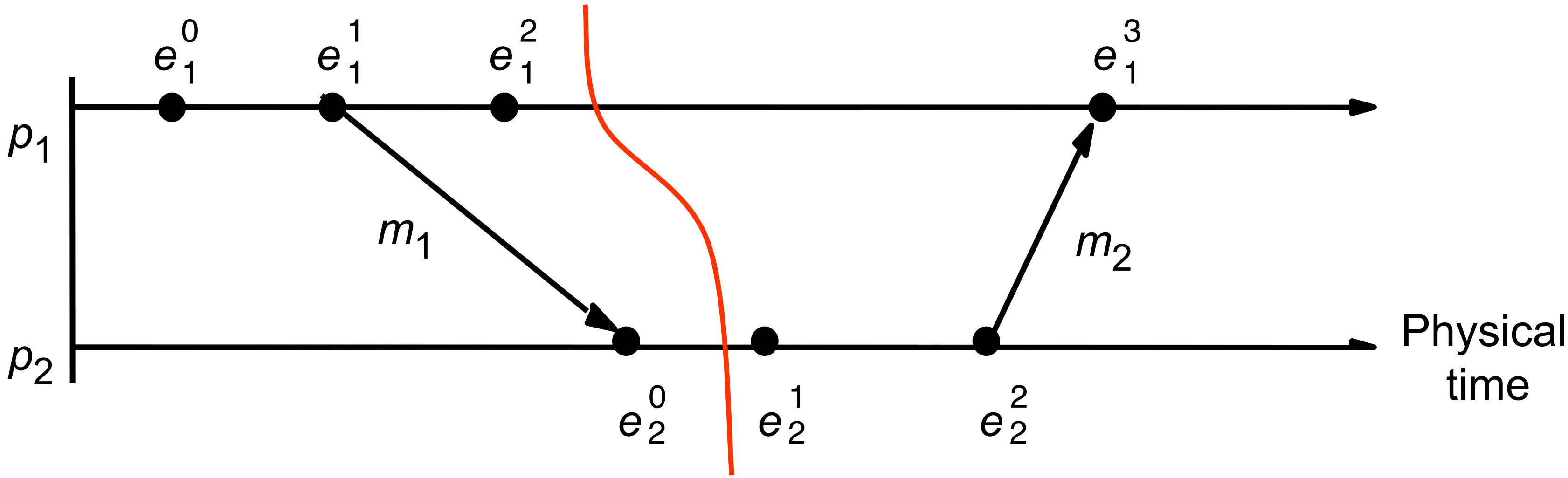


Order at p_1: < e_1^0, e_1^1, e_1^2, e_1^3 > Order at p_2: < e_2^0, e_2^1, e_2^2 >

Causal order across p_1 and p_2: < e_1^0, e_1^1, e_2^0, e_2^1 e_2^2 , e_1^3 >

Linearization: < e_1^0, e_1^1, e_1^2, e_2^0, e_2^1 e_2^2 , e_1^3 >

Example

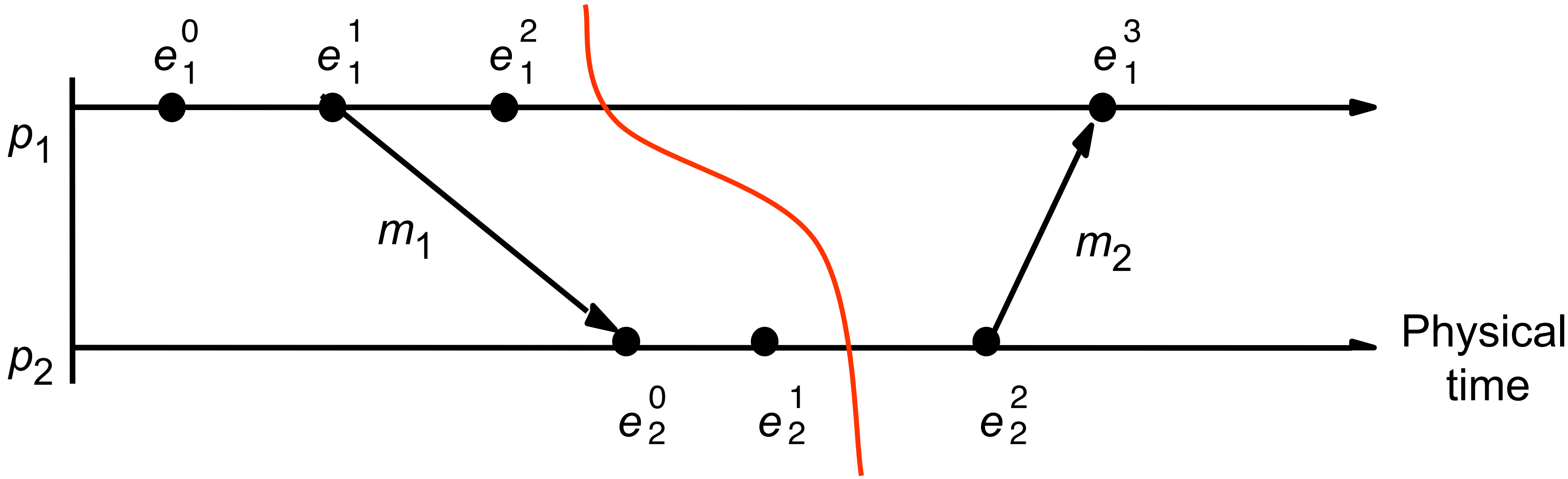


Order at p_1: < $e_1^0, e_1^1, e_1^2, e_1^3$ >   Order at p_2: < e_2^0, e_2^1, e_2^2 >

Causal order across p_1 and p_2: < $e_1^0, e_1^1, e_2^0, e_2^1$ e_2^2 , e_1^3 >

Linearization: < $e_1^0, e_1^1, e_1^2, e_2^0$, | e_2^1 e_2^2 , e_1^3 >

Example

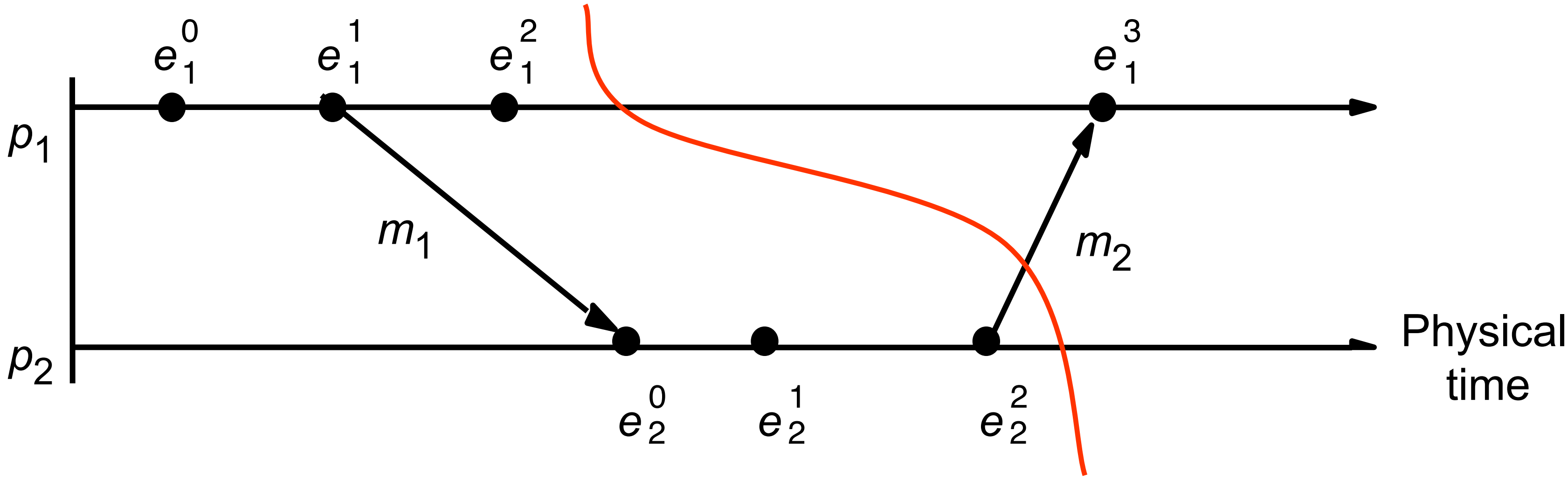


Order at p_1: $< e_1^0, e_1^1, e_1^2, e_1^3 >$ Order at p_2: $< e_2^0, e_2^1, e_2^2 >$

Causal order across p_1 and p_2: $< e_1^0, e_1^1, e_2^0, e_2^1\ e_2^2, e_1^3 >$

Linearization: $< e_1^0, e_1^1, e_1^2, e_2^0, e_2^1 \mid e_2^2, e_1^3 >$

Example

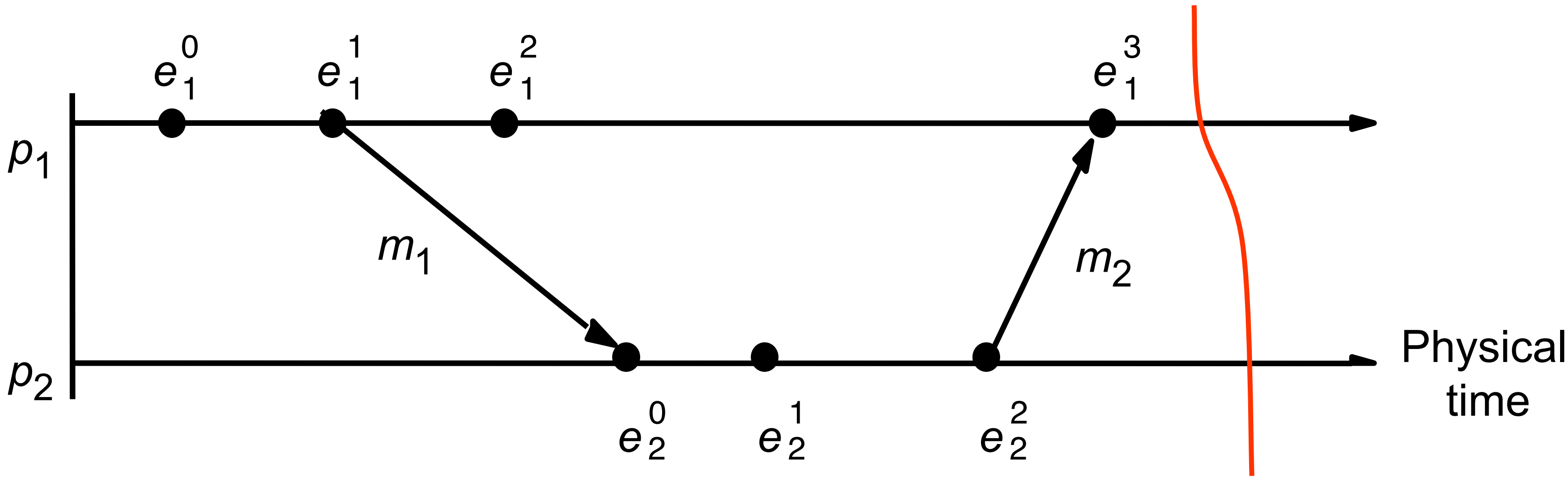


Order at p_1: < $e_1^0, e_1^1, e_1^2, e_1^3$ >   Order at p_2: < e_2^0, e_2^1, e_2^2 >

Causal order across p_1 and p_2: < $e_1^0, e_1^1, e_2^0, e_2^1\ e_2^2, e_1^3$ >

Linearization: < $e_1^0, e_1^1, e_1^2, e_2^0, e_2^1\ e_2^2, e_1^3$ >

Example

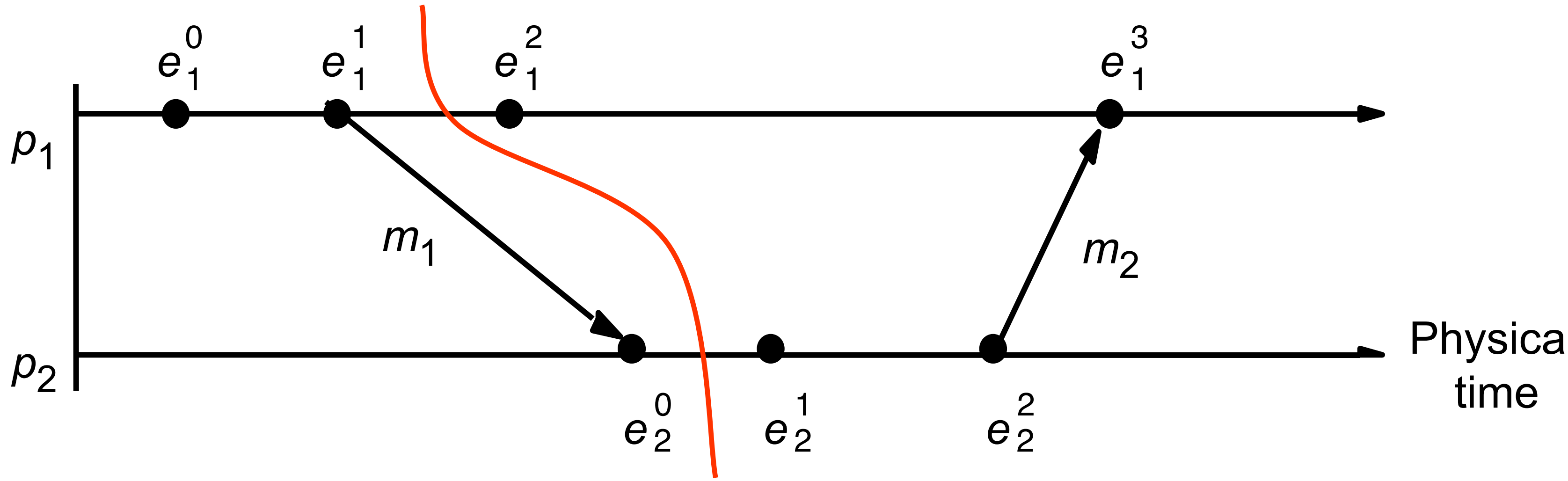


Order at p_1: < $e_1^0, e_1^1, e_1^2, e_1^3$ >   Order at p_2: < e_2^0, e_2^1, e_2^2 >

Causal order across p_1 and p_2: < $e_1^0, e_1^1, e_2^0, e_2^1\ e_2^2, e_1^3$ >

Linearization: < $e_1^0, e_1^1, e_1^2, e_2^0, e_2^1\ e_2^2, e_1^3$ >

Example



Order at p_1: $< e_1^0, e_1^1, e_1^2, e_1^3 >$ Order at p_2: $< e_2^0, e_2^1, e_2^2 >$

Causal order across p_1 and p_2: $< e_1^0, e_1^1, e_2^0, e_2^1\ e_2^2, e_1^3 >$

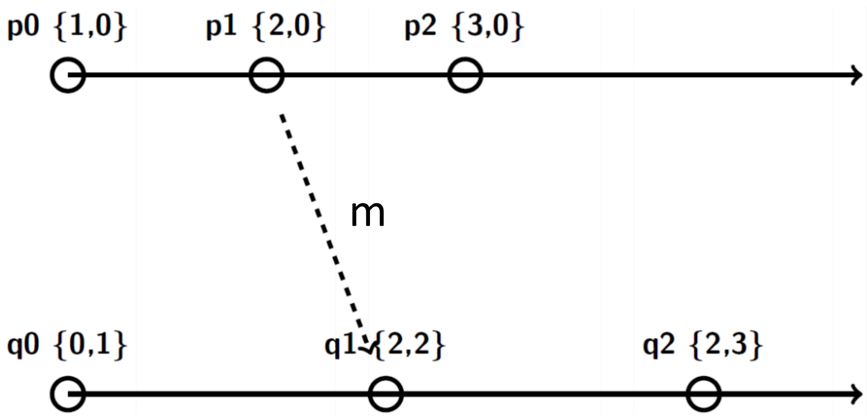
Linearization: $< e_1^0, e_1^1, e_1^2, e_2^0, e_2^1\ e_2^2, e_1^3 >$

Linearization $< e_1^0, e_1^1, e_2^0, e_2^1, e_1^2, e_2^2, e_1^3 >$

More notations and definitions

- Linearizations pass through consistent global states.
- A global state $\mathbf{S_k}$ is reachable from global state $\mathbf{S_i}$, if there is a linearization that passes through $\mathbf{S_i}$ and then through $\mathbf{S_k}$.
- The distributed system evolves as a series of transitions between global states S_0 , S_1 , ….

State Transitions: Example



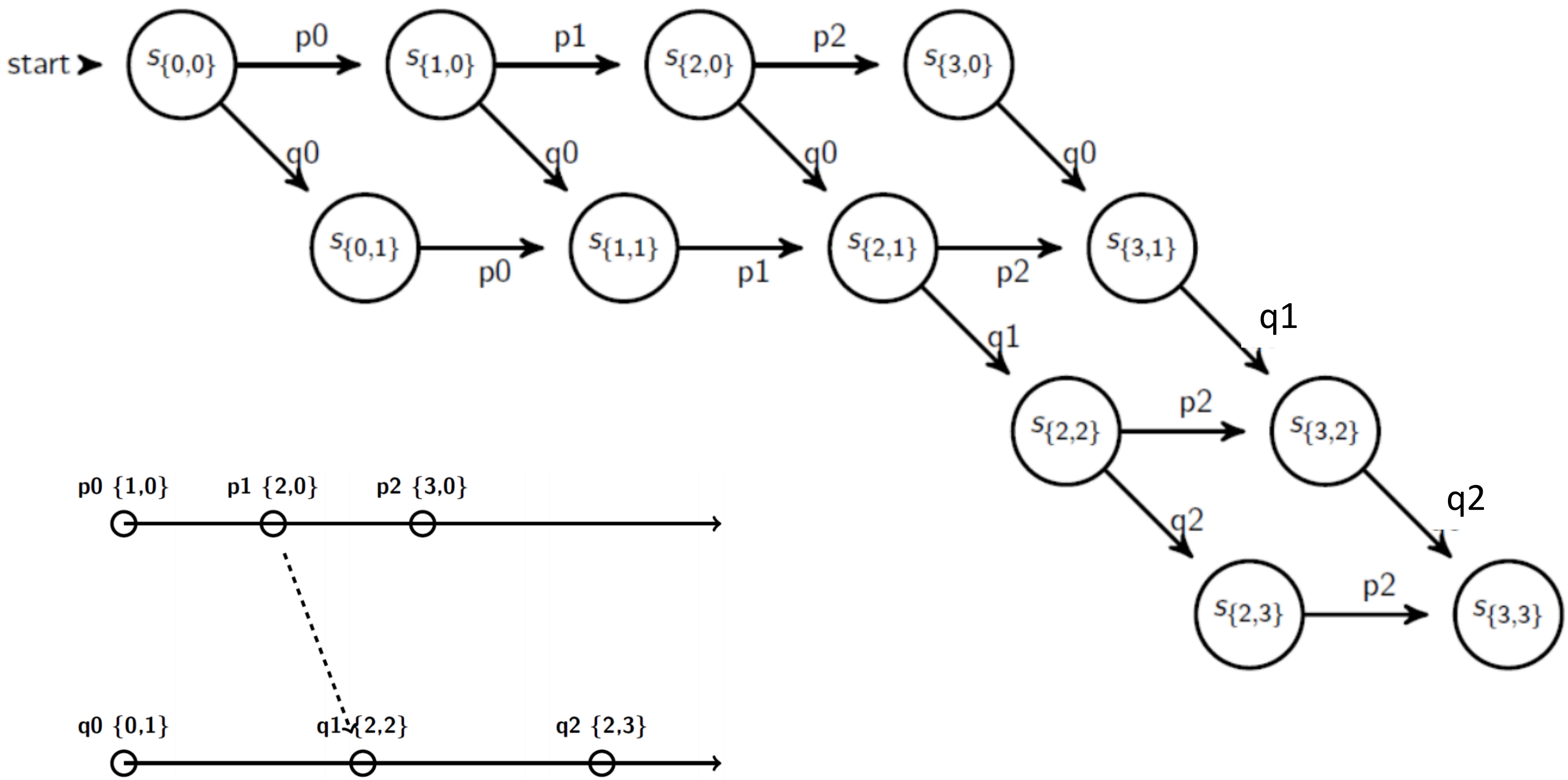
- Causal order:
  - p0 → p1 → p2
  - q0 → q1 → q2
  - p0 → p1 → q1 → q2
- Concurrent:
  - p0 || q0
  - p1 || q0
  - p2 || q0, p2 || q1, p2 || q2

Many linearizations:

- < p0, p1, p2, q0, q1, q2>
- < p0, q0, p1, q1, p2, q2>
- <q0, p0, p1, q1, p2, q2 >
- <q0, p0, p1, p2, q1,q2 >
- ……

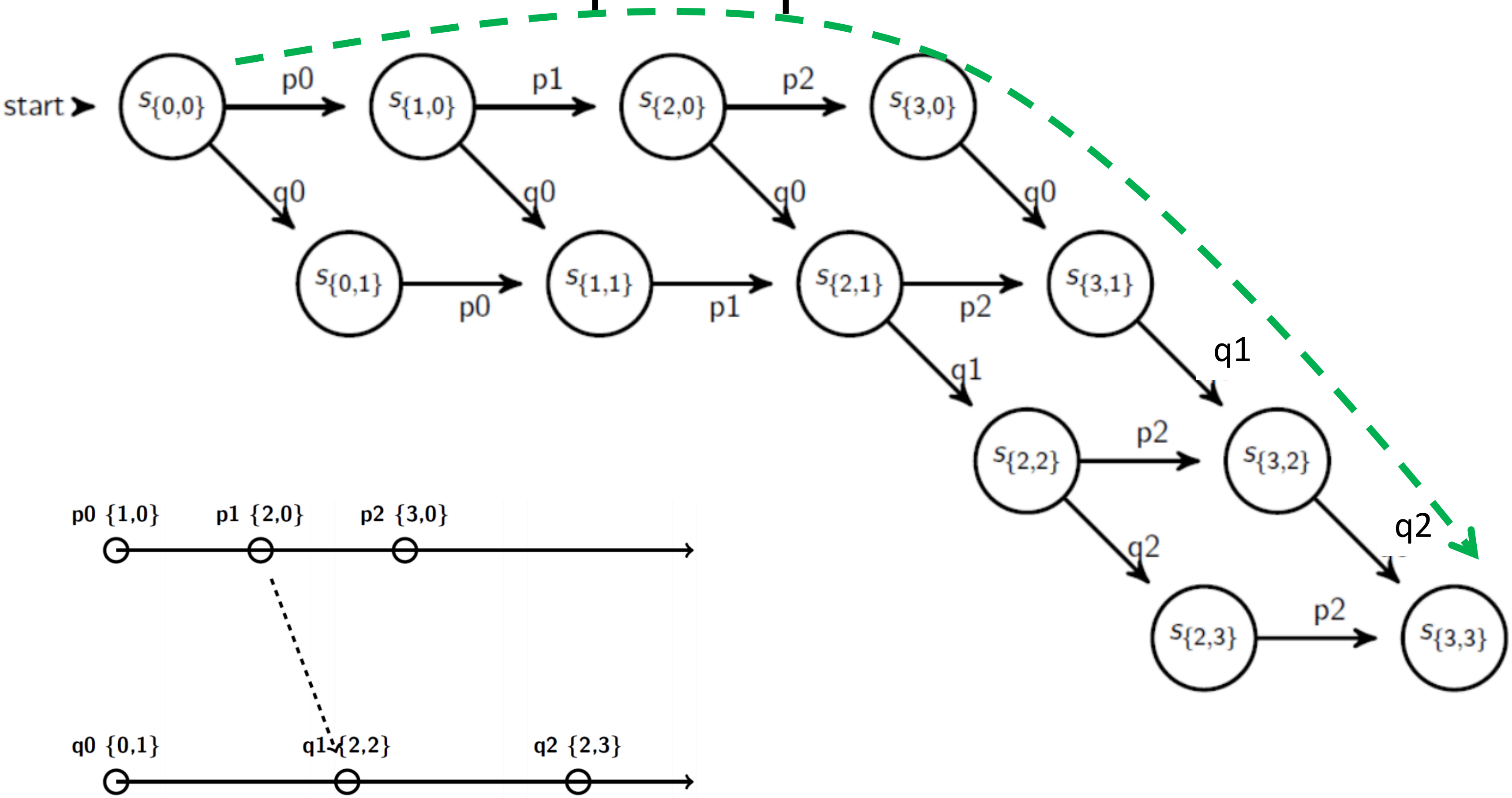
State Transitions: Example

Execution Lattice. Each path represents a linearization.



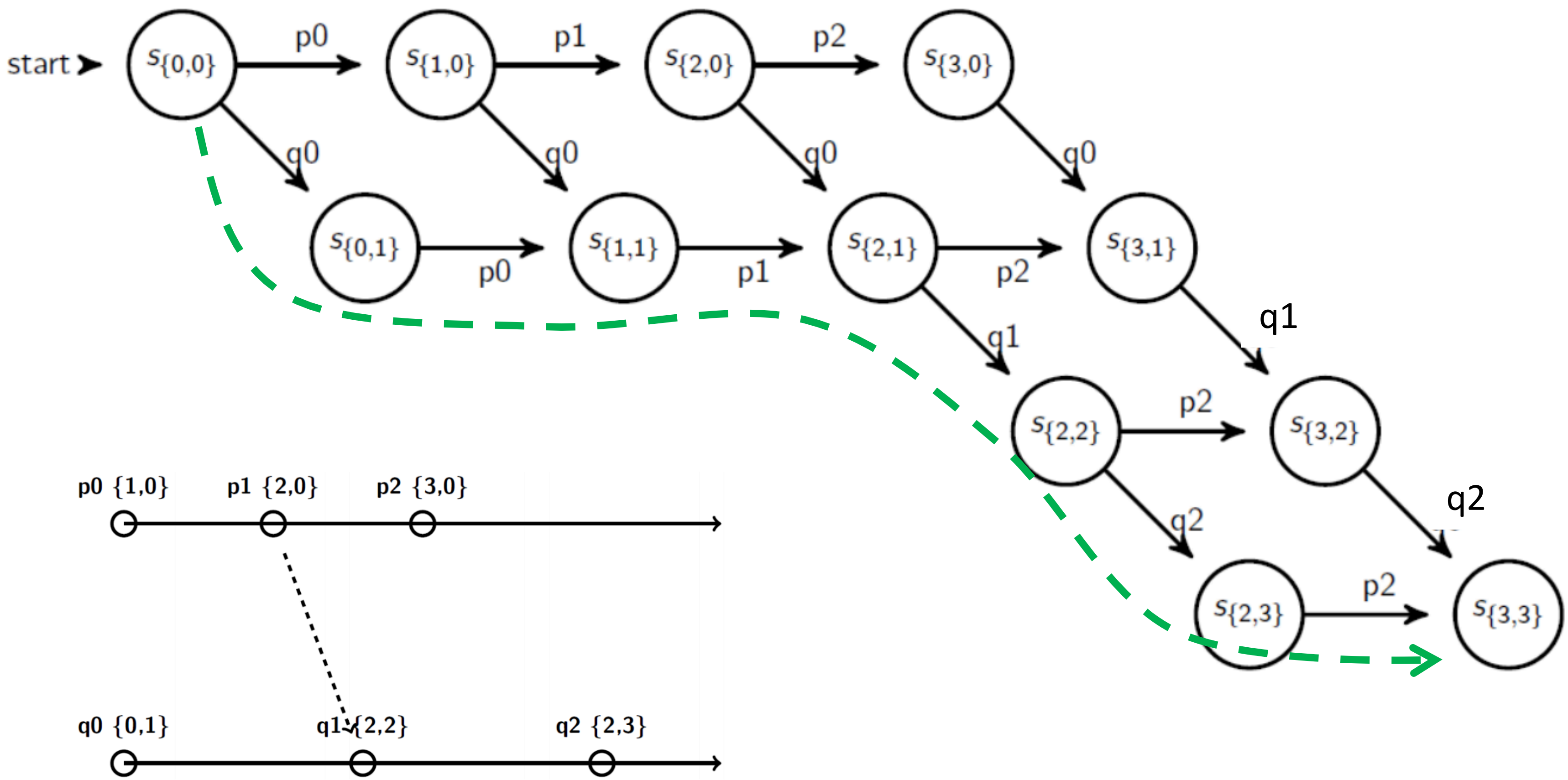
State Transitions: Example

Execution Lattice. Each path represents a linearization.



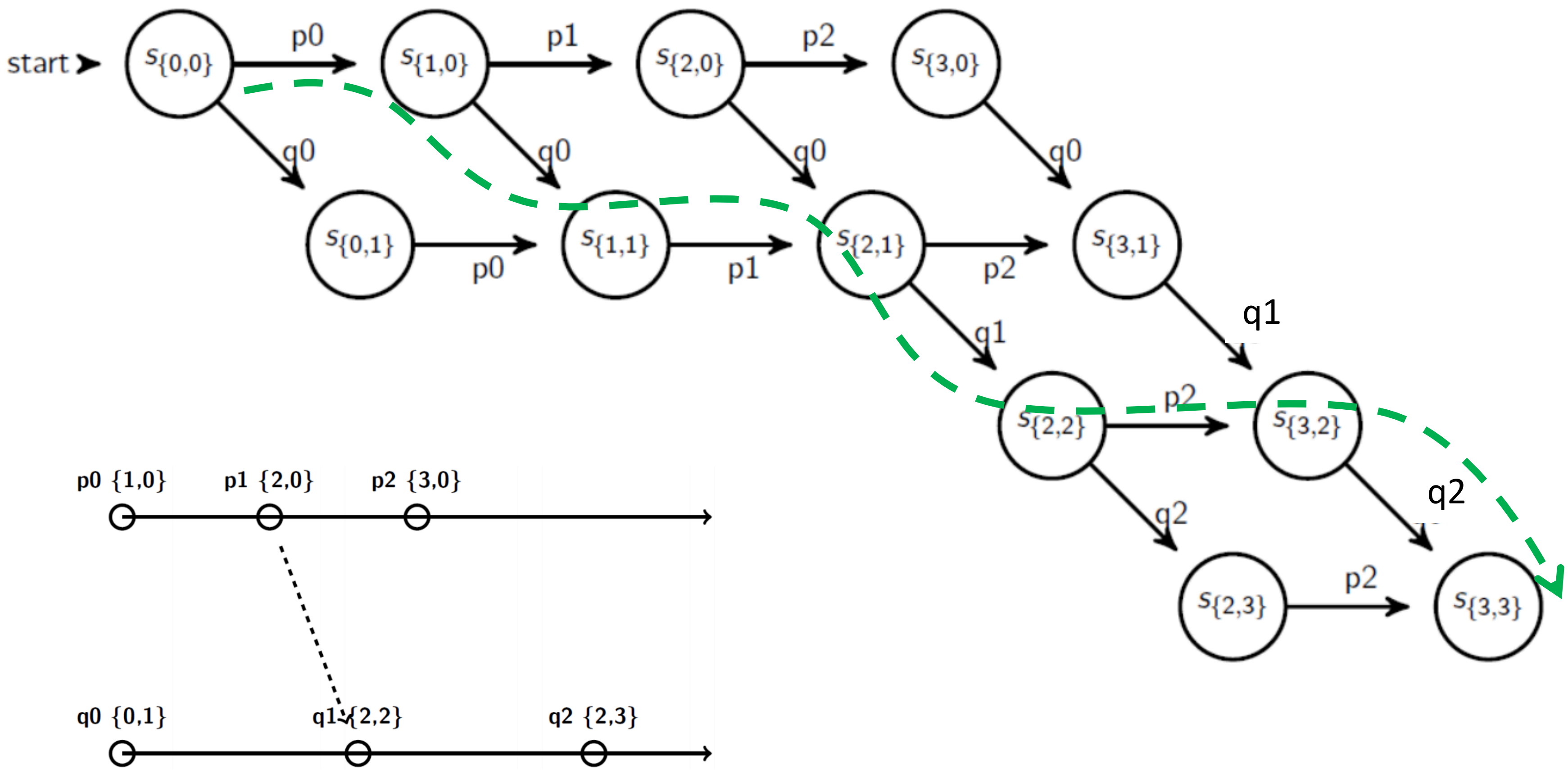
State Transitions: Example

Execution Lattice. Each path represents a linearization.

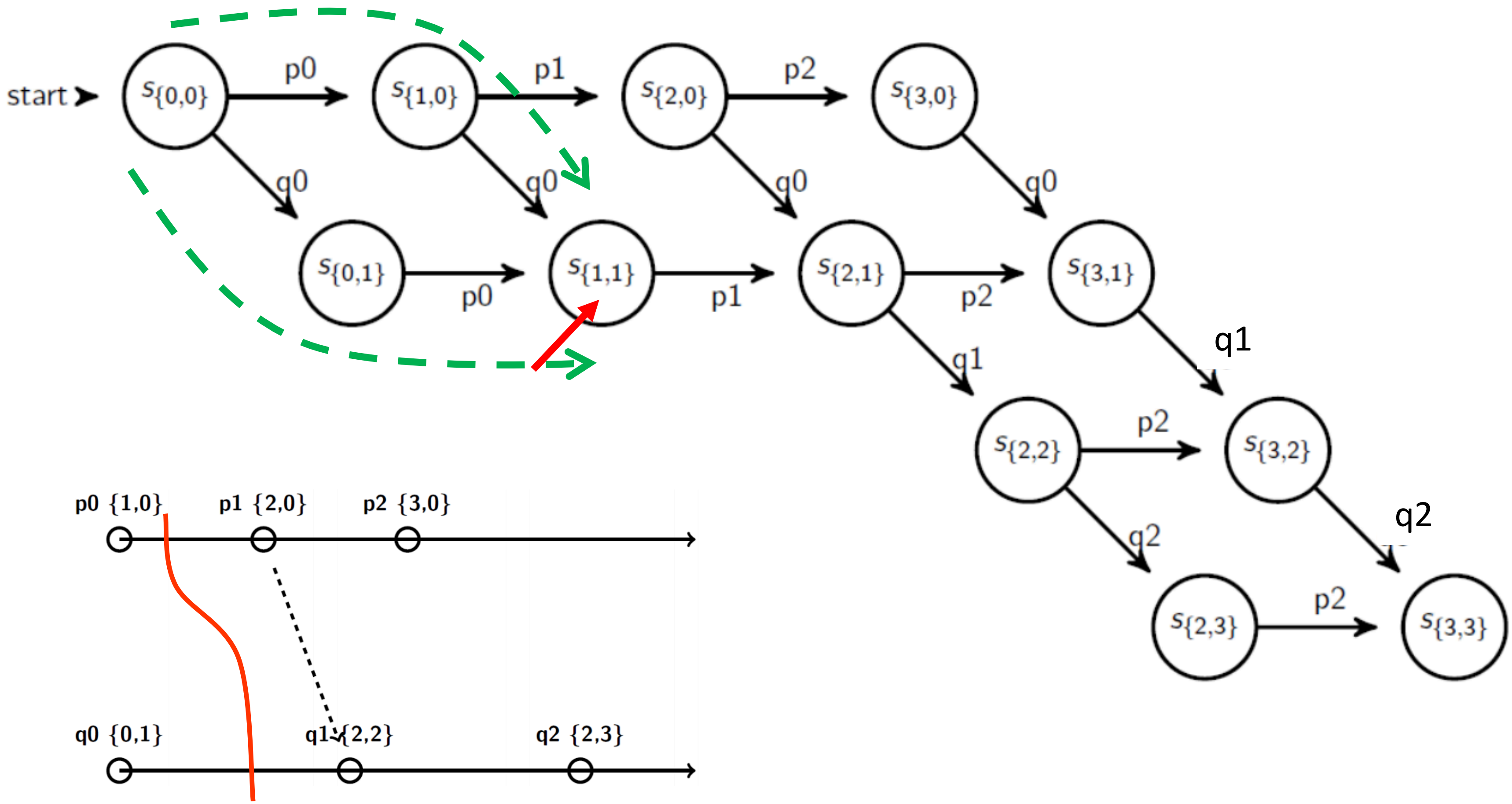


State Transitions: Example

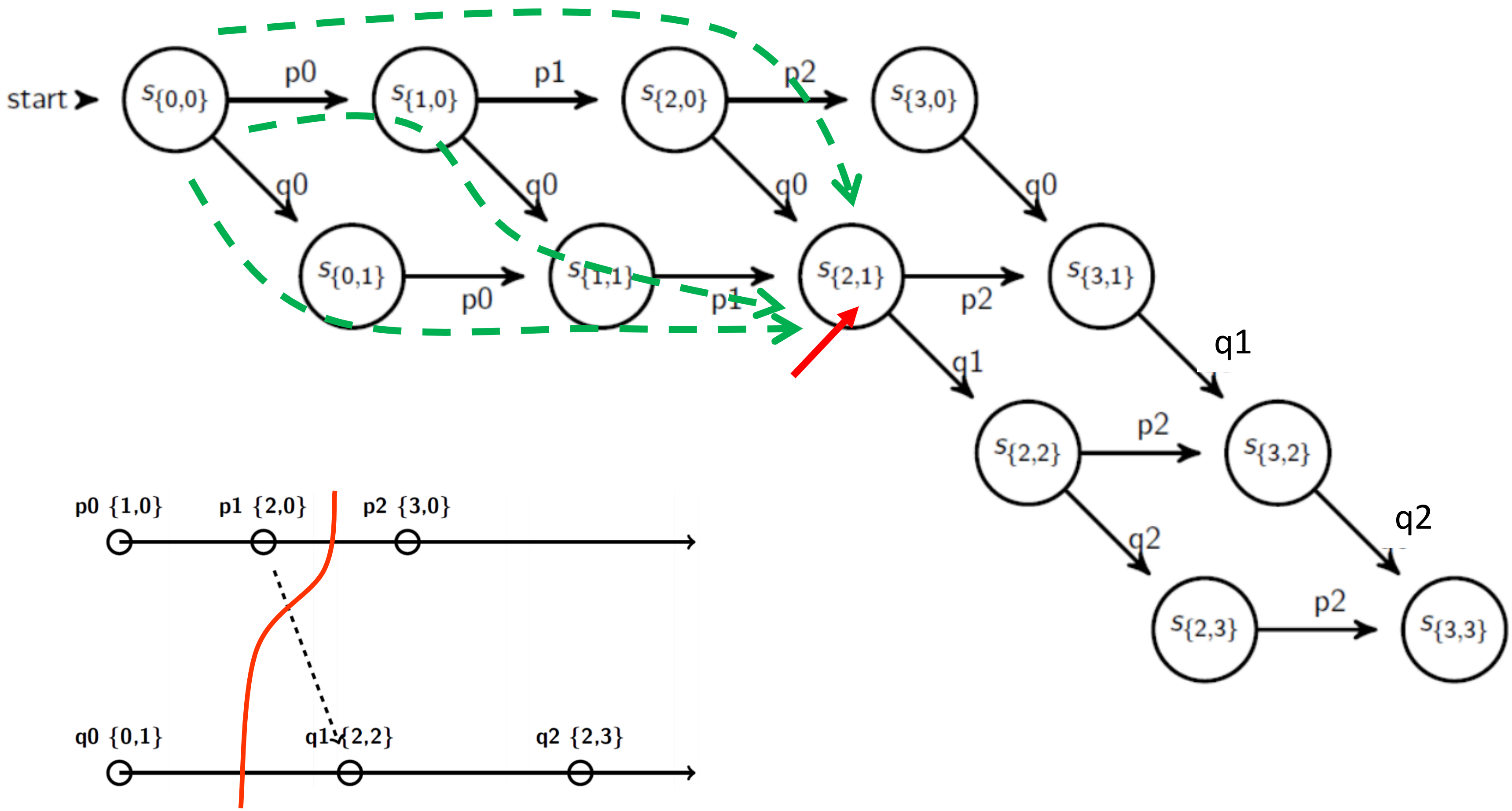
Execution Lattice. Each path represents a linearization.



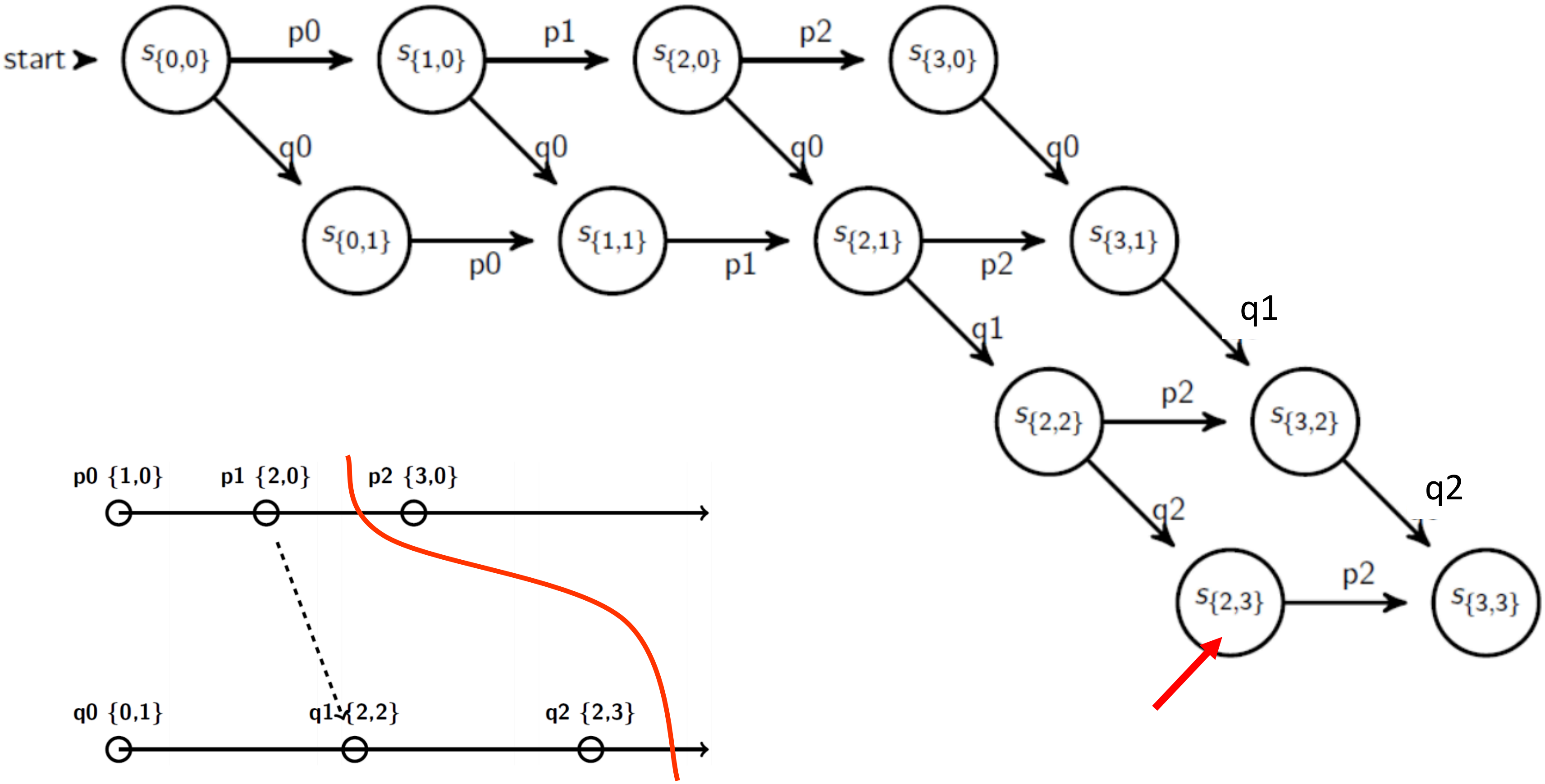
State Transitions: Example



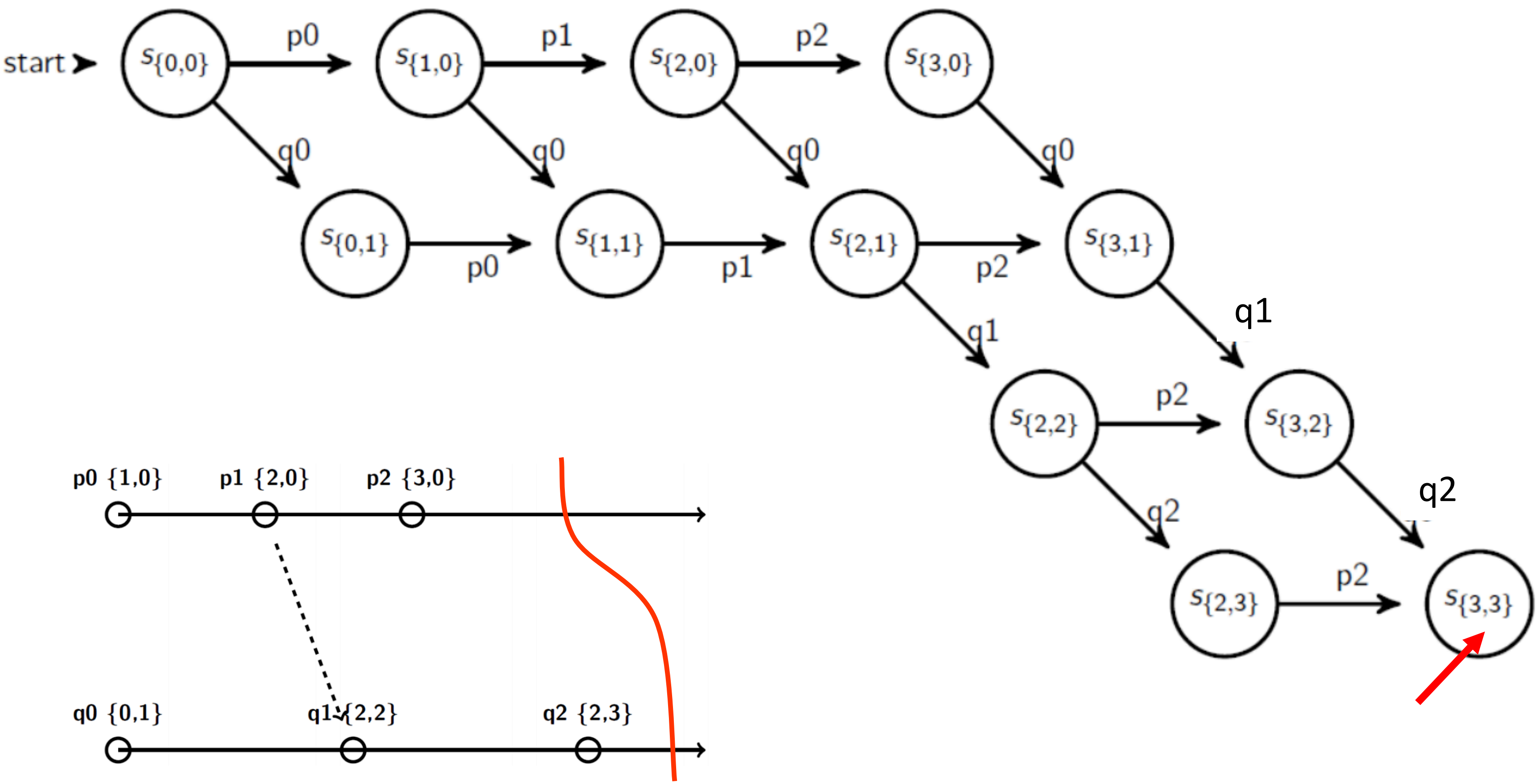
State Transitions: Example



State Transitions: Example



State Transitions: Example



More notations and definitions

- A run is a total ordering of events in H that is consistent with each $\mathbf{h_i}$'s ordering.
- A linearization is a run consistent with happens-before ($\rightarrow$) relation in H.
- Linearizations pass through consistent global states.
- A global state $\mathbf{S_k}$ is reachable from global state $\mathbf{S_i}$, if there is a linearization that passes through $\mathbf{S_i}$ and then through $\mathbf{S_k}$.
- The distributed system evolves as a series of transitions between global states S_0 , S_1 , ….

Global State Predicates

- A global-state-predicate is a property that is *true* or *false* for a global state.
  - Is there a deadlock?
  - Has the distributed algorithm terminated?

- Two ways of reasoning about predicates (or system properties) as global state gets transformed by events.
  - Liveness
  - Safety

Liveness

- Liveness = guarantee that something good will happen, eventually

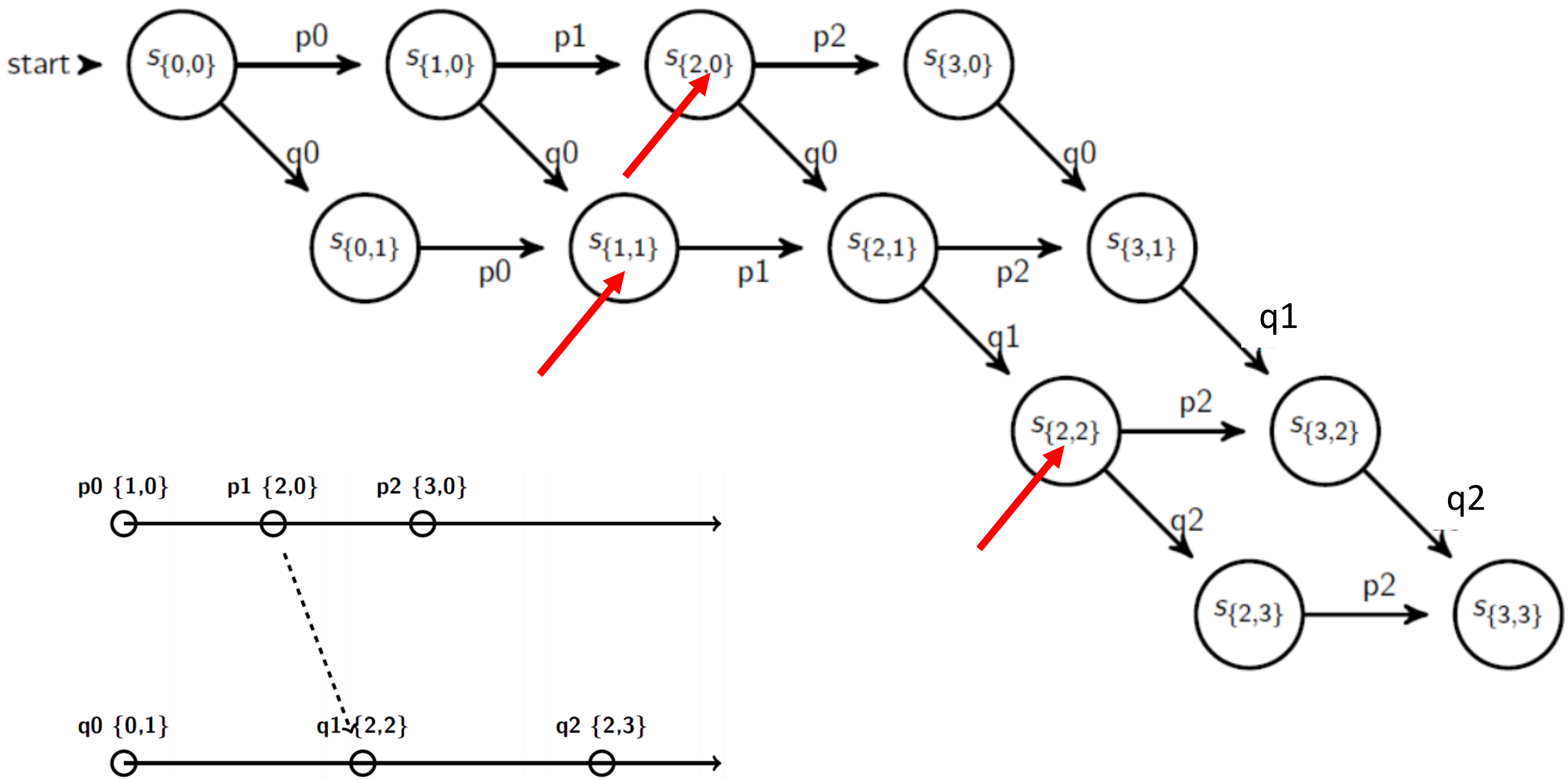
- Examples:
  - A distributed computation will terminate.
  - "Completeness" in failure detectors: the failure will be detected.
  - All processes will eventually decide on a value.

- A global state S_0 satisfies a **liveness** property P iff:
  - liveness(P(S_0)) $\equiv$ $\forall L \in$ linearizations from S_0, L passes through a S_L & P(S_L) = true
  - For all linearizations starting from S_0, P is true for some state S_L reachable from S_0.

Liveness Example

If predicate is true only in the marked states, does it satisfy liveness?

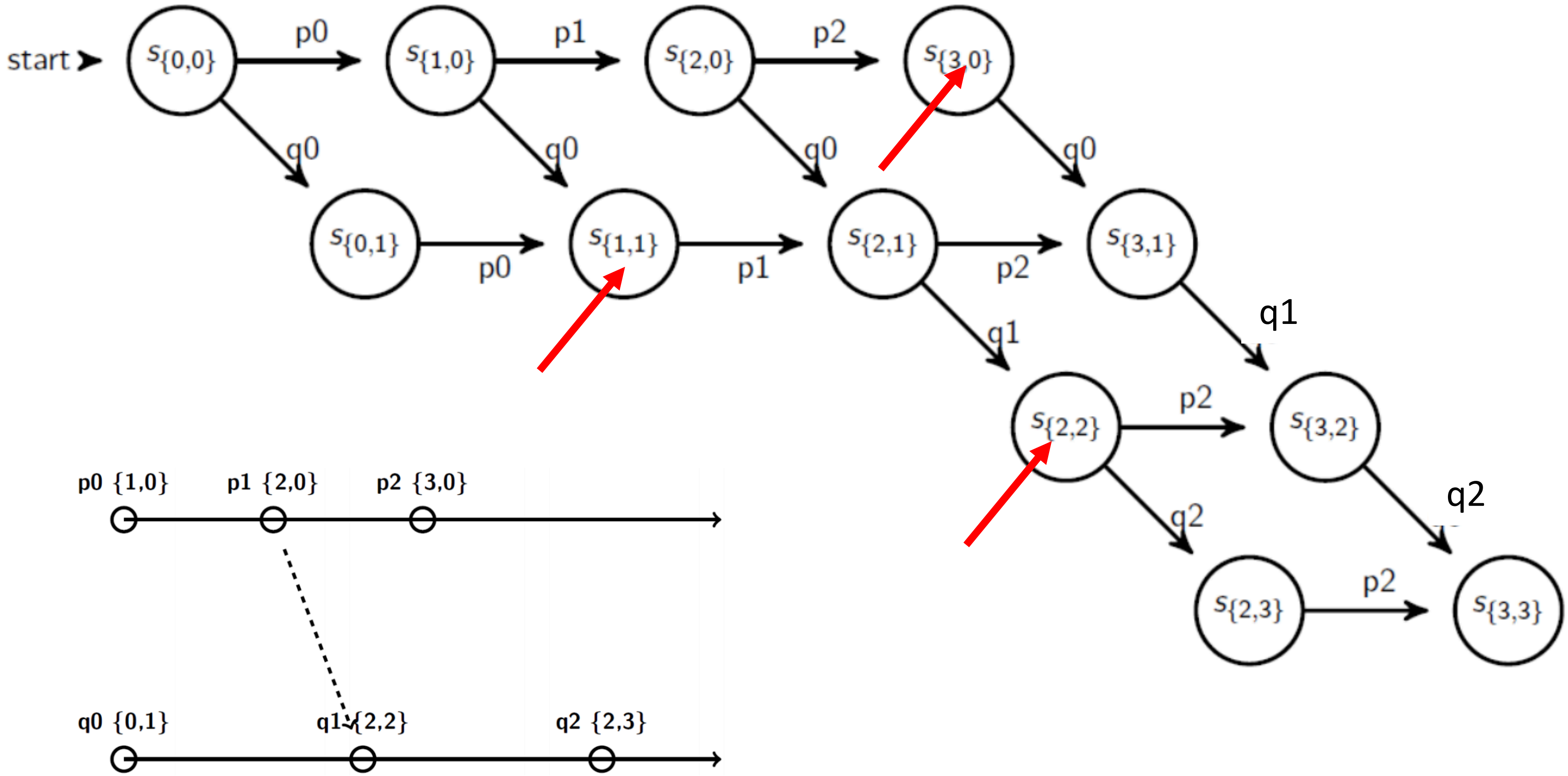
Yes



Liveness Example

If predicate is true only in the marked states, does it satisfy liveness?

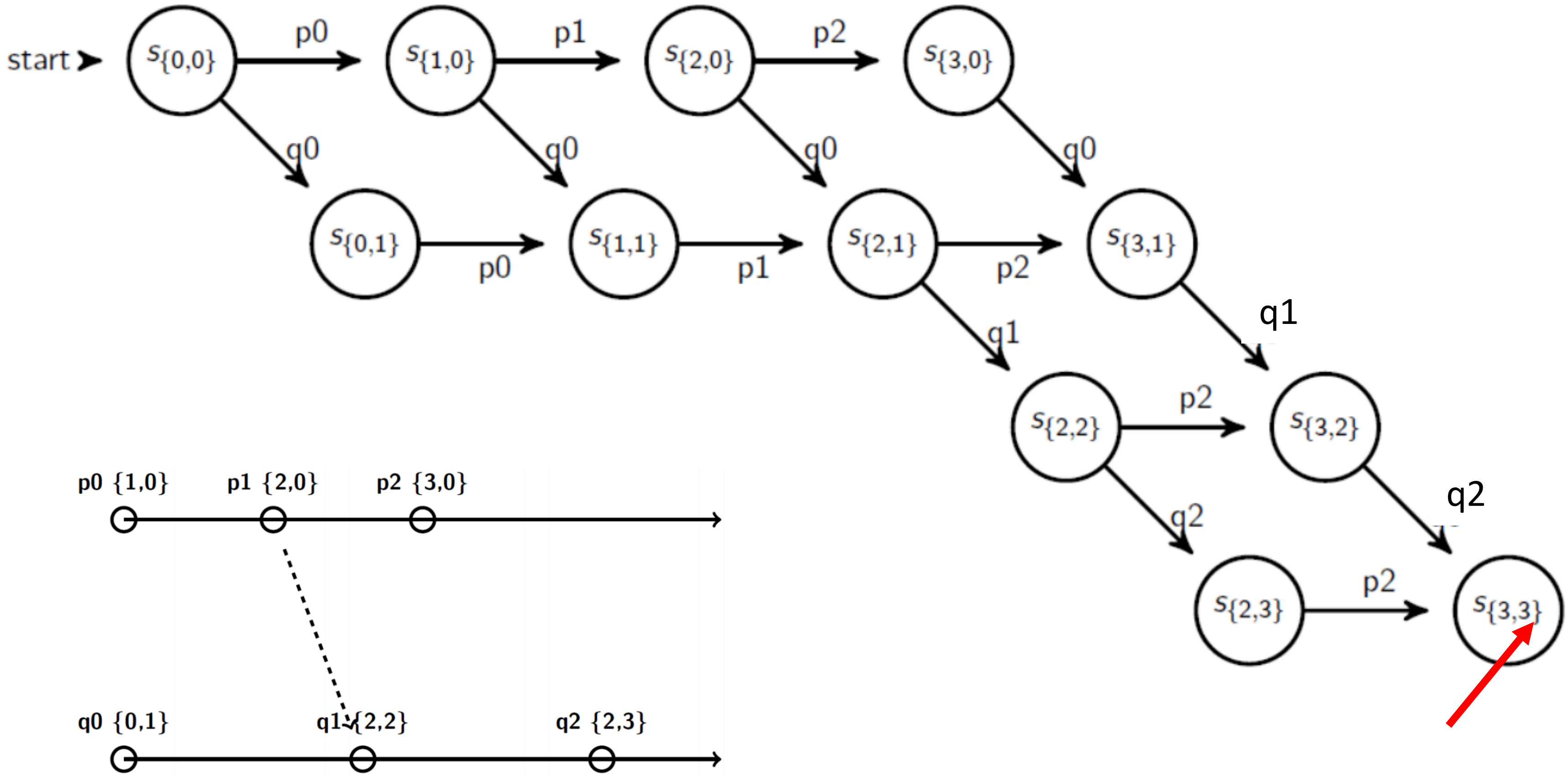
No



Liveness Example

If predicate is true only in the marked states, does it satisfy liveness?

Yes



Liveness

- **Liveness = guarantee that something good will happen, eventually**
- **Examples:**
  - A distributed computation will terminate.
  - "Completeness" in failure detectors: the failure will be detected.
  - All processes will eventually decide on a value.
- A global state S_0 satisfies a **liveness** property P iff:
  - liveness(P(S_0)) ≡ ∀L∈ linearizations from S_0, L passes through a S_L & P(S_L) = true
  - For any linearization starting from S_0, P is true for some state S_L reachable from S_0.

Safety

- **Safety = guarantee that something bad will never happen.**

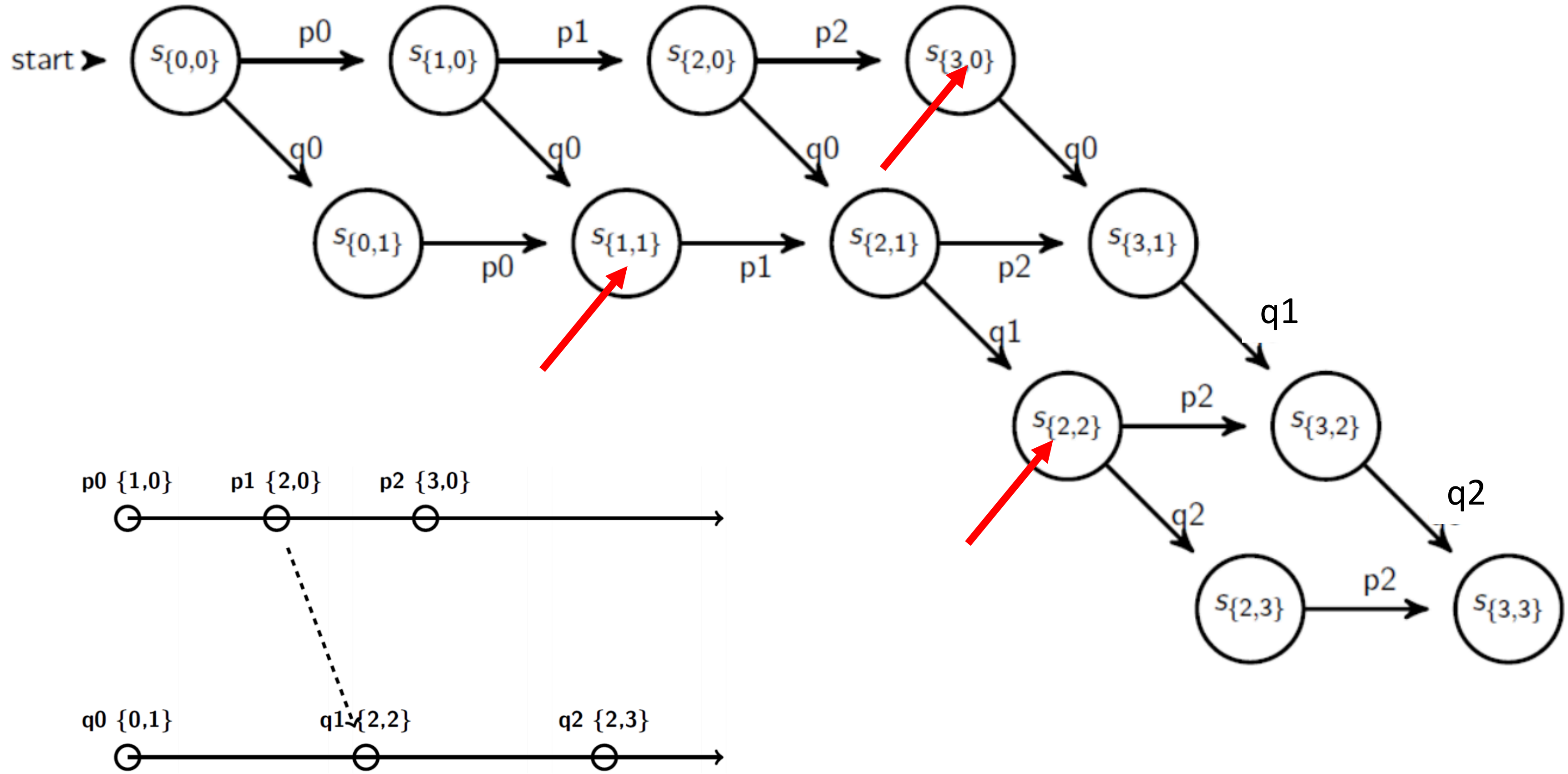
- **Examples:**
  - There is no deadlock in a distributed transaction system.
  - "Accuracy" in failure detectors: an alive process is not detected as failed.
  - No two processes decide on different values.

- A global state S_0 satisfies a **safety** property P iff:
  - $\text{safety}(P(S_0)) \equiv \forall S$ reachable from S_0, $P(S) = \text{true}$.
  - For all states S reachable from S_0, P(S) is true.

Safety Example

If predicate is true only in the marked states, does it satisfy safety?

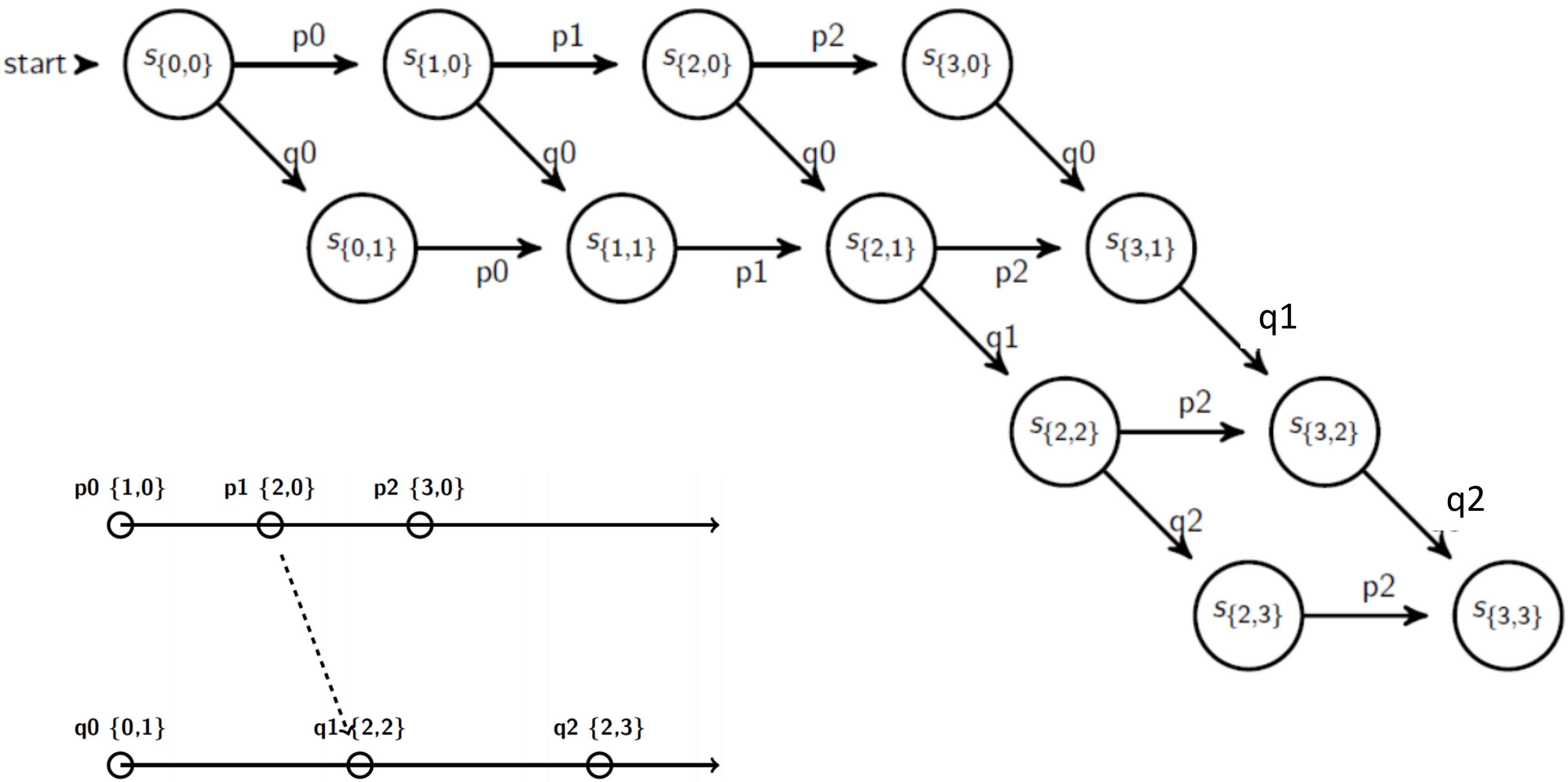
No



Safety Example

If predicate is true only in the **unmarked** states, does it satisfy safety?

Yes

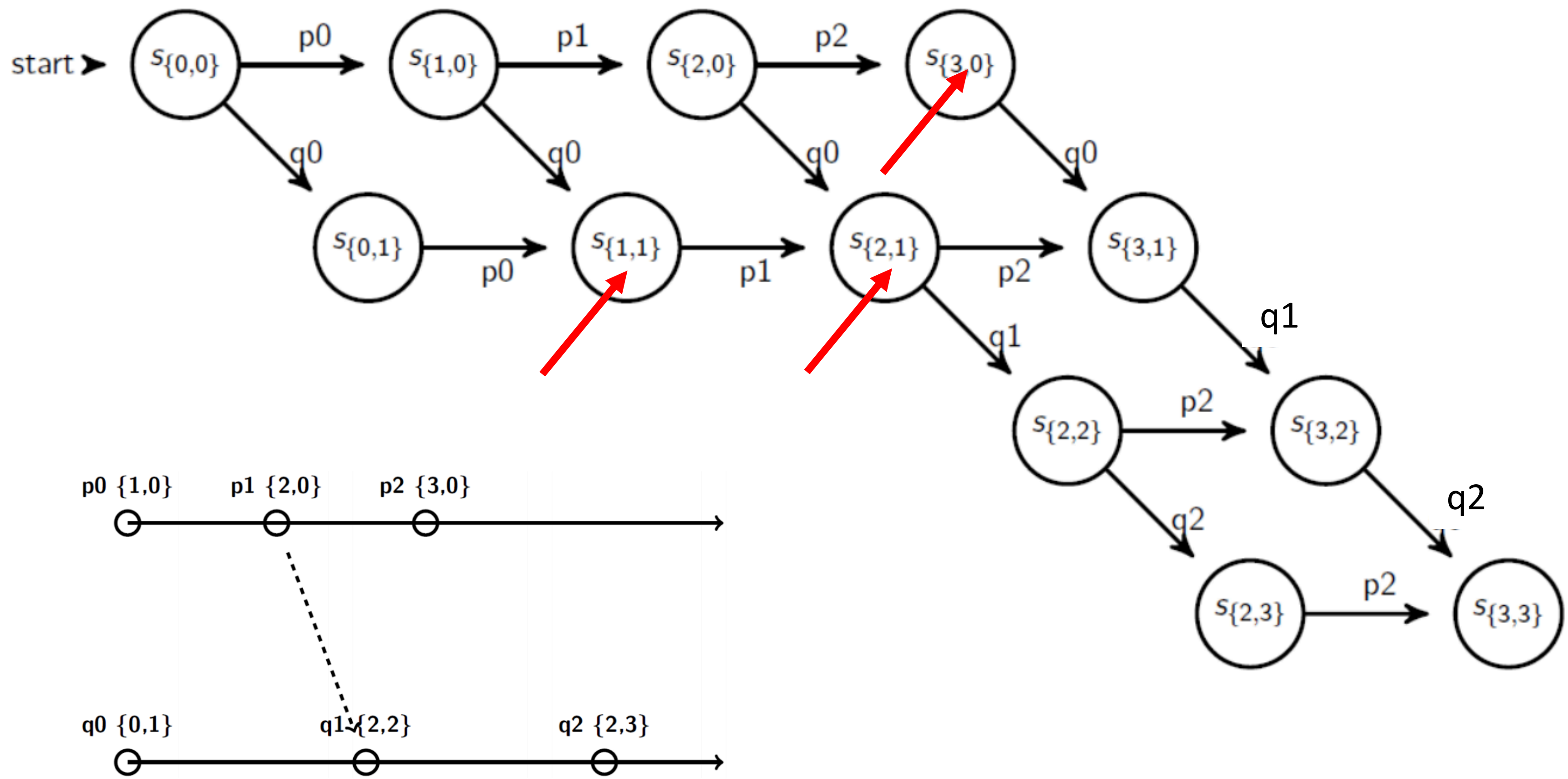


Safety

- **Safety = guarantee that something bad will never happen.**
- **Examples:**
  - There is no deadlock in a distributed transaction system.
  - "Accuracy" in failure detectors: an alive process is not detected as failed.
  - No two processes decide on different values.
- A global state S_0 satisfies a **safety** property P iff:
  - $\text{safety}(P(S_0)) \equiv \forall S$ reachable from S_0, $P(S) = \text{true}$.
  - For all states S reachable from S_0, P(S) is true.

Liveness Example

Technically satisfies liveness, but difficult to capture or reason about.



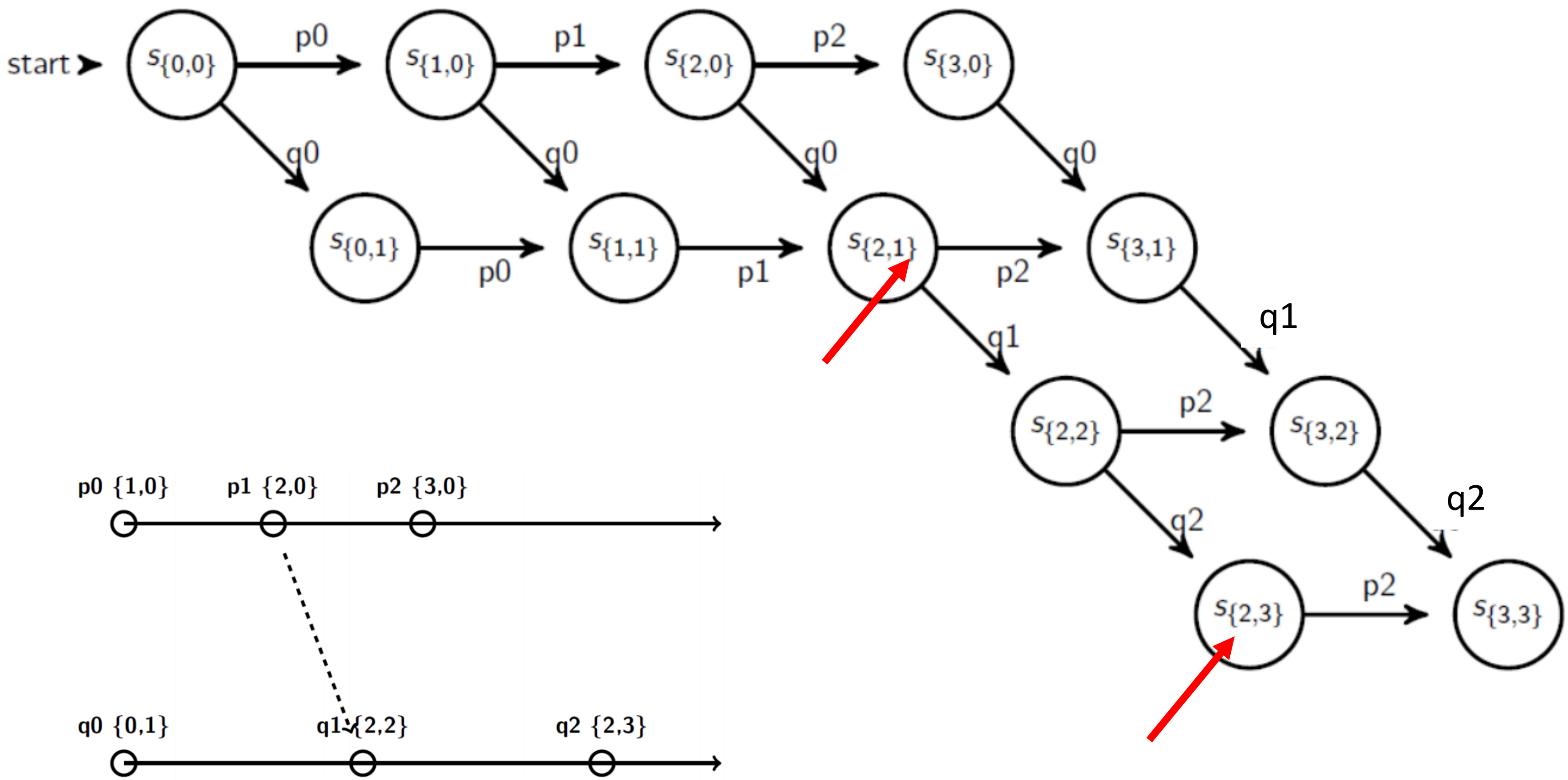
Stable Global Predicates

- once true, stays true forever afterwards (for stable liveness)

Stable Global Predicates

If predicate is true only in the marked states, is it stable?

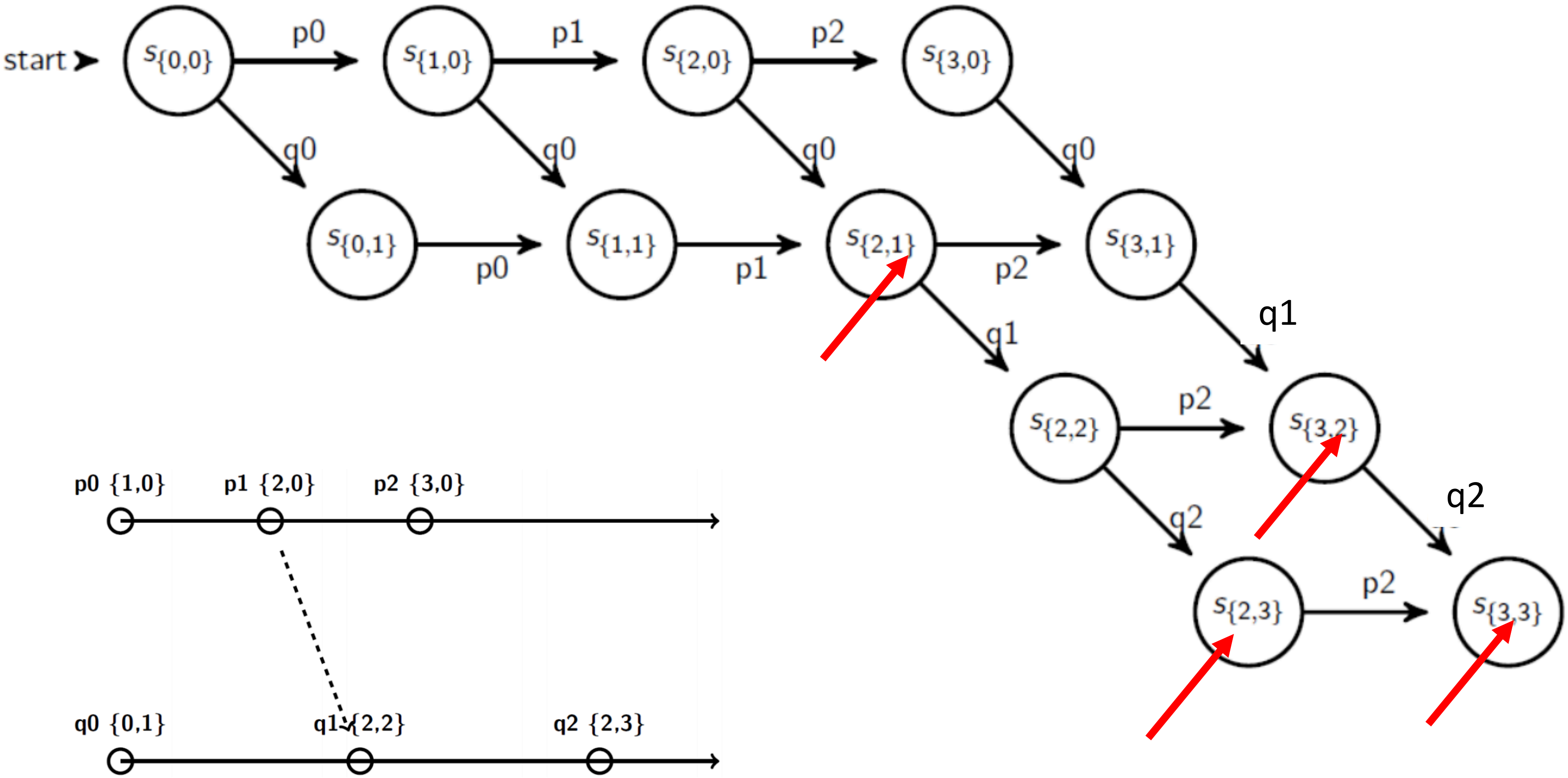
No



Stable Global Predicates

If predicate is true only in the marked states, is it stable?

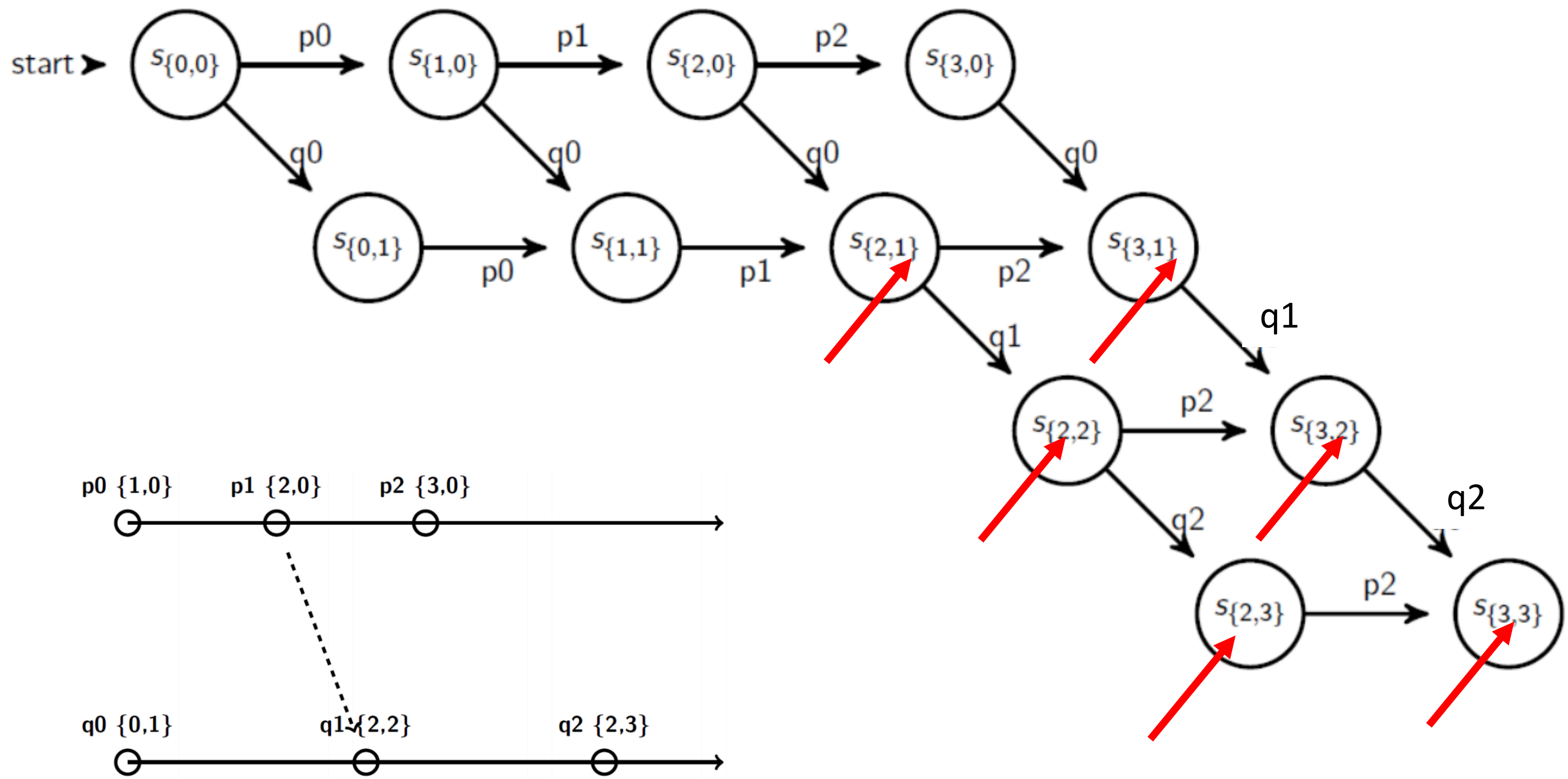
No



Stable Global Predicates

If predicate is true only in the marked states, is it stable?

Yes



Stable Global Predicates

- once true, stays true forever afterwards (for stable liveness)
- once false, stays false forever afterwards (for stable non-safety)
- Stable liveness examples (once true, always true)
  - Computation has terminated.
- Stable non-safety examples (once false, always false)
  - There is no deadlock.
  - An object is not orphaned.

- *All stable global properties can be detected using the Chandy-Lamport algorithm.*

Global Snapshot Summary

- The ability to calculate global snapshots in a distributed system is very important.
- But don't want to interrupt running distributed application.
- Chandy-Lamport algorithm calculates global snapshot.
- Obeys causality (creates a consistent cut).
- Can be used to detect global properties.
- Safety vs. Liveness.