Distributed Computing

José Orlando Pereira

Grupo de Sistemas Distribuídos
Departamento de Informática
Universidade do Minho

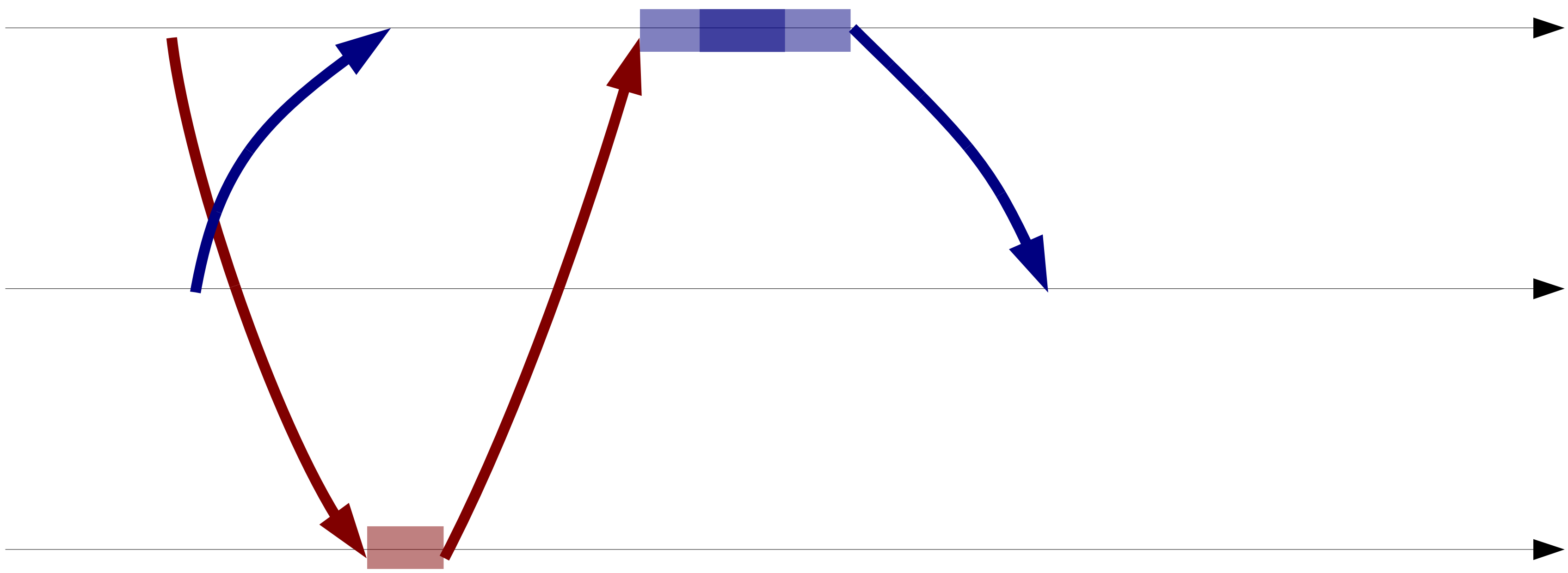
2009/2010

Example: Distributed deadlock

- Remote invocation with single threaded dispatching:
  - All processes request and reply to invocations
  - When waiting for a reply, cannot handle requests
- Distributed deadlock possible when multiple processes invoke each other

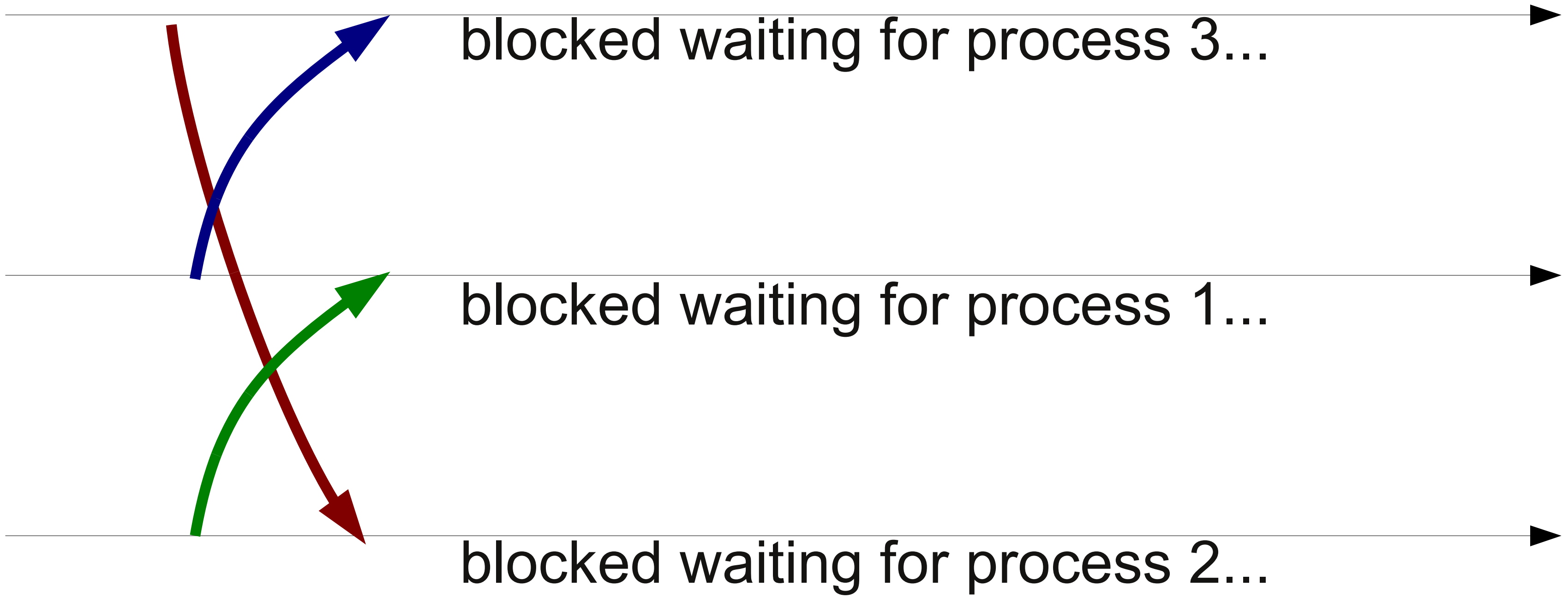
Example: Distributed deadlock

- Deadlock-free run:



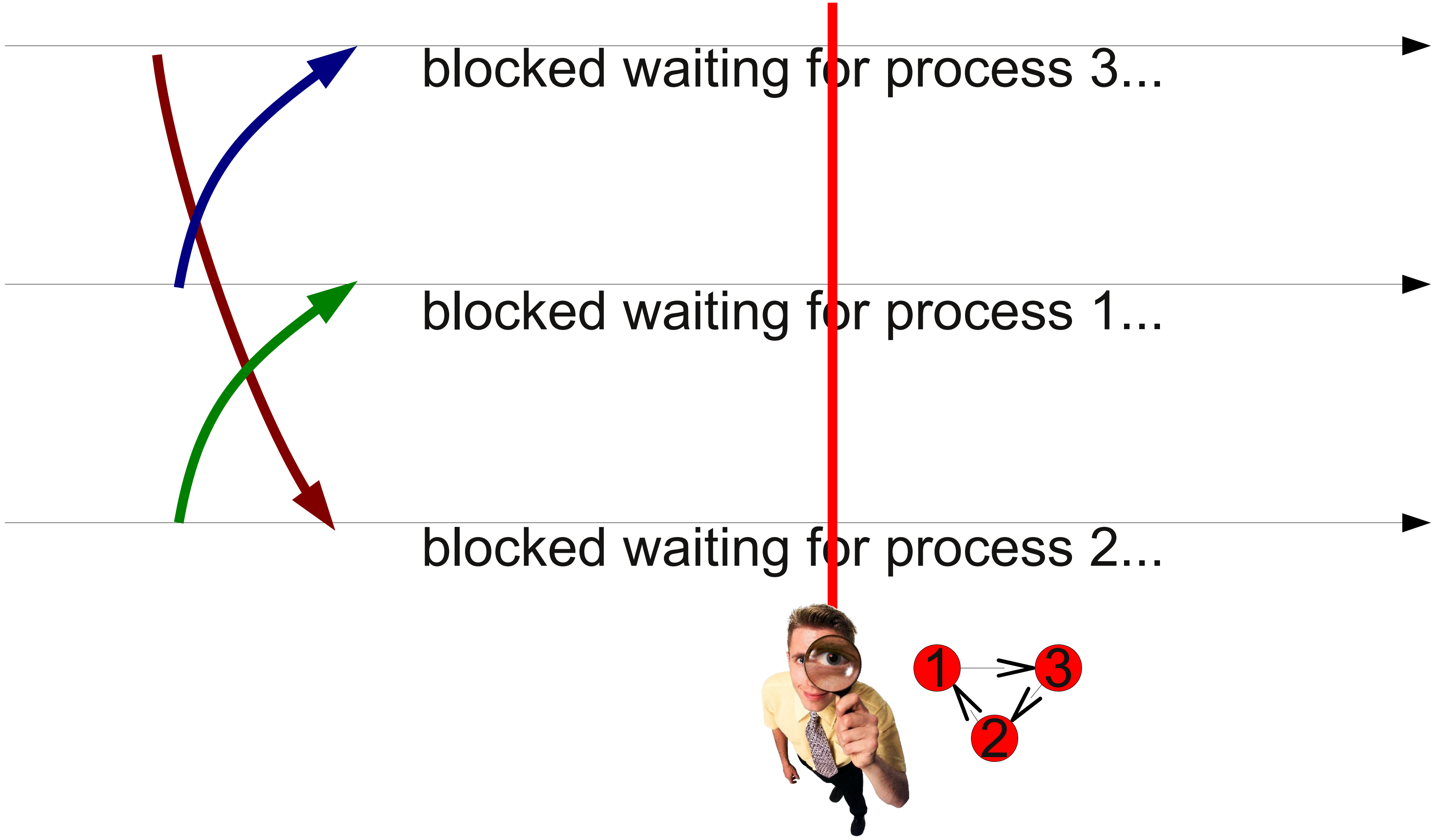
Example: Distributed deadlock

- Distributed deadlock:



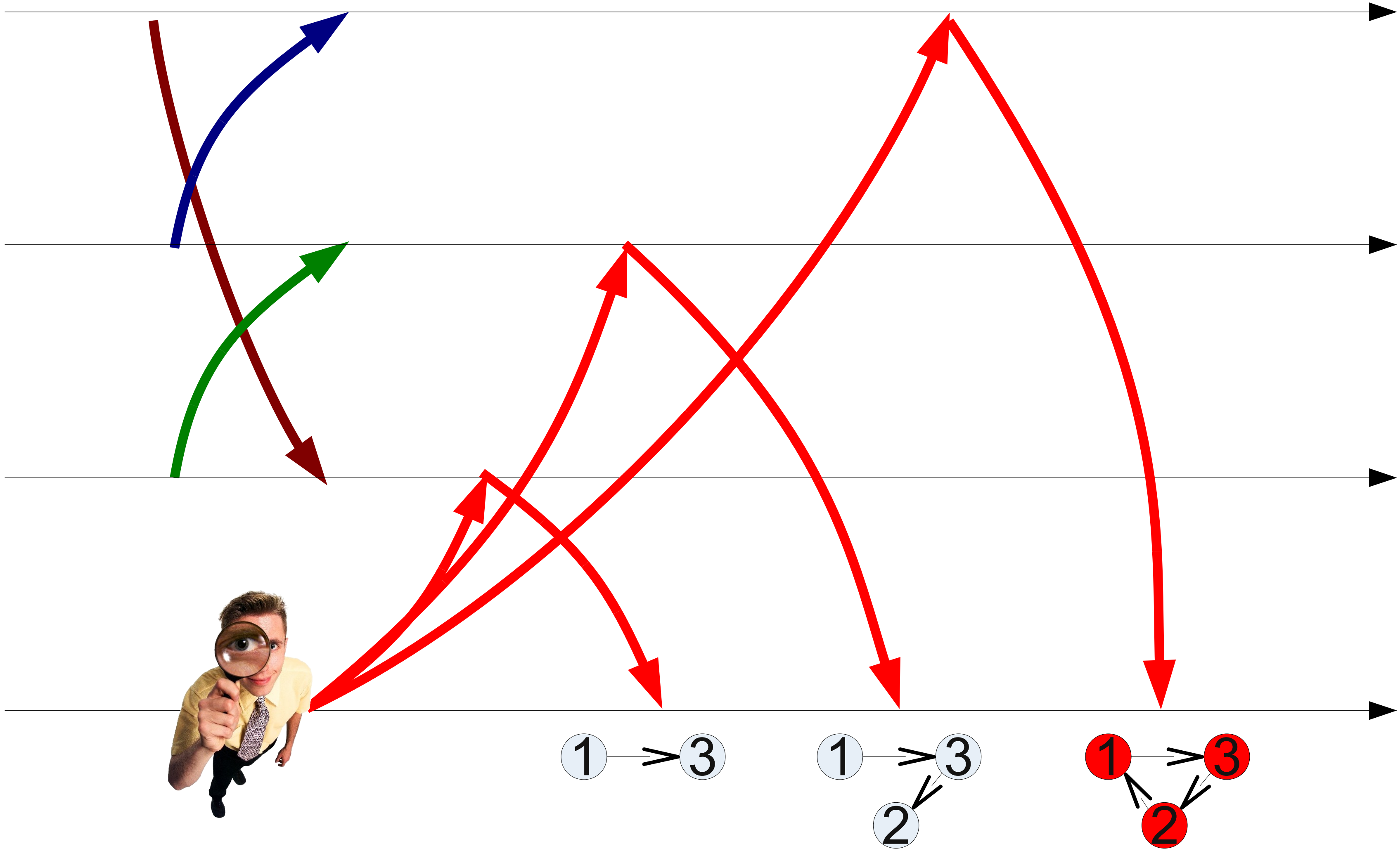
Example: Distributed deadlock

- Instant observation is impossible:



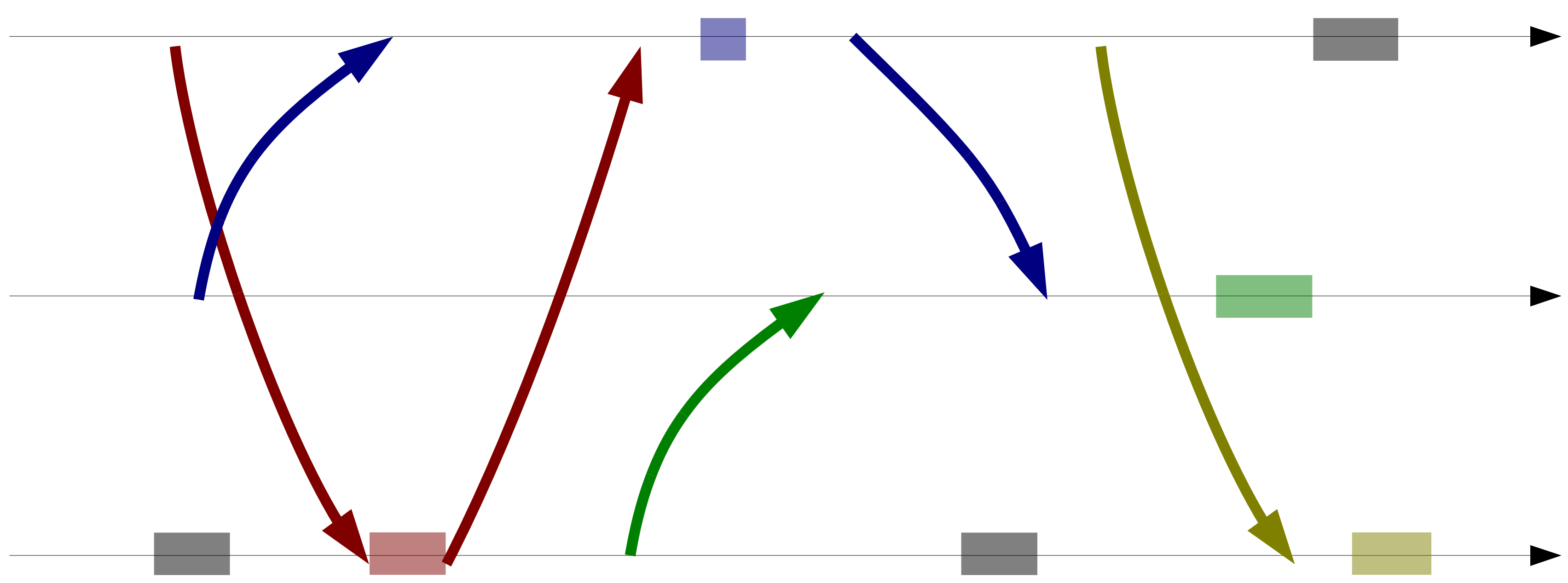
Example: Distributed deadlock

- Deadlock detection with a “wait for” graph:



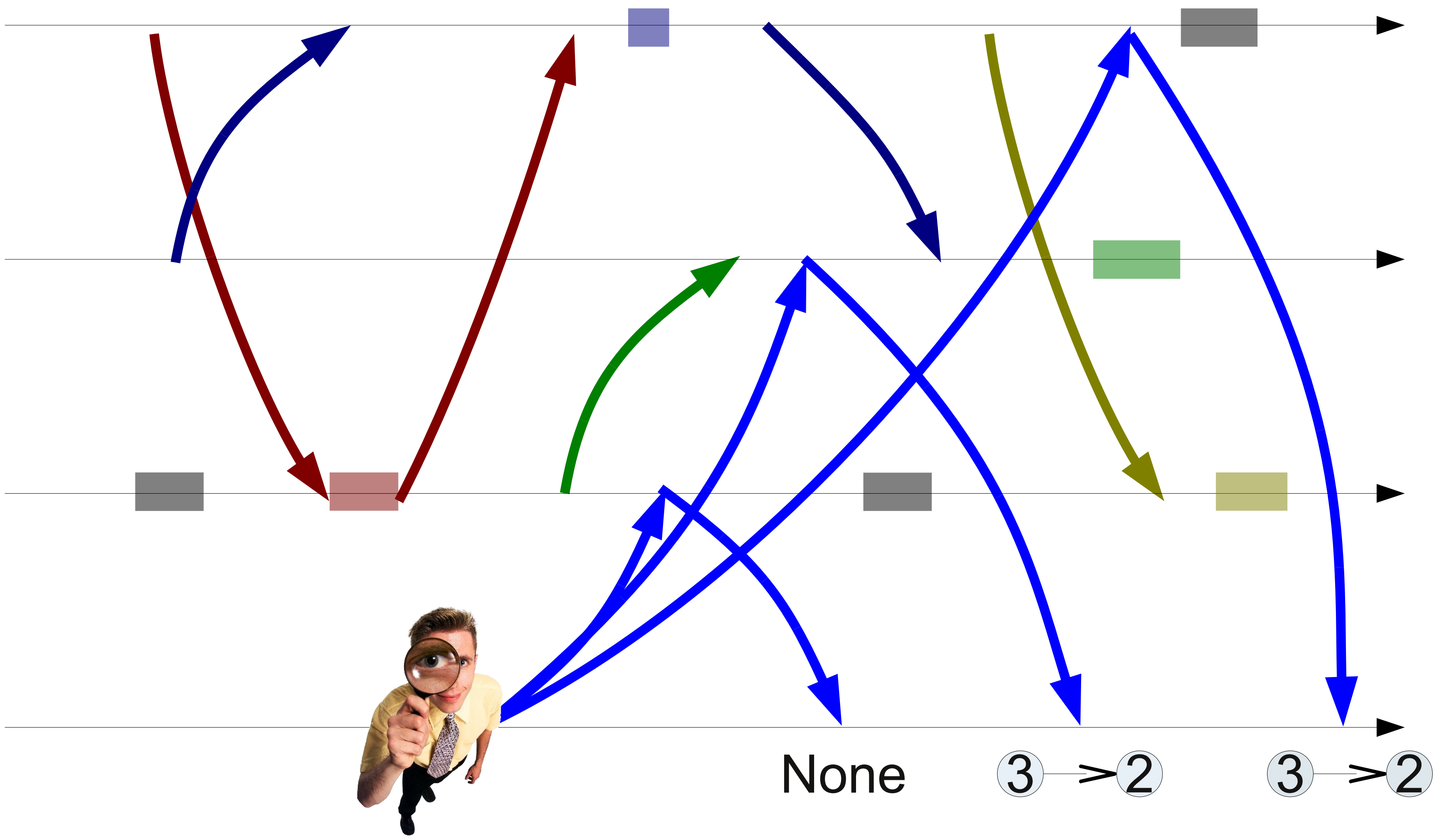
Example: Distributed deadlock

- A more complex deadlock-free run:



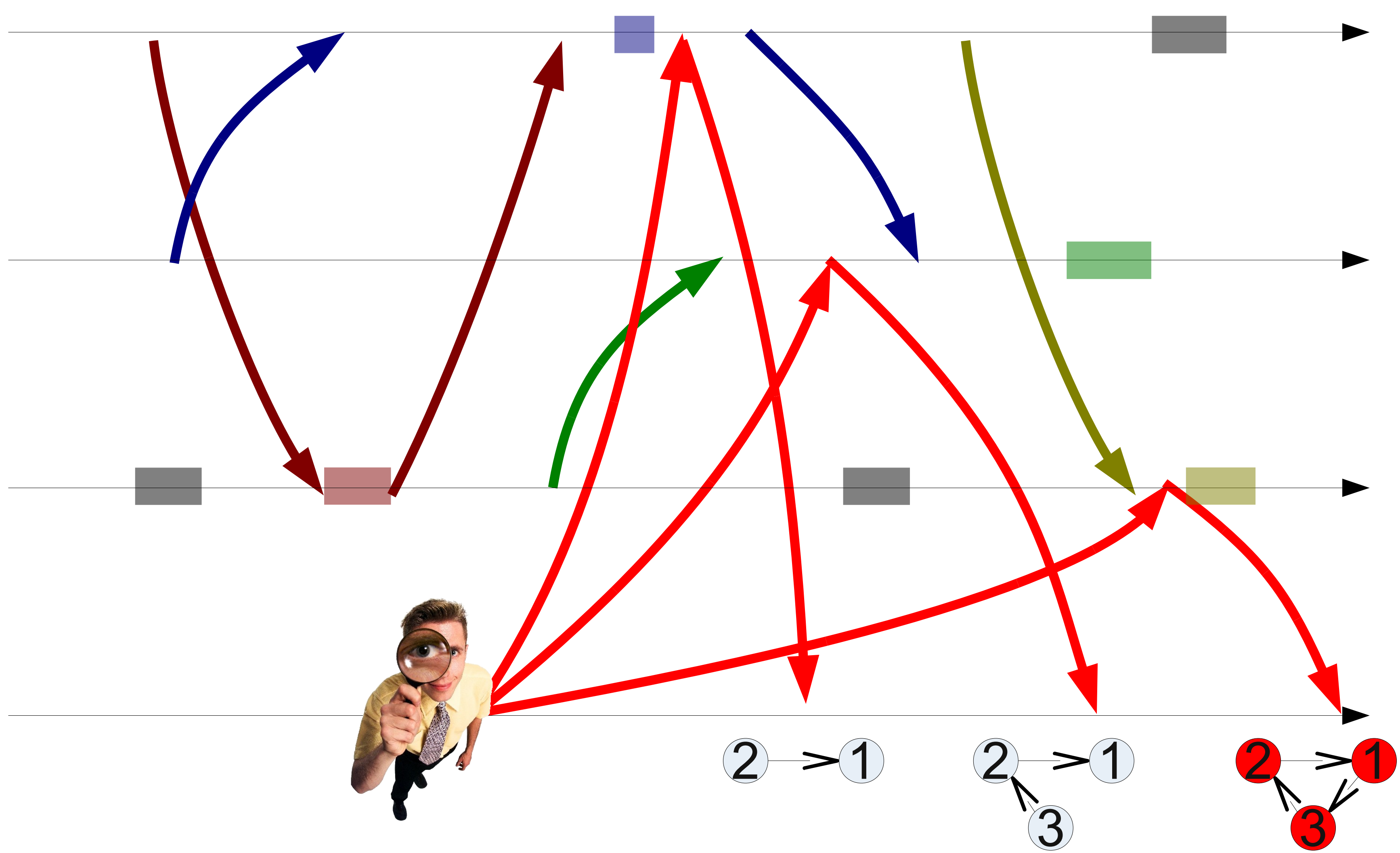
Example: Distributed deadlock

- A deadlock-free WFG:



Example: Distributed deadlock

- A WFG with a ghost deadlock:



Other examples

- Garbage collection: Discovering abandoned objects
- Debugging: Breakpoints in a distributed program
- Checkpointing and restarting: Creating a backup of a distributed application's state
- ...

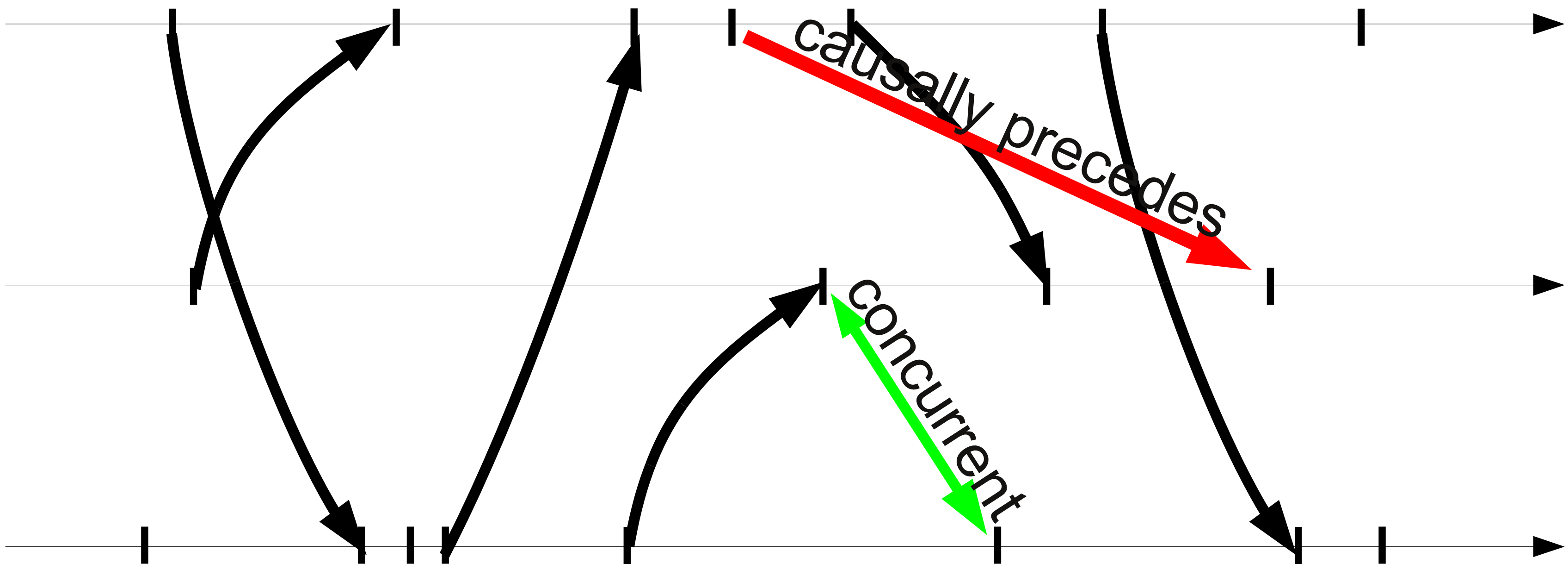
Global Property Evaluation

- All these problems are instances of the Global Property Evaluation (GPE) problem
- Can it be solved in an asynchronous system?
- Methods that can be used? Relative cost?

Causality

- Events i and j are causally related ($i \rightarrow j$) iff:
  - i precedes j in some process p
  - for some m, i=send(m) and j=receive(m)
  - for some k, $i \rightarrow k$ and $k \rightarrow j$ (transitivity)
- Events i and j are concurrent ($i || j$) iff neither $i \rightarrow j$ or $j \rightarrow i$

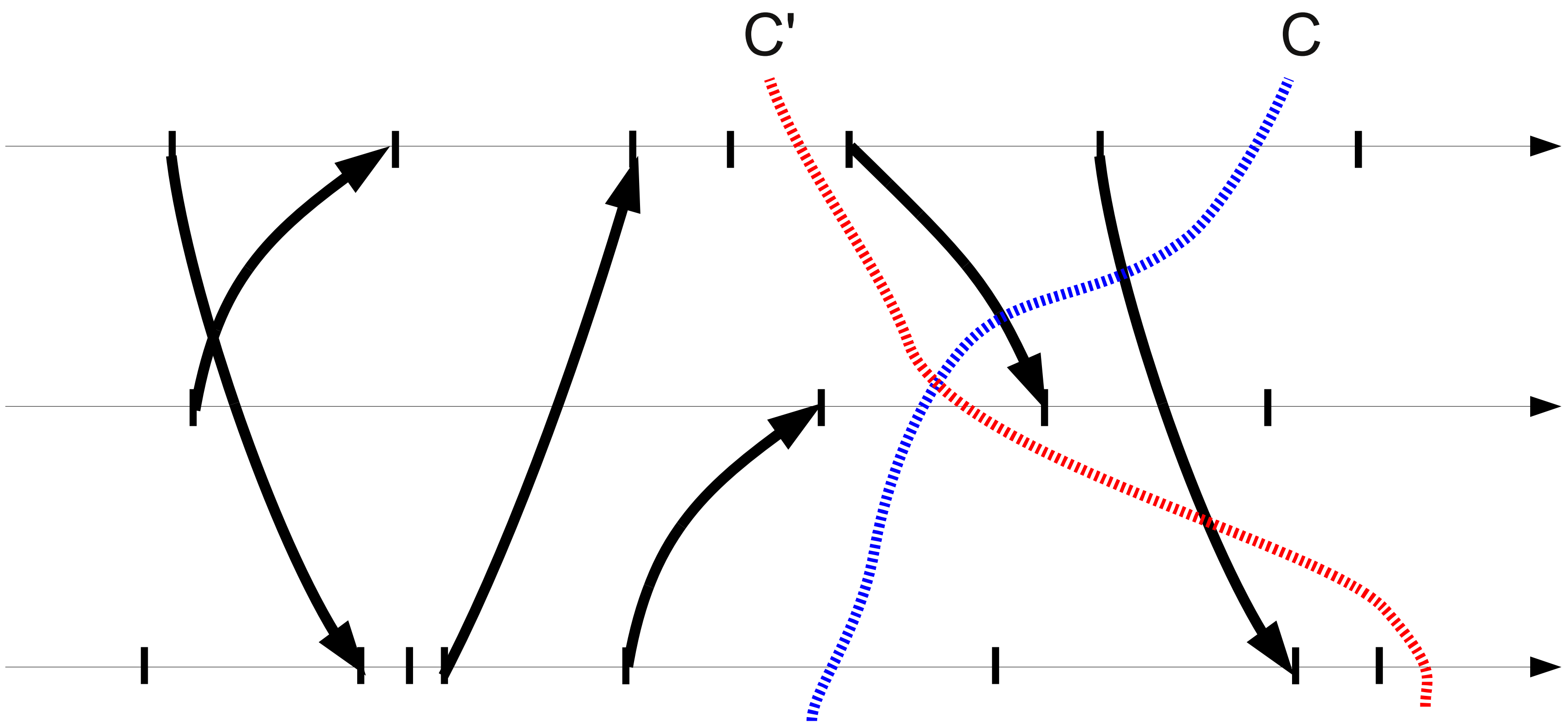
Causality



Cuts and consistency

- A cut is the union of prefixes of process history
- A consistent cut includes all causal predecessors of all events in the cut
- Intuitive methods:
  - If a cut is an instant, there are no messages from the future
  - In the diagram, no arrows enter the cut
  - All events in the frontier are concurrent

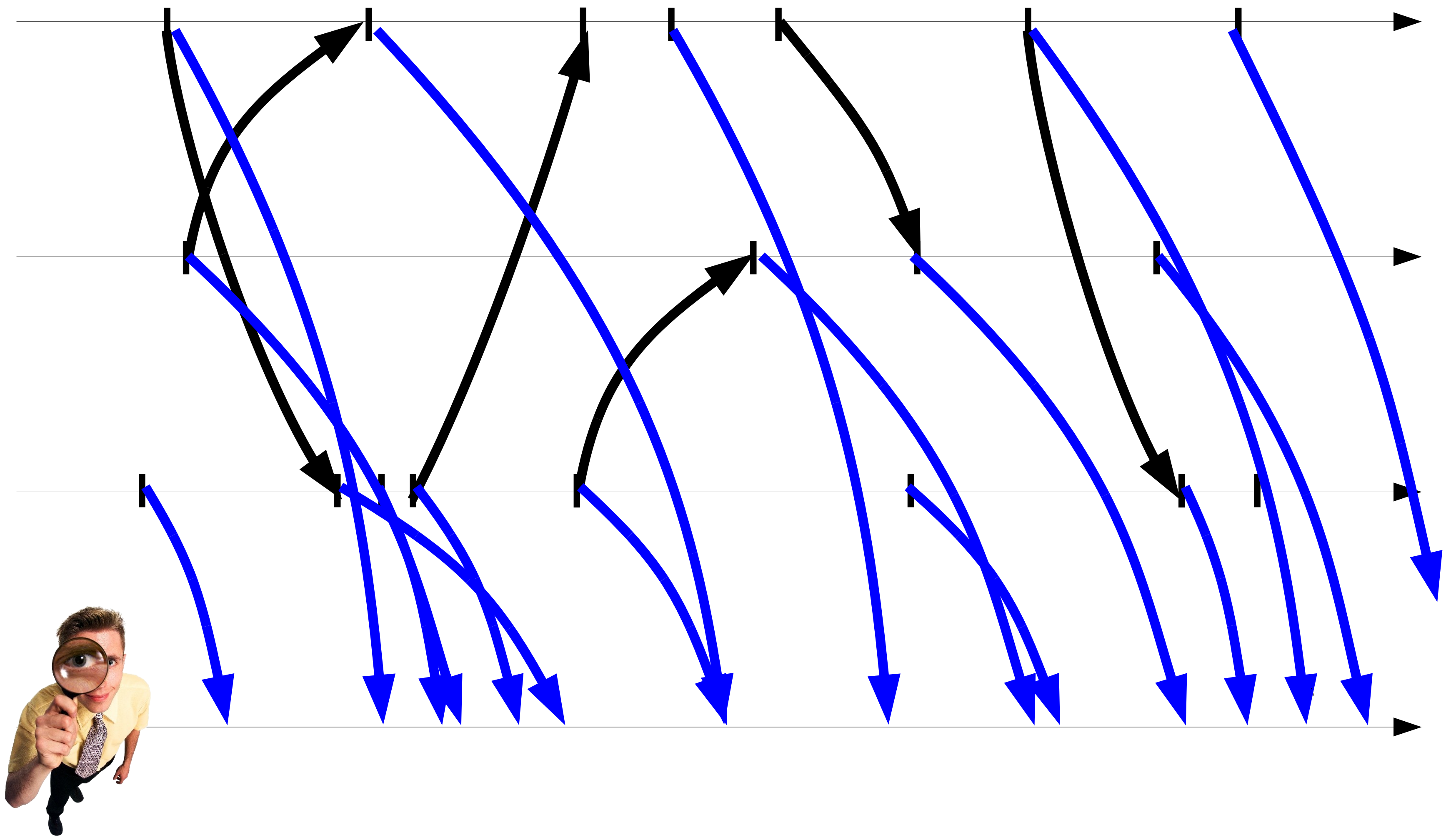
Cuts



- GPE possible only in consistent cuts!

Passive monitor process

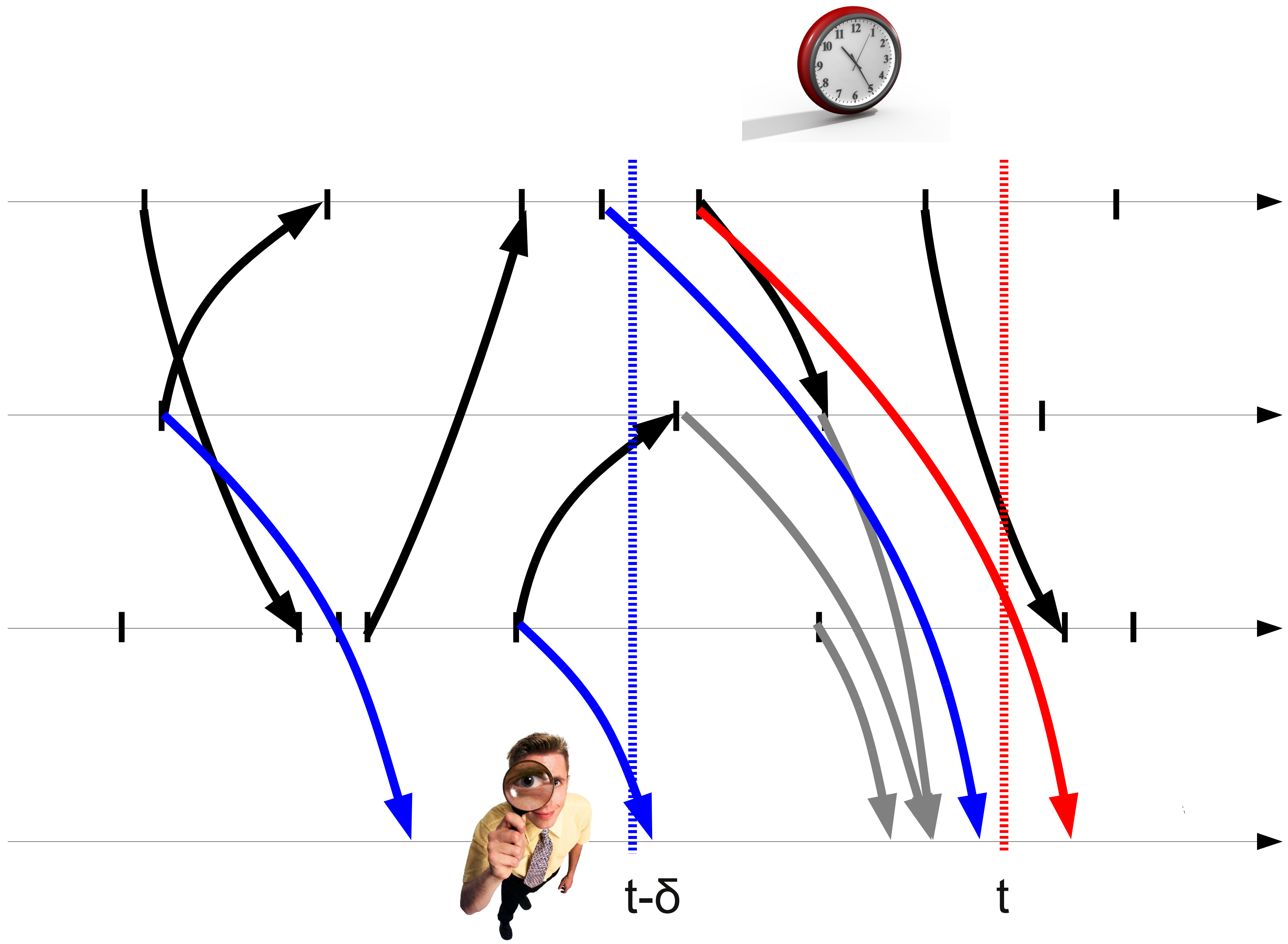
- Report all events to monitor:



First try: Synchronous system

- Global clock, δ upper bound on message delay
- Tag events with real time
- Consider events only up to t-δ

First try: Synchronous system



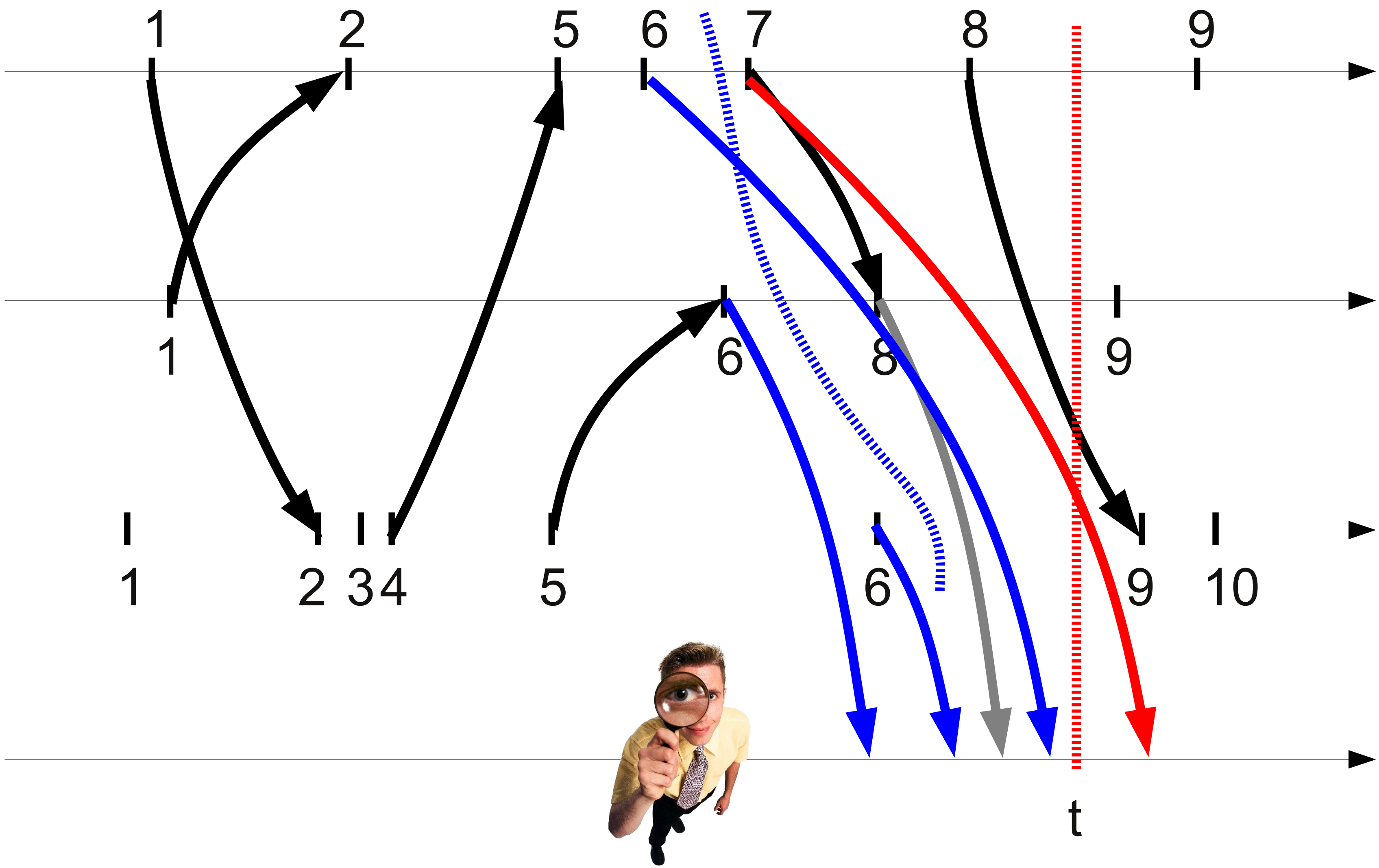
Clock properties

- Consider RC(i) the time at which i happened
- If i→j then RC(i)<RC(j)
- For some event j:
  - When we are sure that there is no unknown i such that RC(i)<RC(j)
  - Then there is no j such that j→i
- Can we build a logical clock with the same property?

Second try: Logical clock

- Tag events as follows:
  - Local events: increment counter
  - Send events: increment and then tag with counter
  - Receive events: update local counter to maximum and then increment
- Use FIFO channels
- Consider events only up to the minimum of maximum tags

Second try: Logical clock



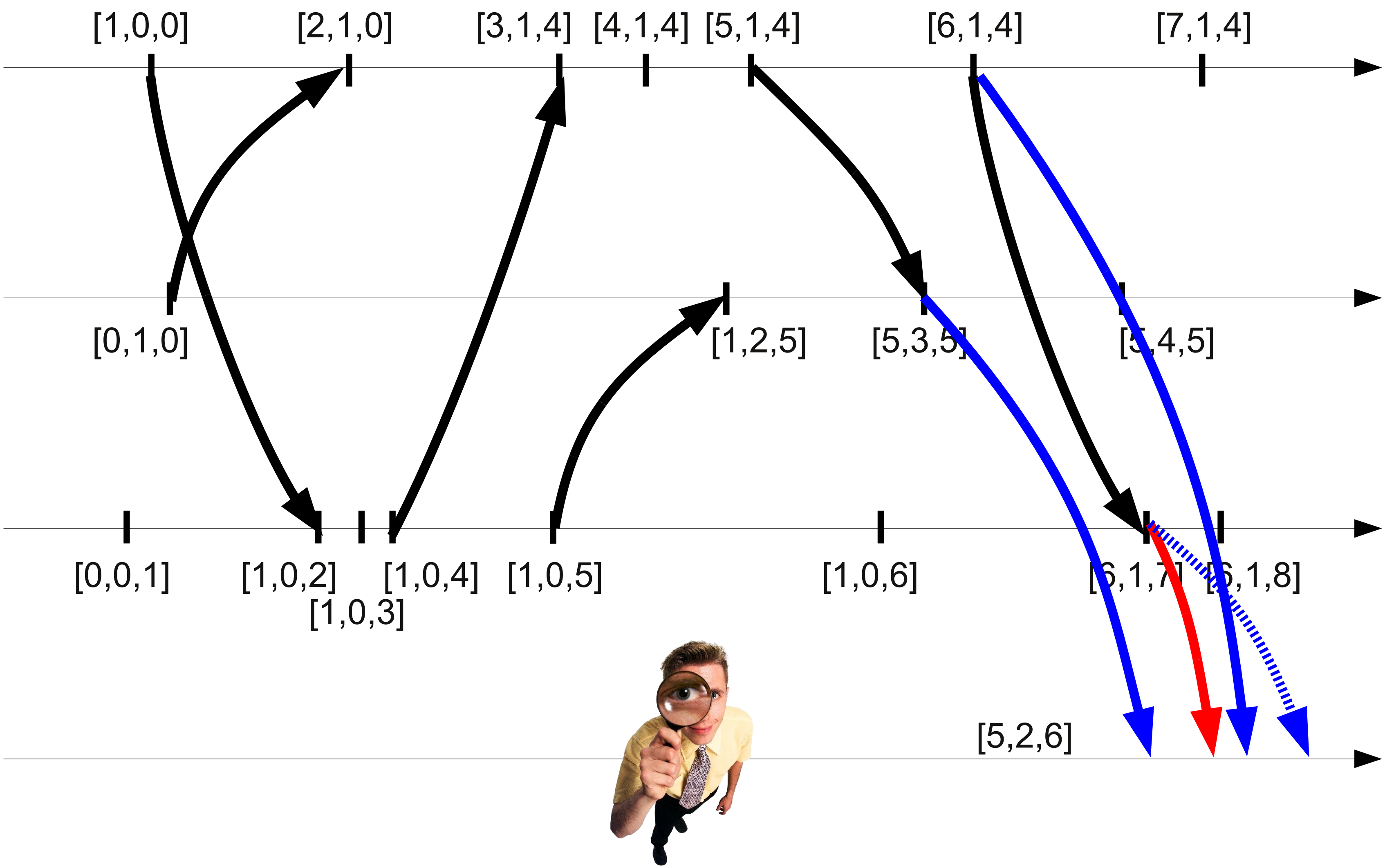
Scalar clocks

- Synchronous system (RC):
  - Delay δ to consistency
- Asynchronous system (LC):
  - Possible unbounded delay to consistency
  - Blocks if some process stops sending messages

Third try: Vector clock

- Tag events with a vector as follows:
  - Local event at i: increment counter i
  - Send event at i: increment counter i and tag with vector
  - Receive event at i: update each counter to maximum and increment counter i

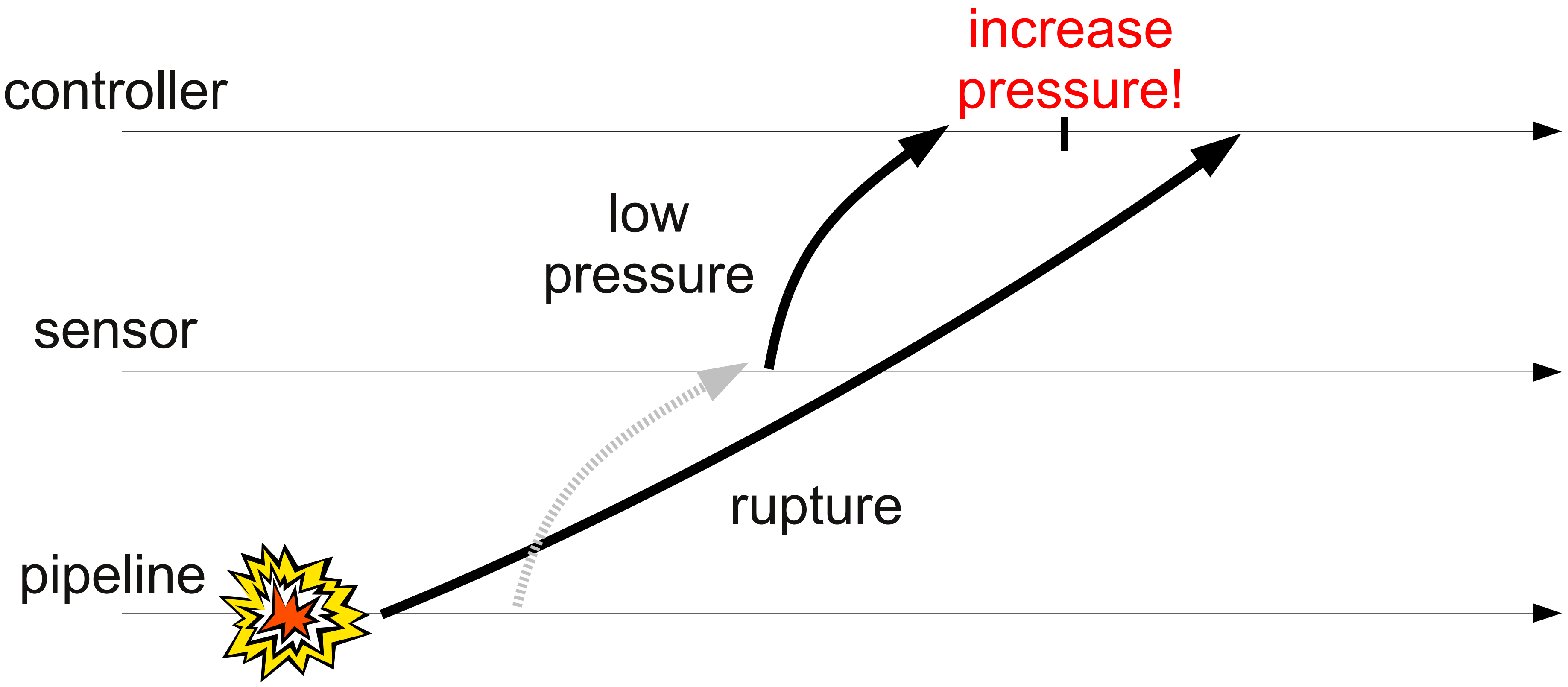
Third try: Vector clock



Causal delivery

- The monitor delivers events as follows:
  - With local vector l[...]
  - For some r[...] from i
  - Wait until:
    - l[i]+1=r[i]
    - For all j≠i: r[i]≤l[i]
- The monitor is always in a consistent cut

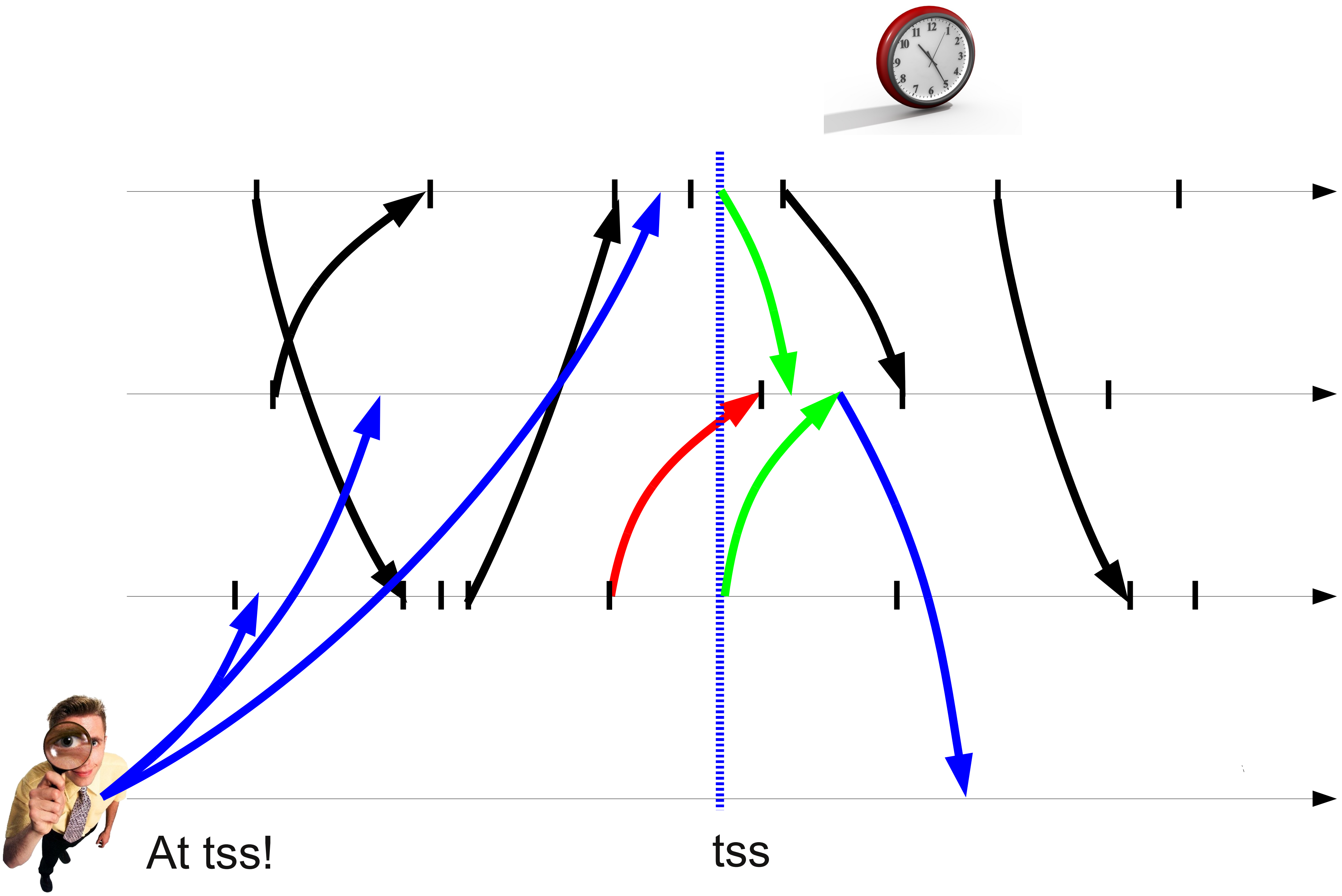
Hidden channels



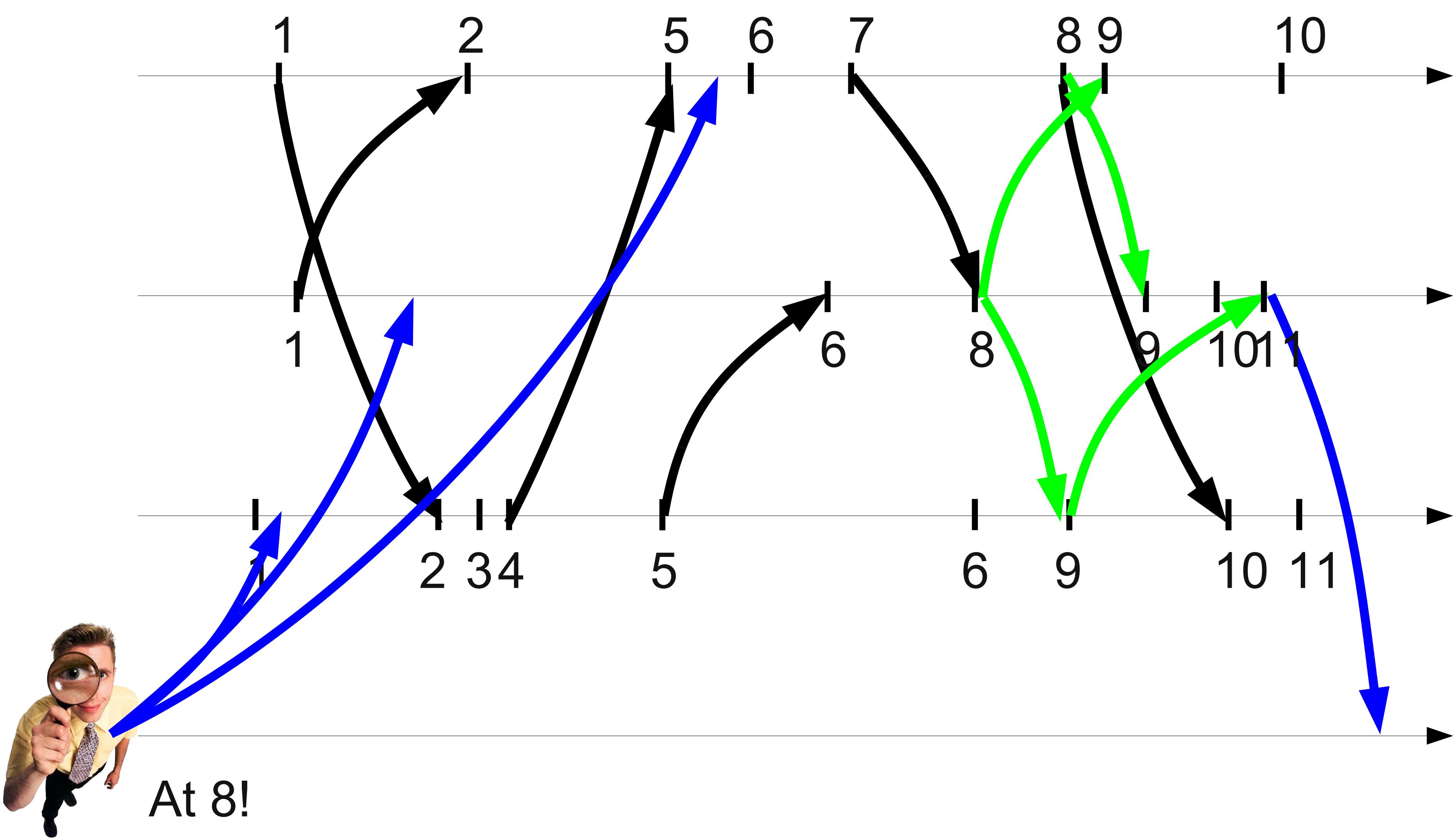
Fourth try: No reporting, synchronous

- Reporting all events to a monitor causes a large overhead
- Consider a monitor within the computation
- Monitor broadcasts tss in the future
- At tss, each process:
  - Records state
  - Sends messages to all others
  - Starts recording messages until receiving a message with RC > tss
- After stopping, sends all data to monitor

Fourth try: No reporting, synchronous



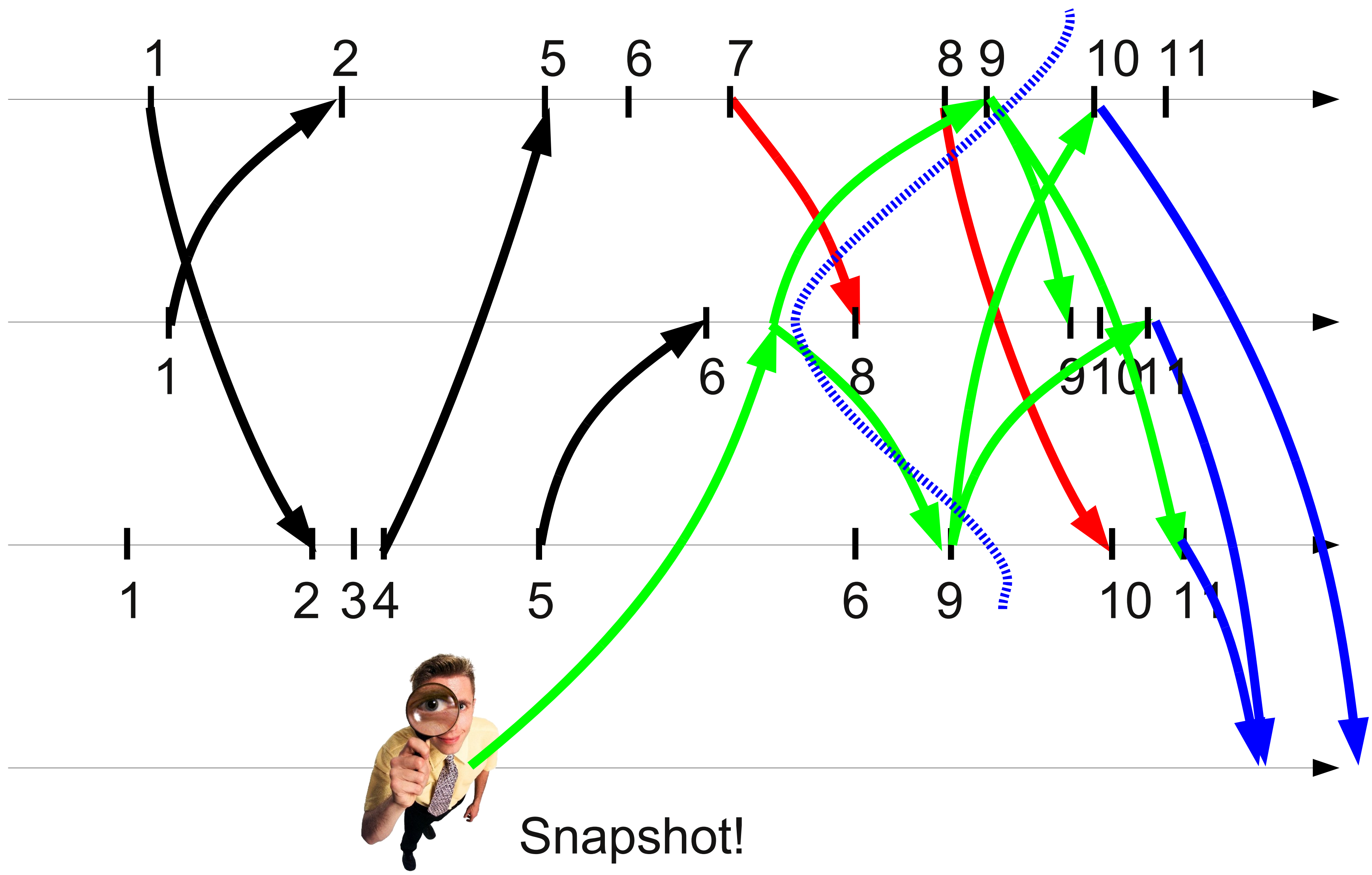
Fifth try: No reporting, logical clock



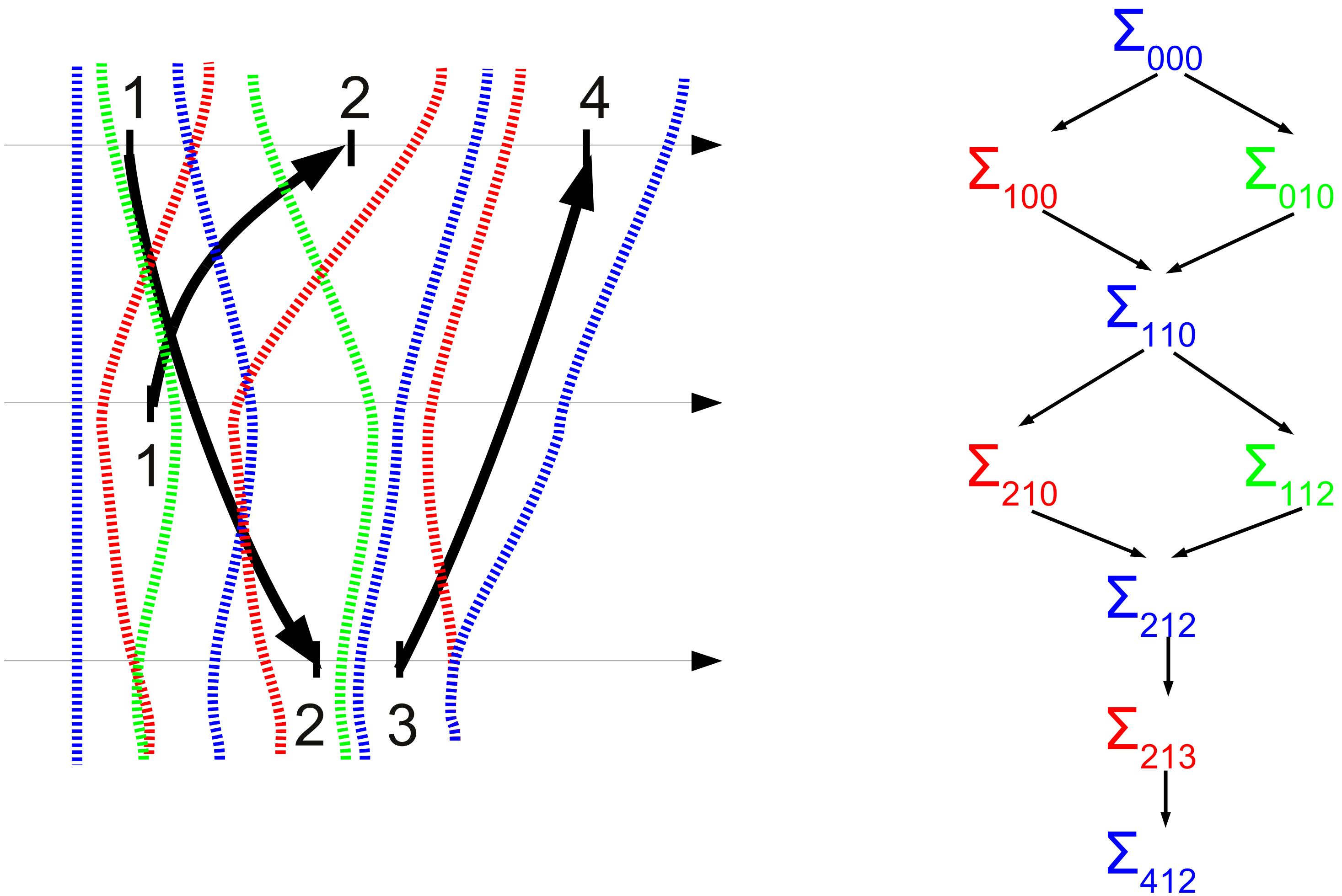
Chandy and Lamport

- Send a "Snapshot" message to some process
- Upon receiving for the first time:
  - Records state
  - Relays "Snapshot" to all others
  - Starts recording on each channel until receiving "Snapshot"
- Send all data to monitor

Chandy and Lamport

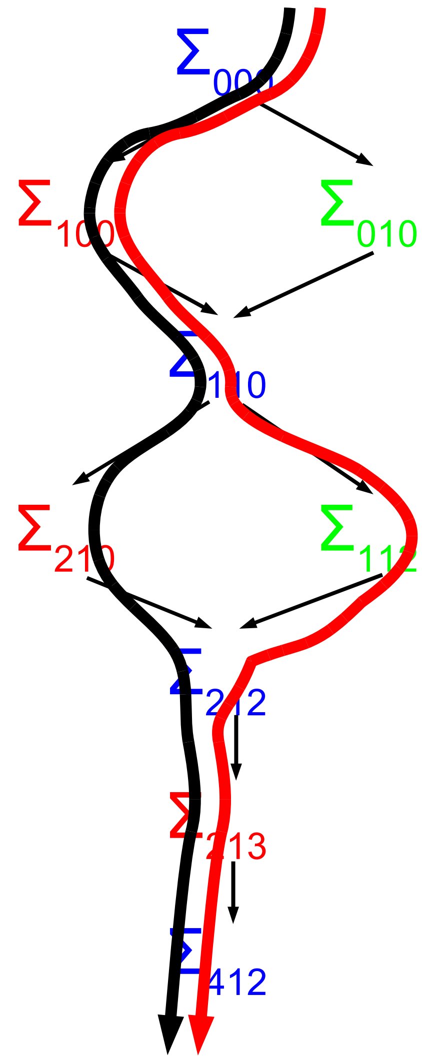


Consistent global states



Consistent global states

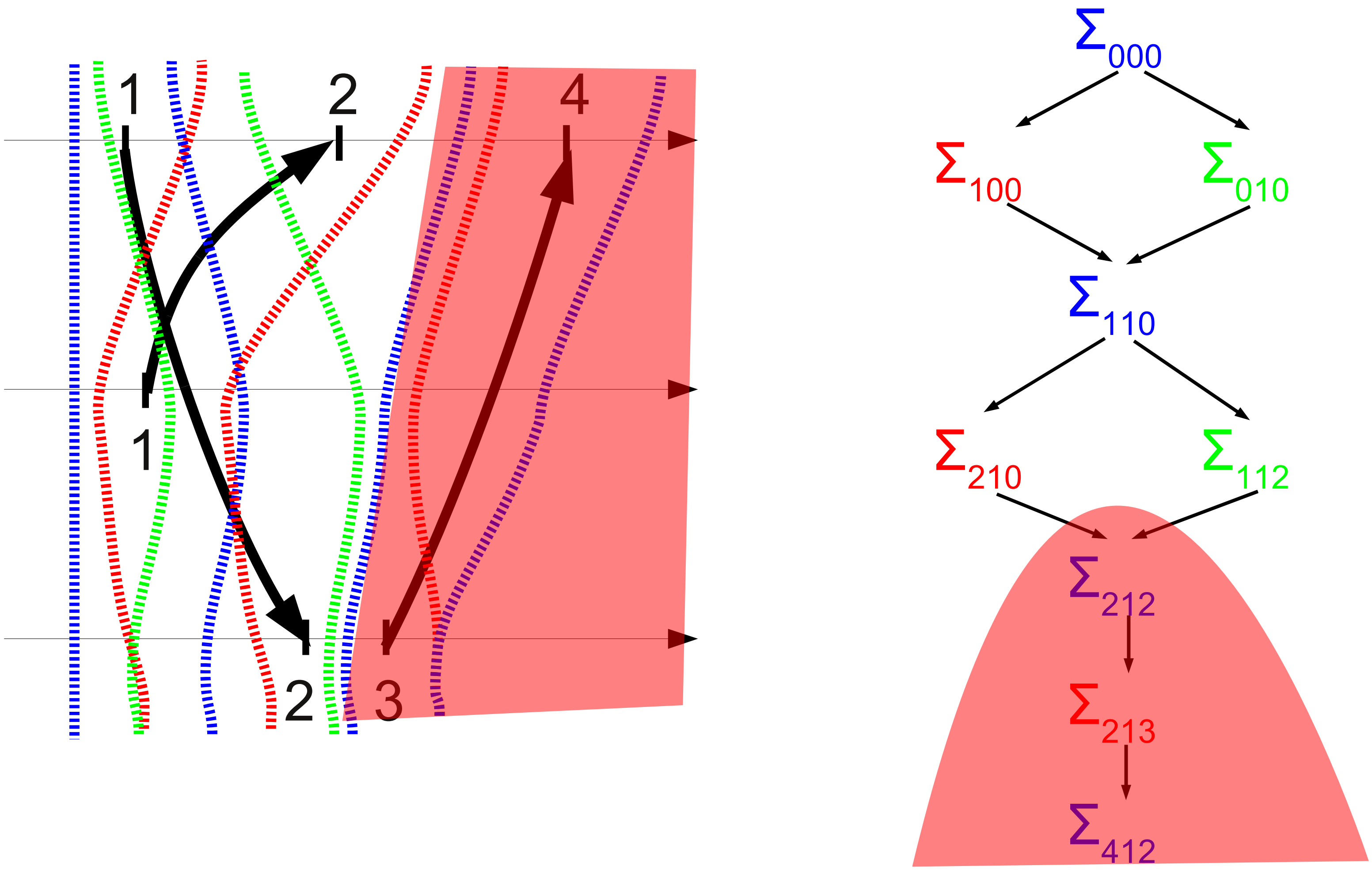
- Includes the true sequence of states in the system
- An observer within the system cannot deny any of the possible paths



Stable predicates

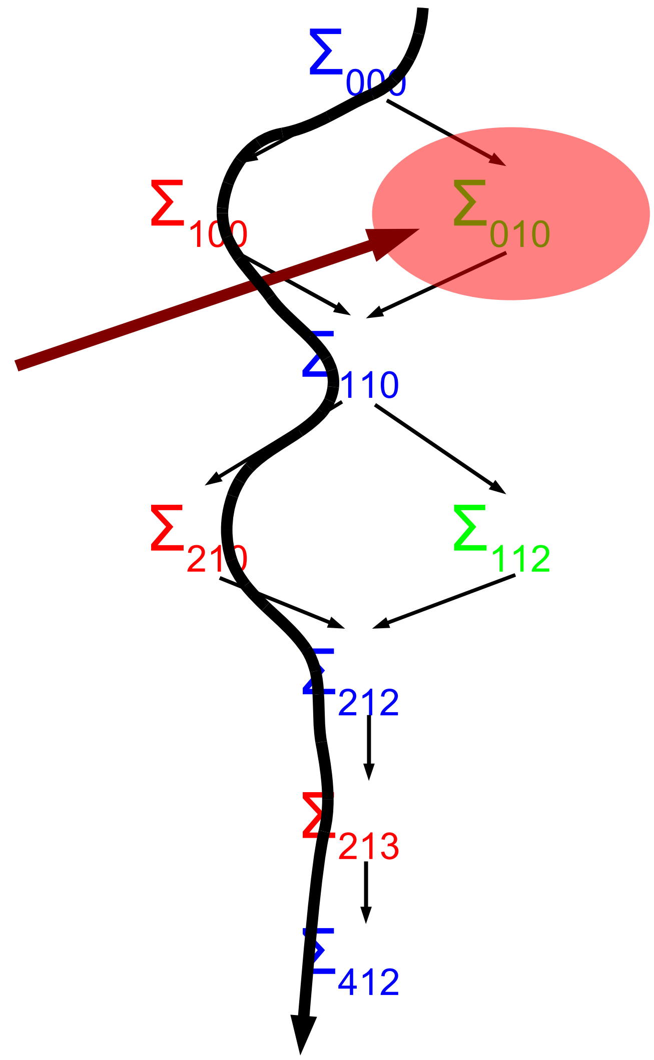
- Once true, always true
- Examples:
  - Deadlock detection
  - Termination
  - Loss of token
  - Garbage collection
- Can be evaluated periodically on snapshots

Stable predicates



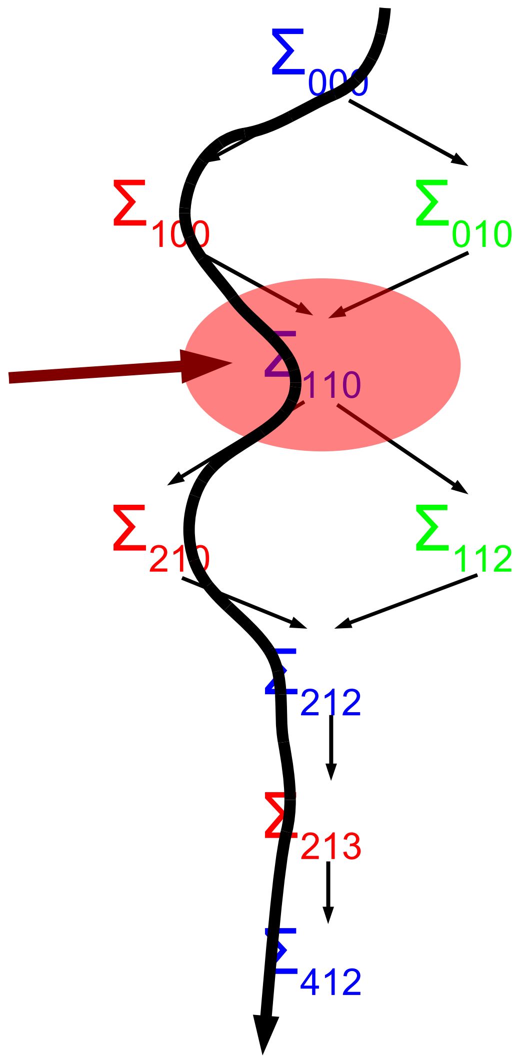
Nonstable predicates

- True in a subset of observable states
- Some are possibly true: an observer in the system cannot deny having been true
- The predicate does not hold on some paths



Nonstable predicates

- True in a subset of observable states
- Some are <u>definitely true</u>: an observer in the system is sure of having been true
- The predicate holds on all possible paths



Nonstable predicates

- Examples:
  - Total size of queues in the system
  - Number of messages in transit
  - Amount of memory used
- Can be detected by full monitoring of all (relevant) events

References

- O. Babaoglu and K. Marzullo, “Consistent Global States of Distributed Systems: Fundamental Concepts and Mechanisms”. In Sape Mullender (Ed.), *Distributed Systems*, 2nd Edition, Addison-Wesley, 1993.
- N. Lynch, “Distributed Algorithms”. Ch. 19 and 20. Morgan-Kaufmann, 1996.