Time and Coordination in Distributed Systems

INF 5040 autumn 2015

lecturer: Roman Vitenberg

We live in an asynchronous world

- Each entity or process has its own pace
  - Computers, OS processes, mobile devices, vehicles, satellites, planets
- Communication latencies are not decreasing over time
- Clocks are inherently unsynchronized
  - Perfect synchronization is theoretically impossible
- Gets worse as the scale increases
  - Nuisance for people, issue for node clusters, big issue for satellites, ultimate obstacle for inter-planet communication
- Yet, many applications require synchronization
  - Timeouts, video, audio, GPS/Galileo, air traffic control, …

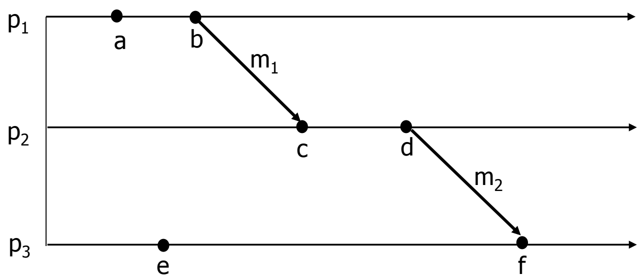
Time in Distributed Systems

- Uses of time
  - Real-time synchronization
  - Relative order of events
    - The only way to infer in an asynchronous system is through causality
- Logical time
  - Attempts to capture dependencies due to message exchange and local process ordering
    - Possible false positives
    - Does not capture dependencies that are due to a cause other than message exchange

"Happened-before" relation

- Notation
  - $x \rightarrow^{p} y$: x happened before y at process p
  - $x \rightarrow y$: x happened before y
- Condition 1
  - If ∃ process p : $x \rightarrow^{p} y$, then $x \rightarrow y$
- Condition 2
  - For each process m : send(m) → rcv(m)
- Condition 3
  - If x, y, and z are events such that $x \rightarrow y$ and $y \rightarrow z$, then $x \rightarrow z$

"Happened-before" illustrated

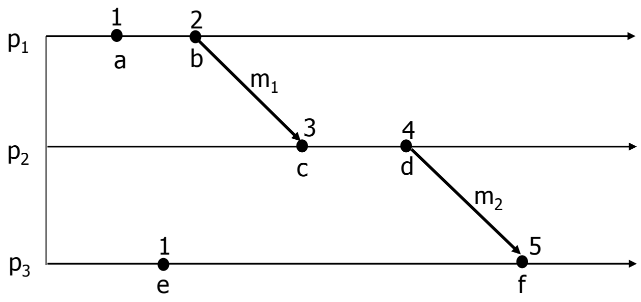


- Events that are not related by the "happened-before" relation are called concurrent: a || e

Logical clock

- Each process p maintains its own logical clock C_p
  - Monotonically increasing counter
  - Used to timestamp events
- C_p(a) : the timestamp of event a at process p
- Rules for logical clock
  - LC1:
    - C_p is incremented by 1 before each event is issued at process p
  - LC2:
    - When a process p sends a message m, it piggybacks C_p on m
    - When (m,t) is received by q, q computes C_q:=max(C_q,t) and applies LC1 before timestamping the event rcv(m).

Example for Logical Clocks

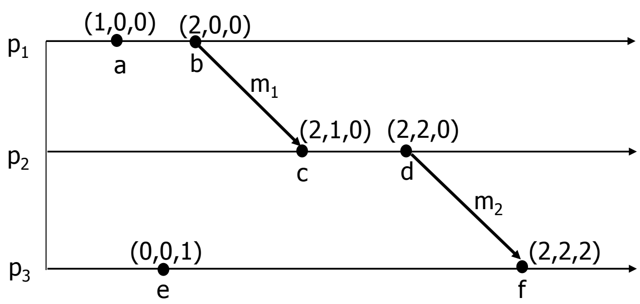


$x \rightarrow y \Rightarrow C(x) < C(y)$ (not equivalent!!)

Vector clocks

- Assumption: N processes whose ids are totally ordered
- Each process `p` maintains its vector clock V_p of size N
- $V_p(a)$: the timestamp of event `a` at process `p`
- Rules for vector clock
  - VC1: $V_p[j]$ is initially 0 for all j
  - VC2:
    - `p` sets $V_p[p] := V_p[p]+1$ before timestamping each event
  - VC3:
    - When `p` sends a message `m`, it piggybacks V_p on `m`
  - VC4: When `(m,t)` is received by `q`, `q` computes $V_q[j] := max(V_q[j], t[j])$ for all j and applies VC2

Example for Vector Clocks



$x \rightarrow y \Leftrightarrow V(x) < V(y)$ (equivalent)

INF5040, Roman Vitenberg 9

Local events and states

The history (h) of a process is modelled as a sequence of events and corresponding states:

$$h_i = s_i^0, e_i^1 \leftrightarrow s_i^1, e_i^2 \leftrightarrow s_i^2, e_i^3 \leftrightarrow s_i^3, \cdots$$

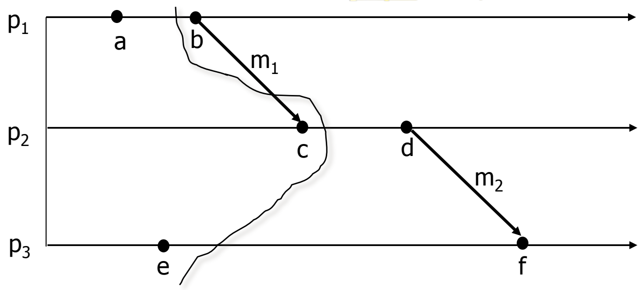
Sometimes it is assumed that sending a message does not alter the local state

Sometimes we are only interested in the events:

$$h_i = e_i^1, e_i^2, e_i^3, \cdots$$

INF5040, Roman Vitenberg 10

Global histories and cuts



- Global history: a collection of local histories, one from each process
- Cut: union of prefixes of process histories
  - May be consistent or not

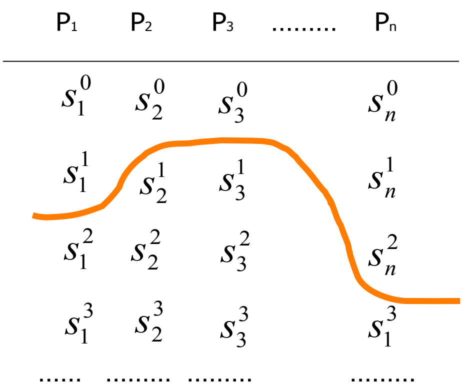
Consistent cuts

Cut C is consistent if

$$e \in C \wedge (f \rightarrow e) \Rightarrow f \in C$$

When reasoning about system execution, we are only interested in consistent cuts!

Global state



If local states do not include message sends, we additionally need to capture messages in transition

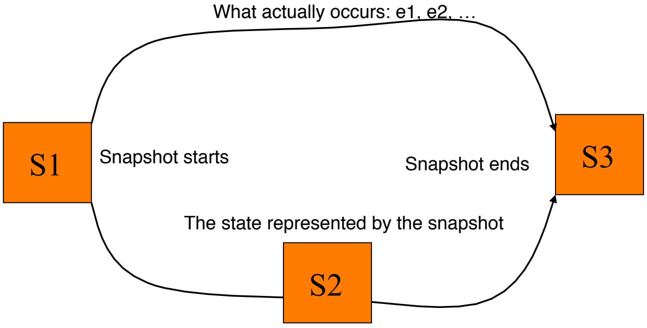
…consistent states correspond to consistent cuts...

Reasoning about global states and its applications

- Linearization is a full ordering of events in a global history that preserves →
- State S' is reachable from state S if there is a linearization that starts in S and ends in S'
- Property: a global state predicate
  - Stable property: if true in S, true in every state reachable from S
  - Safety property: true in every state reachable from S_0
  - Liveness property: in every linearization, there is a state reachable from S_0 in which it is true

The snapshot problem

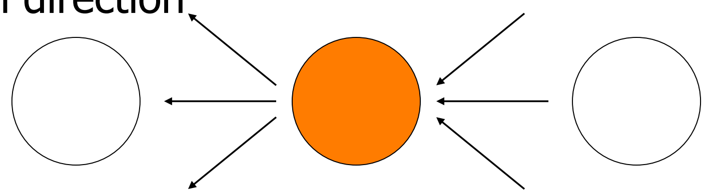
- Finds a consistent global state that may have occurred



- Consistent state reachable from S1 so that S3 is reachable from it

Assumptions for the snapshot algorithm

- No process or network link fails
- Network links preserve FIFO
- Full network: each pair of processes connected by two network links, one in each direction



Responsibility of the processes

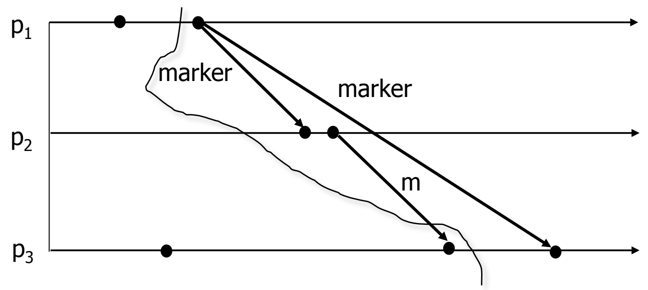
- Every process can initiate a snapshot
  - A process takes initiative to log its own state and sends a marker message on all output channels.
- Each process has responsibility for
  - Logging its own state,
  - Logging the incoming messages on input channels,
  - Sending or forwarding the marker.
- Upon termination, the collection of local states of processes and recorded states of channels should give us a consistent global state

The first attempt

```
P sends a marker over all outgoing links…
P waits until it receives a marker on all input channels
P logs its own state

When another process Q receives a marker
        Q logs its own state
        Q sends the marker back to P…
```

Is the protocol correct?



- The captured state may be inconsistent
- It does not capture messages in transit

INF5040, Roman Vitenberg 20

Correct snapshot protocol

- [Chandy,Lamport 1985]
- The procedure to start the snapshot

```
P logs its own state.
P sends a marker over all outgoing links.
P starts to log incoming messages on all
        input channels
```

INF5040, Roman Vitenberg 21

The procedure upon marker reception

```
When P receives a marker over channel c
  IF P has not recorded its state
     P records its state.
     P forwards the marker over all output channels
     P sets the state of c to the empty set
     P starts to record incoming messages
        on all other input channels
  ELSE
     P records the state of c:
             all the messages that have been
             received on c since P recorded
             its own state, which are said to be
             in transition over the channel
  END
```

INF5040, Roman Vitenberg 22

Proof of protocol correctness

- The recorded state is consistent.
  - If $x \rightarrow y$, and y occurred at p before p recorded its state, then x must have occurred at q before q recorded its state.
- State S2 must be reachable from S1.
- State S3 must be reachable from S2.

INF5040, Roman Vitenberg 23

Distributed consensus

- N processes out of which at most f can be faulty
- Two possible input values, 0 or 1
- Agreement (also called correctness)
  - No two non-faulty processes decide on different values
- Termination
  - If there are non-faulty processes, at least one of them decides
- Integrity (or validity or non-triviality)
  - if all non-faulty processes start with the same initial value v, then v is the only possible decision value for a non-faulty process

Other agreement problems

- Reliable multicast (also called terminating reliable broadcast)
- Group membership
- Leader election
- Distributed locking
- Mutual exclusion
- Atomic transactions
- Coordinated resource allocation

Reliable broadcast

- One sender that sends a single message
- Termination: Every non-faulty process delivers a message (possibly $\perp$)
- Agreement: No two non-faulty processes deliver different messages
- Validity: no spurious messages
- Integrity: If the sender is non-faulty, it delivers the message it sent

Group membership

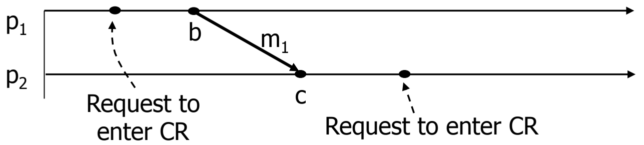
- Each process starts with a list of processes it considers correct
- Agreement on the list of participating processes
- Validity 1: If a process is in all input lists, then it will be in the decided list
- Validity 2: If a process is in no input list, then it will not be in the decided list

Known impossibility results for distributed consensus

- Impossible to solve if at least a third of all processes are malicious
  - Can be alleviated by using digital signatures
- Impossible to solve in asynchronous systems where processes can fail
  - Can be circumvented by masking faults
  - Or by designating the process that adds to asynchrony as faulty
  - Or by using randomization

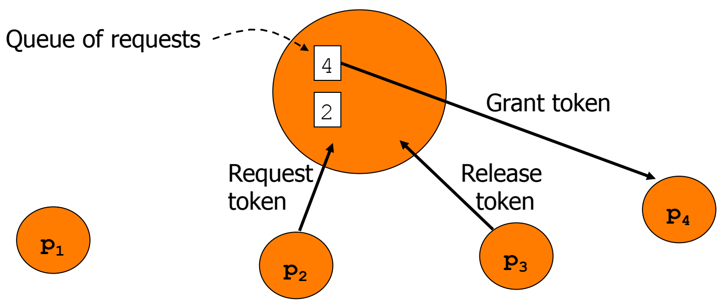
Mutual exclusion problem

- Safety:
  - At most one process can be in a critical section at a time
- Liveness:
  - Each request to enter or exit the critical section eventually succeeds (as long as the process that executes in the critical section eventually requests to leave it)
- Ordering:
  - Entrance to the CS must observe the “happened-before” relation



Central server algorithm

- Central server that grants entrance to the critical section
- protocol
  - enter() -- enter critical section - blocks if necessary
    exit() -- leaves critical section - other processes can now enter



INF5040, Roman Vitenberg 30

Evaluation of the central server algorithm

- Are safety and liveness satisfied?
- Is ordering satisfied?
  - How to ensure it?
- Shortcomings of the algorithm
  - Performance bottleneck
  - The server can fail
    - We can make one of the clients a new server
    - Requires distributed election
    - How to ensure that the old order preceding the failure is preserved?
- Client with the token may fail
  - How to ensure that the token becomes accessible again?

INF5040, Roman Vitenberg 31

Ring-based algorithm

- A token rotating in one direction
- A process can enter the critical section when it has the token
- When a process that has not requested to enter receives a token, it passes the token on

INF5040, Roman Vitenberg 32

Evaluation of the ring-based algorithm

- No central bottleneck
  - Redundant messages are sent if no process attempts to enter the critical section
  - A process may have to wait a long time for a token
- Safety and liveness are trivially satisfied in absence of failures, but ordering requires an additional mechanism
- Fault-tolerance
  - Problematic when a node crashes
    - Mend the ring
    - Ensure that the ring contains exactly one token

INF5040, Roman Vitenberg 33

Distributed algorithm based on logical clocks

- Basic idea [Ricart & Agrawala, 1981]:
  - A process that wishes to enter a critical section, multicasts a message to all the processes
  - A process can enter a CS when it gets acks from all the processes
  - Rules wrt when to send an ack in order to ensure fulfillment of the requirements
- Assumptions
  - Processes know each other addresses
  - Every sent message will eventually be delivered
- Properties
  - Each process maintains a logical clock
  - Timestamps include processId: $<T,p>$ (i.e., total ordering)
  - Each process maintains its state wrt token possession
    - RELEASED, WANTED, HELD

Ricart & Agrawala algorithm

Upon initialization
    state := RELEASED;
To enter the critical section
    state:= WANTED;
    Multicast a timestamped request to all the processes
    T := the current timestamp;
    wait until ((n-1) acks are received);
    state := HELD;
Upon receiving a request with $<T_i,p_i>$ at p_j $(i \neq j)$
    if (state=HELD *or* (state=WANTED *and* $(T,p_j) < (T_i,p_i)$))
        queue the request from p_i without replying
    else
        send an ack to p_i
    end if
Upon exiting from the critical section
    state := RELEASED
    reply to all queued messages

Evaluation of the Ricart & Agrawala algorithm

- Are safety and liveness satisfied?
- Is ordering satisfied?
- Shortcomings
  - Many messages are sent in order to enter critical section
    - 2(n-1) messages without network support for multicast
    - n messages with native network support for multicast
  - Not resilient to process crashes

Summary of distributed mutual exclusion algorithms

- Little resilience to failures
  - Can be improved by additional mechanisms
  - But it will never be perfect in an asynchronous system
- Central server requires the lowest number of messages but can become a bottleneck

Requirements for distributed leader election

- In many distributed algorithms, one of the participating processes will play the role of a central coordinator
  - Central server in the mutual exclusion algorithms
  - Coordinator of a distributed transaction
- If a coordinator fails, one of the remaining processes can be elected to take over the central role
  - In order to provide better fault-tolerance
- The main requirements
  - Safety: only one leader may exist at a time
  - Liveness: a leader will eventually be elected

The "Bully" algorithm

- [Silberschatz et al, 1993]
- Prerequisites
  - The processes know each other identities and addresses
  - Process identifiers are totally ordered
  - The algorithm selects the process with the biggest identifier
- Message types
  - `election`: announces an election
  - `answer`: is sent as a reply to the election message
  - `coordinator`: announces the identity of the new coordinator

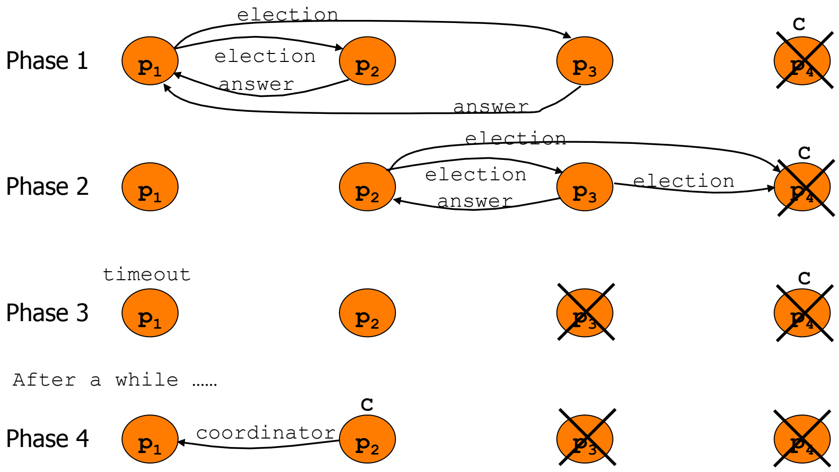
The "Bully" algorithm II

- Election procedure
  - The process (that detects that the coordinator has failed) sends the `election` message to the processes that have a bigger identifier
  - It then waits a limited amount of time for the `answer` message
  - If no `answer` message is received, the process considers itself as a new coordinator and sends a `coordinator` message to all the processes with smaller identifiers
  - If an `answer` message is received, the process waits a limited amount of time for a `coordinator` message. If none arrives, it starts a new election.

The "Bully" algorithm III

- Election procedure (continued)
  - If a process receives a `coordinator` message, it memorizes the identifier included in the message and considers the process as the new coordinator
  - If a process receives an `election` message, it sends back an `answer` message and starts a new election - unless the process has already started one
  - When a process recovers or joins the system, it starts a new election. If it has the biggest identifier, it makes itself a coordinator and announces it, even if there is another functioning coordinator

Illustration for the "bully" algorithm



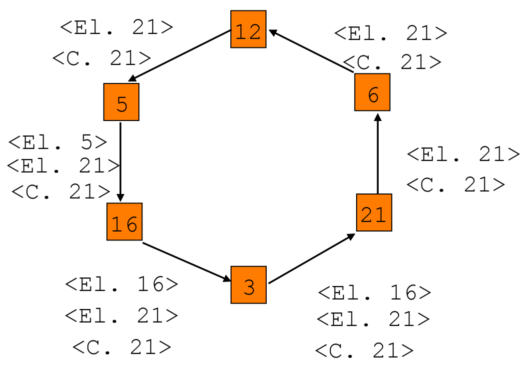
INF5040, Roman Vitenberg 42

Evaluation of the "bully" algorithm

- Best case: n-2 `coordinator` messages
  - Occurs when the process with the second highest id detects that the coordinator has failed
- Worst case: $O(n^2)$ messages
  - Occurs when the process with the lowest id detects that the coordinator has failed
  - => (n-1) processes start an election
- Ring-based algorithm is more efficient wrt the number of messages

INF5040, Roman Vitenberg 43

The ring-based algorithm



When a message has made a full circle without changing the id, the process will know that it has the highest number

Then it must inform all other processes that it is the leader

…it is possible to handle multiple elections that have been started concurrently