Distributed Systems

Coordination and Agreement

Overview

- Co-ordination in distributed systems
  - why needed, sources of problems
- Mutual exclusion
  - ring-based
  - multicast & logical clocks [Ricart & Agrawala]
- Leader election
  - ring-based [Chang & Roberts]

Co-ordination algorithms

- are fundamental in distributed systems:
- for resource sharing: concurrent updates of
  - records in a database (record locking)
  - files (file locks in stateless file servers)
  - a shared bulletin board
- to agree on actions: whether to
  - commit/abort database transaction
  - agree on a readings from a group of sensors
- to dynamically re-assign the role of master
  - choose primary time server after crash
  - choose co-ordinator after network reconfiguration

Why difficult?

- Centralised solutions not appropriate
  - communications bottleneck
- Fixed master-slave arrangements not appropriate
  - process crashes
- Varying network topologies
  - ring, tree, arbitrary; connectivity problems
- Failures must be tolerated if possible
  - link failures
  - process crashes
- Impossibility results
  - in presence of failures, esp. asynchronous model

Co-ordination problems

- Mutual exclusion
  - distributed form of critical section problems
  - must use message passing
- Leader elections
  - after crash failure has occurred
  - after network reconfiguration
- Consensus (also called Agreement)
  - similar to coordinated attack
  - some based on multicast communication
  - variants depending on type of failure, network, etc

Failure assumptions

- Assume reliable links, but possible process crashes.
- Failure detection service
  - processes query if a process has failed
  - how?
    - processes send ‘P is here’ messages every T secs
    - failure detector records replies
  - unreliable, especially in asynchronous systems
- Observations of failures:
  - Suspected: no recent communication, but could be slow
  - Unsuspected: but no guarantee it has not failed since
  - Failed: crash has been determined

Distributed mutual exclusion

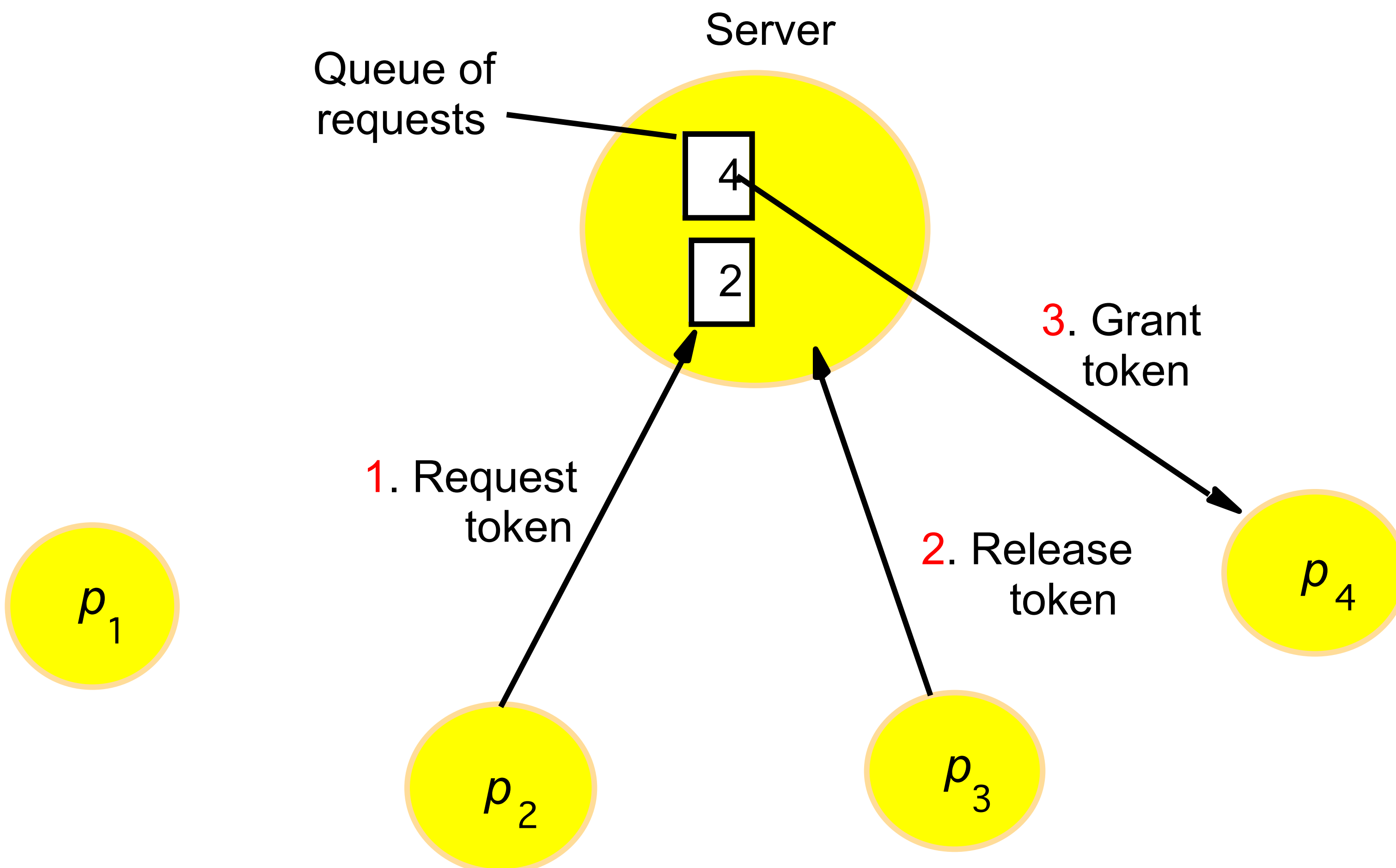
- The problem:
  - N asynchronous processes, for simplicity no failures
  - guaranteed message delivery (reliable links)
  - to execute critical section (CS), each process calls:
    - **enter**()
    - **resourceAccess**()
    - **exit**()
- Requirements

(**ME1**) At most one process is in CS at the same time.

(**ME2**) Requests to **enter** and **exit** are eventually granted.

(**ME3** - Optional, stronger) Requests to **enter** granted according to causality order.

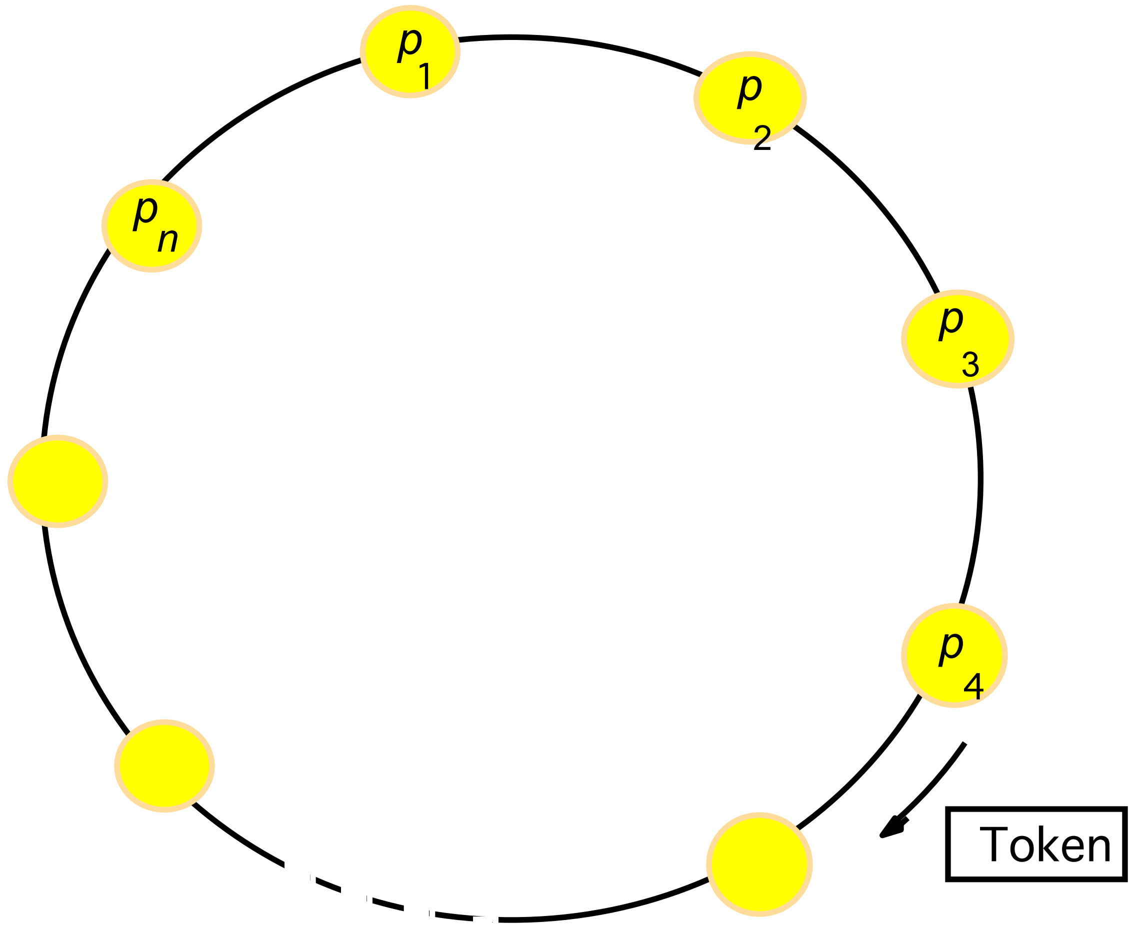
Centralised mutual exclusion



Centralised service

- Single server implements imaginary token:
  - only process holding the token can be in CS
  - server receives **request** for token
  - replies **grant**ing access if CS free; otherwise, request queued
  - when a process **release**s token, oldest request from queue granted
- It works though...
  - does not respect causality order of requests (MC3) - why?
- but
  - server is performance bottleneck!
  - what if server crashes?

Ring-based algorithm



- Arrange processes in a logical ring, let them pass token.

Ring-based algorithm

- No server bottleneck, no master
- Processes:
  - continually pass token around the ring, in one direction
  - if do not require access to CS, pass on to neighbour
  - otherwise, wait for token and retain it while in CS
  - to exit, pass to neighbour
- How it works
  - continuous use of network bandwidth
  - delay to enter depends on the size of ring
  - causality order of requests not respected (ME3) - why?

Ricart-Agrawala algorithm

- Based on multicast communication
  - N inter-connected asynchronous processes, each with
    - unique id
    - Lamport's logical clock
  - processes multicast request to enter critical section:
    - timestamped with Lamport's clock and process id
  - entry granted
    - when all other processes replied
    - simultaneous requests resolved with the timestamp
- How it works
  - satisfies the stronger property (ME3)
  - if hardware support for multicast, only one message to enter

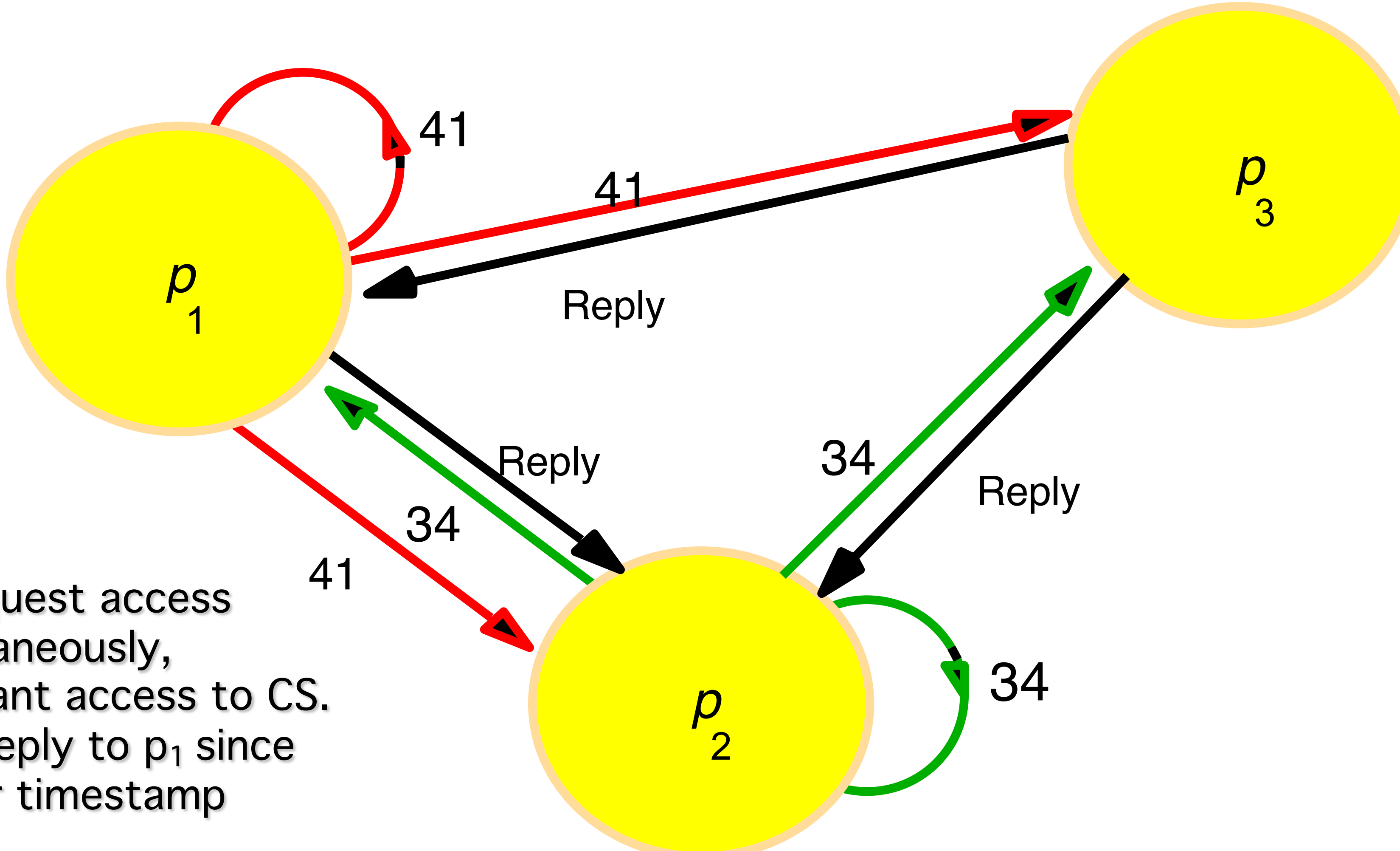
Ricart-Agrawala algorithm

On initialization
    state := RELEASED;
To ***enter the critical section***
    state := WANTED;
    Multicast *request* to all processes;
    T := request's timestamp;
    Wait until (number of replies received = $(N - 1)$);
    state := HELD;

request processing deferred here

On receipt of a request $<T_i, p_i>$ *at* p_j $(i \neq j)$
    if (*state* = HELD) or ((*state* = WANTED) *and* $((T, p_j) < (T_i, p_i))$)
    then
        queue *request* from p_i without replying;
    else
        reply immediately to p_i;
    end if
To ***exit the critical section***
    state := RELEASED;
    reply to any queued requests;

Multicast mutual exclusion



p_1, p_2 request access simultaneously,
p_3 does not want access to CS.
p_2 does not reply to p_1 since has lower timestamp

Mutual exclusion summary

- Performance
  - one request-reply enough to enter
  - relatively high usage of network bandwidth
  - client delay depends on the frequency of access and size of network
- Fault tolerance
  - usually assume reliable links
  - some can be adapted to deal with crashes
- Other solutions
  - sufficient to obtain agreement from certain overlapping subsets of processes voting set (Maekawa algorithm)

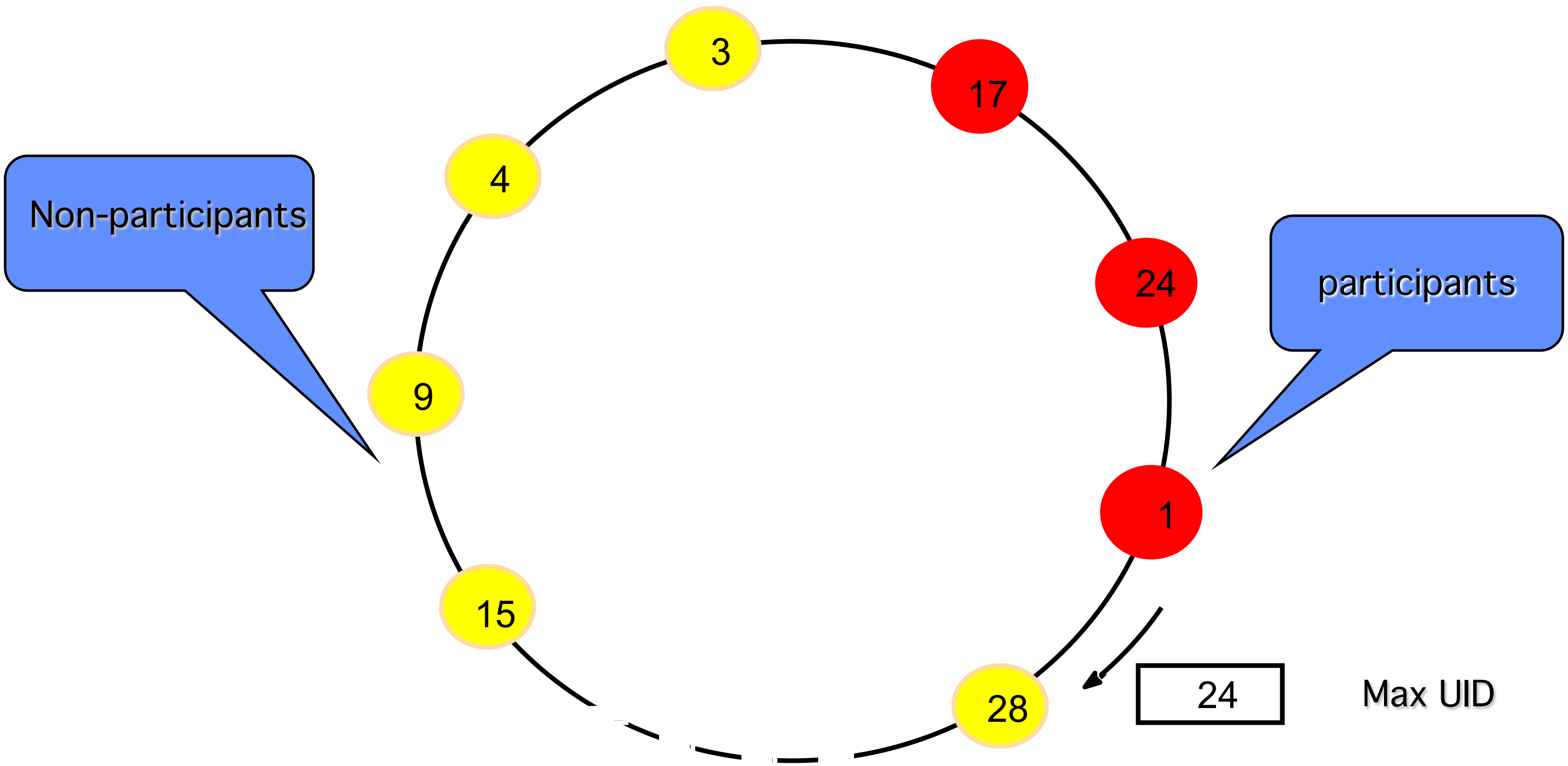
Leader election algorithms

- The problem
  - N processes
  - for simplicity assume no crashes
  - must choose unique master co-ordinator amongst processes
  - election called after failure has occurred
  - one or more processes can call election simultaneously
- Requirements

  (**LE1**) Every process knows P, identity of leader, where P is unique process id (usually maximum) or is yet undefined.

  (**LE2**) All processes participate and eventually discover the identity of the leader (cannot be undefined).

Chang&Roberts algorithm



- Leader election in a ring: asynchronous model, UIDs known.

Chang&Roberts algorithm

- Assumptions
  - unidirectional ring, asynchronous, each process has UID
- Election
  - initially each process non-participant
  - determine leader (*election* message):
    - initiator becomes participant and passes own UID on to neighbour
    - when non-participant receives *election* message, forwards maximum of own and the received UID and becomes participant
    - participant does not forward the *election* message
  - announce winner (*elected* message):
    - when participant receives *election* message with own UID, becomes leader and non-participant, and forwards UID in *elected* message
    - otherwise, records the leader's UID, becomes non-participant and forwards it

Chang&Roberts algorithm

- How it works
  - if UIDs, then identity of leader unique
  - two exchanges around the ring: election, announce winner
  - if one process starts election
    - in worst case 3 round-trips needed - explain?
- but
  - does not tolerate failures (need reliable failure detector)
    - see bully algorithm (synchronous model)
    - works if more than one process simultaneously start election
  - what if no UIDs?
    - nodes on a re-configurable network, ‘hot-pluggable’

Agreement

- Agreement (= Consensus) problems
  - why & where needed
  - definition
- Byzantine generals
  - in synchronous systems
  - in asynchronous systems
- And finally
  - impossibility results!
  - practical implications

Consensus algorithms

- used when it is necessary to agree on actions:
  - in transaction processing
    - commit or abort transaction?
  - mutual exclusion
    - which process should enter the critical section?
  - in control systems
    - proceed or abort based on sensor readings?

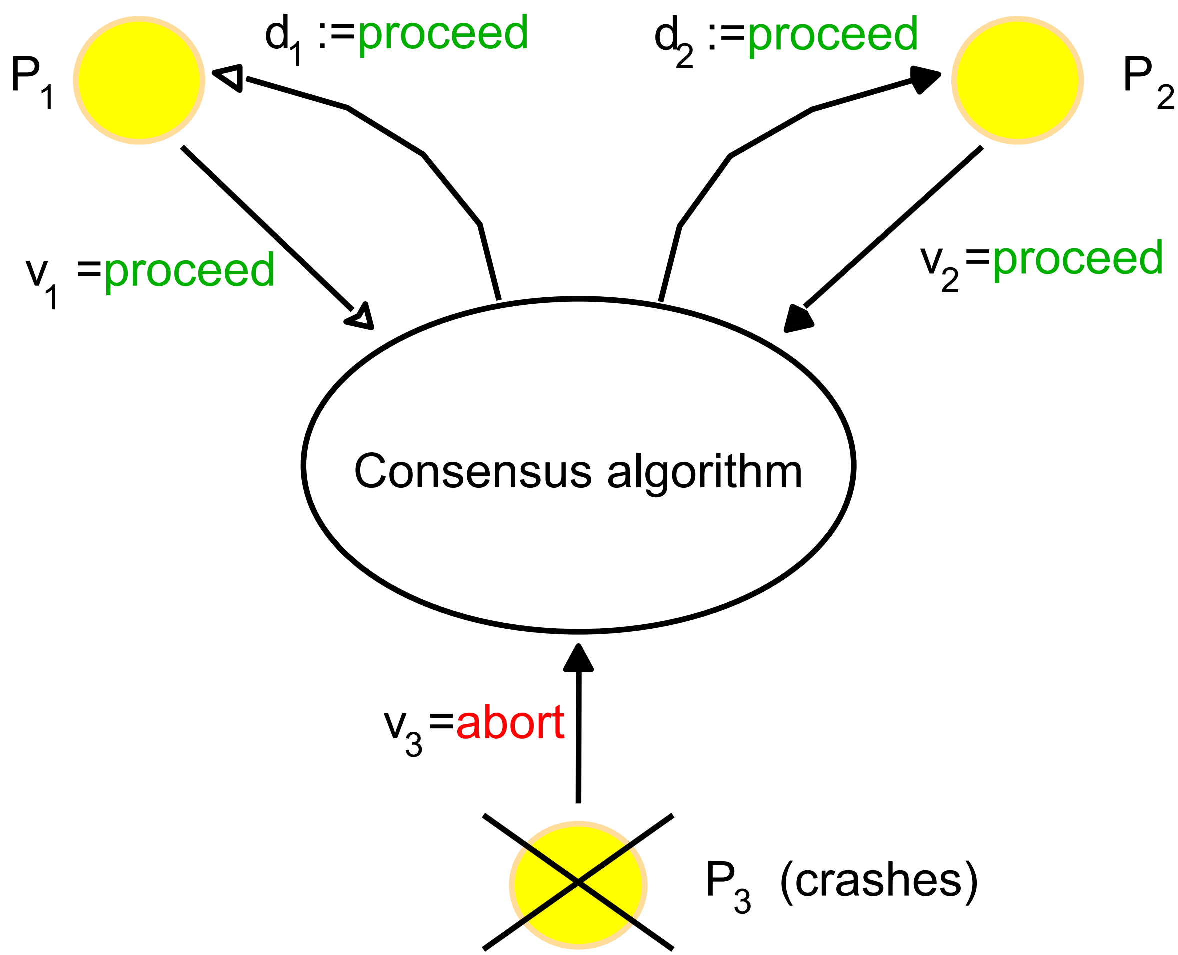
The model & assumptions

- The model
  - N processes
  - message passing
  - synchronous or asynchronous
  - communication reliable
- Failures!
  - process crashes
  - arbitrary (Byzantine) failures
    - processes can be treacherous and lie
- The algorithm
  - works in presence of certain failures

Consensus: main idea

- Initially
  - processes begin in **undecided** state
  - propose an initial value from a set D
- Then
  - processes communicate, exchanging values
  - attempt to decide
  - cannot change the decision value in **decided** state
- The difficulty
  - must reach decision even if crash has occurred
  - or arbitrary failure!

Consensus for three processes



Consensus: requirements

- Termination
  - Eventually each correct process sets its decision value.
- Agreement
  - Any two correct processes must have decided on the same decision value.
- Integrity
  - If all correct processes propose the same value, then any correct process that has decided must have chosen that value.

Towards a solution

- For simplicity, assume no failures
  - processes multicast its proposed value to others
  - wait until all N values collected (including own)
  - decide through majority vote (⊥ special value if none)
    - can also use minimum/maximum
- It works since...
  - all processes end up with the same set of values
  - majority vote ensures Agreement and Integrity
- But what about failures?
  - process crash - stops sending values after a while
  - arbitrary failure - different values to different processes

Byzantine generals

- The problem [Lamport 1982]
  - three or more generals are to agree to **attack** or **retreat**
  - one (commander) issues the order
  - the others (lieutenants) decide
  - one or more generals are treacherous (= faulty!)
    - propose attacking to one general, and retreating to another
    - either commander or lieutenants can be treacherous!
- Requirements
  - Termination, Agreement as before.
  - Integrity: If the commander is correct then all correct processes decide on the value proposed by commander.

Consensus in synchronous system

- Uses basic multicast
  - guaranteed delivery by correct processes assuming the sender does not crash
- Admits process crash failures
  - assume up to f of the N processes may crash
- How it works...
  - f +1 rounds
  - relies on synchrony (timeout!)

Consensus in synchronous system

- Initially
  - each process proposes a value from a set D
- Each process
  - maintains the set of values V_r known to it at round r
- In each round r, where $1 \leq r \leq f+1$, each process
  - multicasts the values to each other (only values not sent before, $V_r - V_{r-1}$)
  - receives multicast messages, recording any new value in V_r
- In round f+1
  - each process chooses minimum V_{f+1} as decision value

Consensus in synchronous system

- Why it works?
  - sets timeout to maximum time for correct process to multicast message
  - can conclude process crashed if no reply
  - if process crashes, some value not forwarded...
- At round f+1
  - all correct process arrive at the same set of values
  - hence reach the same decision value (minimum)
  - at least f+1 rounds needed to tolerate f crash failures
- What about arbitrary failures?

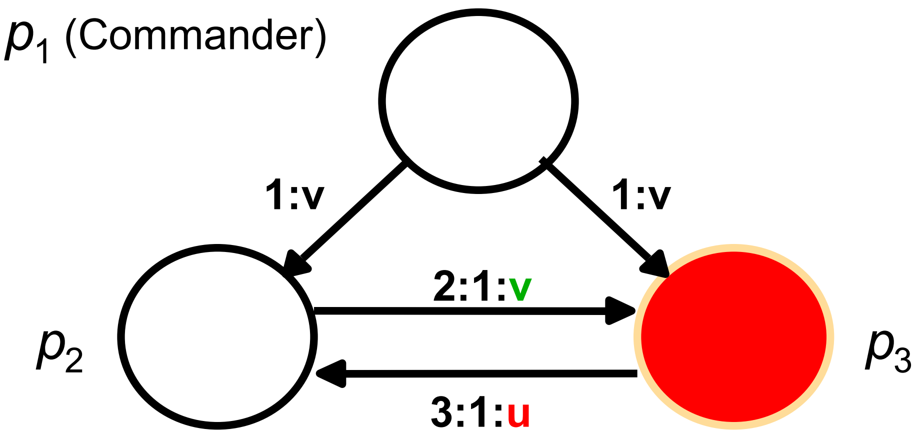
Byzantine generals...

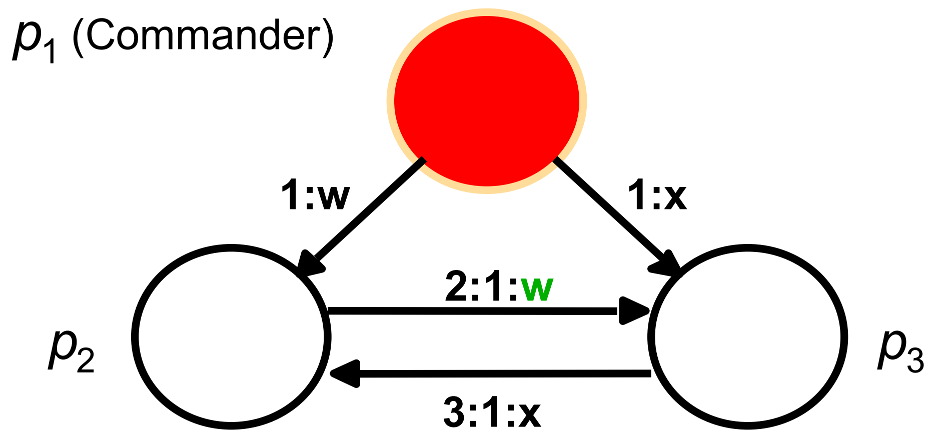
- Processes exhibit arbitrary failures
  - up to f of the N processes faulty
- In a synchronous system
    - can use timeout to detect absence of a message
  - cannot conclude process crashed if no reply
  - impossibility with $N \leq 3f$
- In asynchronous system
  - cannot use timeout to reliably detect absence of a message
  - impossibility with even one failure!!

Impossibility with three generals

- Assume synchronous system
  - 3 processes, one faulty
  - if no message received, assume ⊥
  - proceed in rounds
  - messages '3:1:u' meaning '3 says 1 says u'
- Problem! '1 says v' and '3 says 1 says u'
  - cannot tell which process is telling the truth!
  - goes away if digital signatures used...
- Show
  - no solution to agreement for N=3 and f=1
- Can generalise to impossibility for $N \leq 3f$

Three Byzantine generals





Faulty processes are shown in Colour

p_3 sends illegal value to p_2
p_2 cannot tell which value sent by commander

commander faulty
p_2 cannot tell which value sent by commander

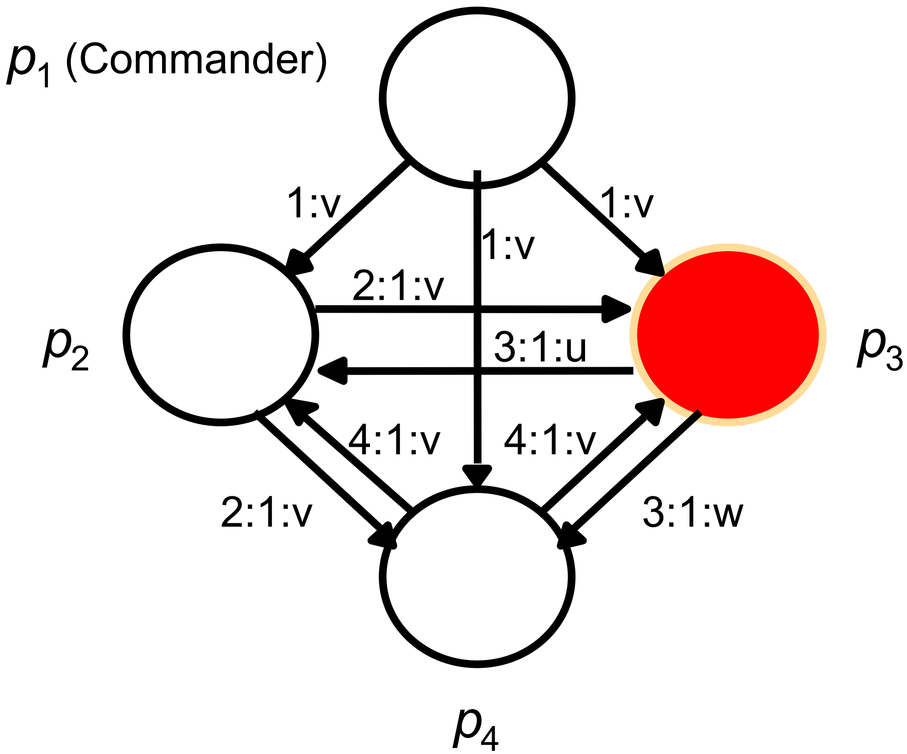
Impossibility with three generals

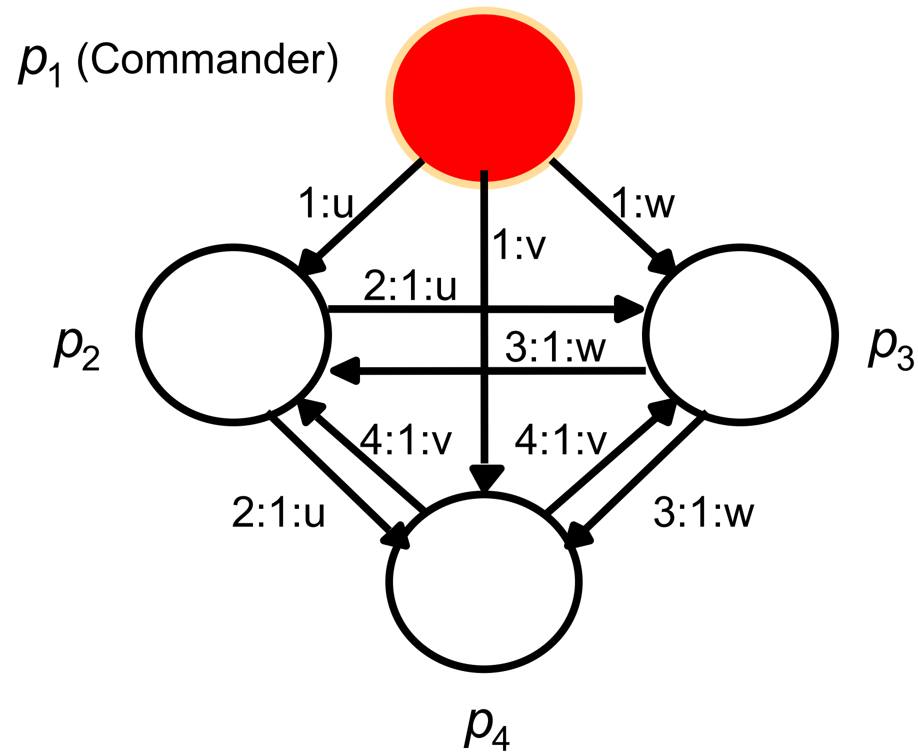
- So, if a solution exists
  - p_2 decides on value sent by commander (v) when the commander is correct
  - and also when commander faulty (w), since cannot distinguish between the two scenarios
- Apply the same reasoning to p_3
  - obtain p_3 must decide on x when commander faulty
- Thus
  - Contradicts agreement principle! since p_2 decides on w, p_3 on x if commander faulty
  - no solution exists

But...

- Solution exists for 4 processes with one faulty
  - commander sends value to each of the lieutenants
  - each lieutenant sends value it received to its peers
  - if commander faulty, then correct lieutenants have gathered all values sent by the commander
  - if one lieutenant faulty, the each correct lieutenant receives 2 copies of the value from the commander
- Thus
  - correct lieutenants can decide on majority of the values received
- Can generalise to $N \geq 3f + 1$

Four Byzantine generals





Faulty processes are shown red

p_2 decides majority(v,u,v) = v
p_4 decides majority(v,v,w) = v

p_2, p_3 and p_4 decide $\perp$
(no majority exists)

In asynchronous systems...

- No guaranteed solution exists even for one failure!!! [Fisher, Lynch, Paterson ‘85]
  - does not mean never reach consensus in presence of failures
  - but that can reach it with positive probability
- But…
  - Internet asynchronous, exhibits arbitrary failures and uses consensus?
- Solutions exist using
  - partially synchronous systems
  - randomisation [Aspnes&Herlihy, Lynch, etc]