

# CS 425 / ECE 428

## Distributed Systems

### Fall 2016

Indranil Gupta (Indy)

Oct 18, 2016

*Lecture 17: Leader Election*

# WHY ELECTION?

- Example 1: Your Bank account details are replicated at a few servers, but one of these servers is responsible for receiving all reads and writes, i.e., it is the **leader** among the replicas
  - What if there are two leaders per customer?
  - What if servers disagree about who the leader is?
  - What if the leader crashes?  
*Each of the above scenarios leads to Inconsistency*

# MORE MOTIVATING EXAMPLES

- Example 2: (A few lectures ago) In the sequencer-based algorithm for total ordering of multicasts, the “sequencer” = leader
- Example 3: Group of NTP servers: who is the root server?
- Other systems that need leader election:  
Apache Zookeeper, Google’s Chubby
- Leader is useful for coordination among distributed servers

# LEADER ELECTION PROBLEM

- In a group of processes, elect a *Leader* to undertake special tasks
  - And *let everyone know* in the group about this Leader
- What happens when a leader fails (crashes)
  - Some process detects this (using a Failure Detector!)
  - Then what?
- Focus of this lecture: *Election algorithm*. Its goal:
  1. Elect one leader only among the non-faulty processes
  2. All non-faulty processes agree on who is the leader

# SYSTEM MODEL

- $N$  processes.
- Each process has a unique id.
- Messages are eventually delivered.
- Failures may occur during the election protocol.

# CALLING FOR AN ELECTION

- Any process can **call** for an **election**.
- A process can call for **at most one election at a time**.
- Multiple processes are allowed to call an election simultaneously.
  - **All of them together must yield only a single leader**
- The result of an election should not depend on which process calls for it.

# ELECTION PROBLEM, FORMALLY

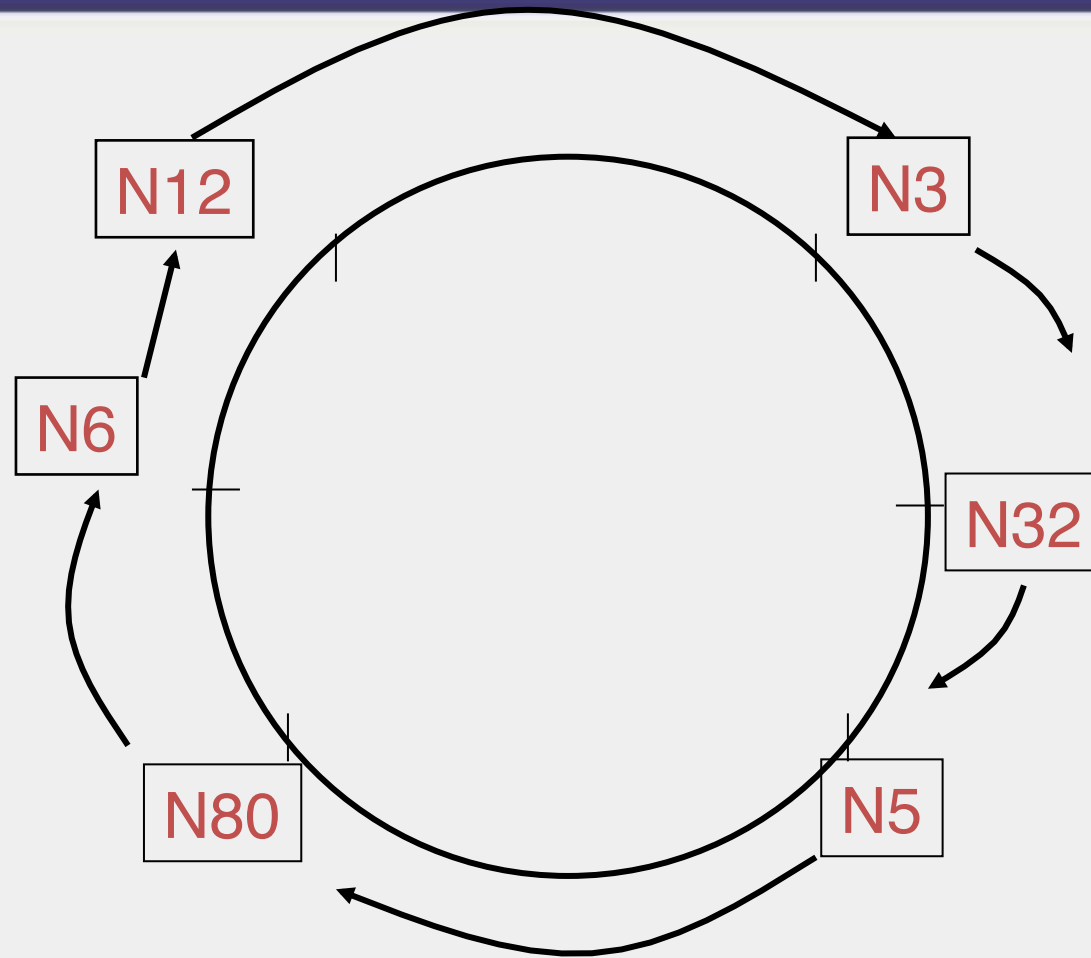
- A run of the election algorithm must always guarantee at the end:
  - **Safety:** For all non-faulty processes  $p$ : ( $p$ 's elected = ( $q$ : a particular non-faulty process with the best attribute value) or Null)
  - **Liveness:** For all election runs: (election run terminates)  
& for all non-faulty processes  $p$ :  $p$ 's elected is not Null
- At the end of the election protocol, the non-faulty process with the best (highest) election attribute value is elected.
  - Common attribute : leader has highest id
  - Other attribute examples: leader has highest IP address, or fastest cpu, or most disk space, or most number of files, etc.

# FIRST CLASSICAL ALGORITHM: RING ELECTION

- $N$  processes are organized in a logical ring
  - Similar to ring in Chord p2p system
  - $i$ -th process  $p_i$  has a communication channel to  $P_{(i+1) \bmod N}$
  - All messages are sent clockwise around the ring.



# THE RING



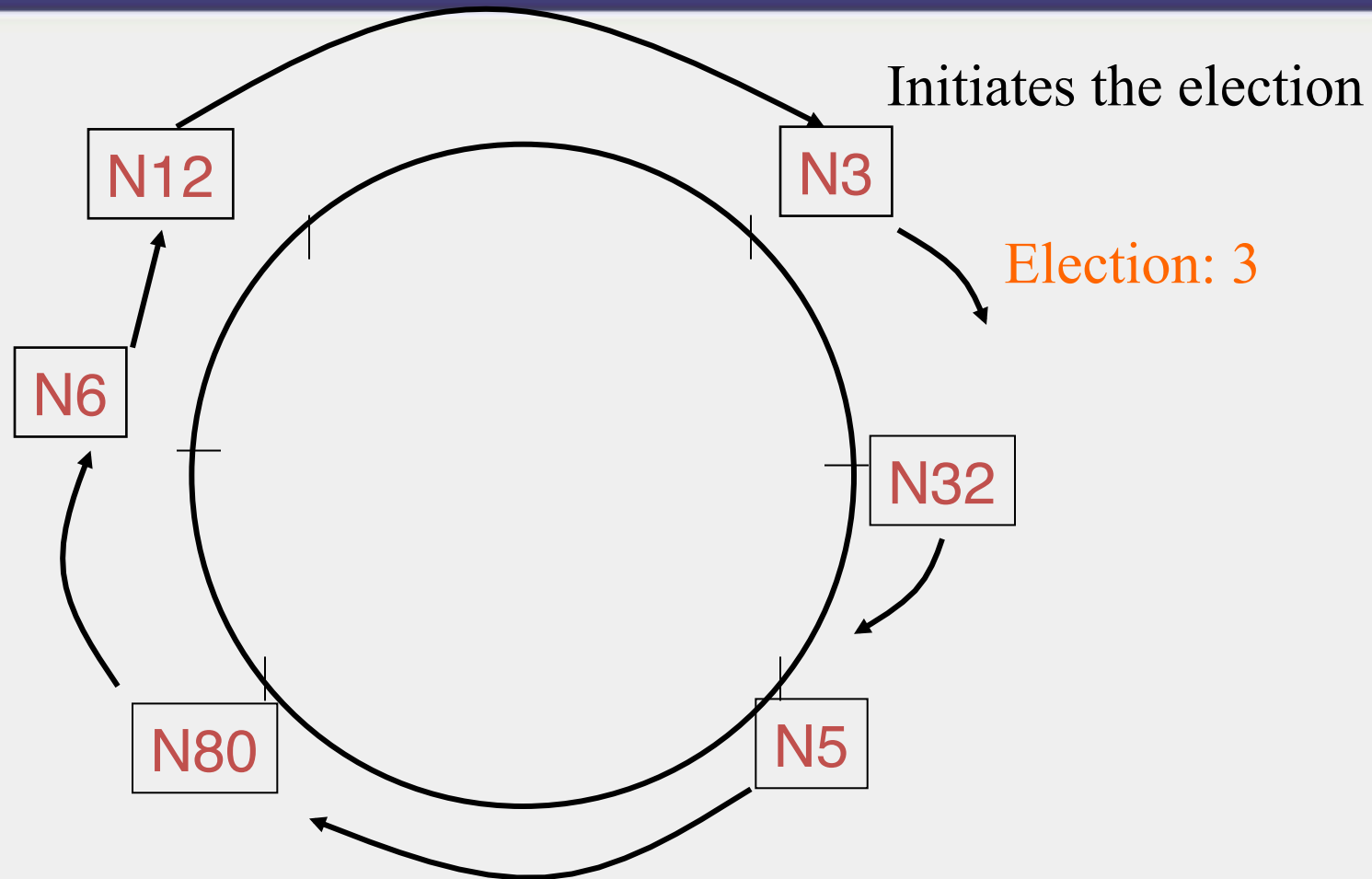
# THE RING ELECTION PROTOCOL

- Any process  $p_i$  that discovers the old coordinator has failed initiates an “Election” message that contains  $p_i$ ’s own id:attr. This is the *initiator* of the election.
- When a process  $p_i$  receives an “Election” message, it compares the attr in the message with its own attr.
  - If the arrived attr is greater,  $p_i$  forwards the message.
  - If the arrived attr is smaller and  $p_i$  has not forwarded an election message earlier, it overwrites the message with its own id:attr, and forwards it.
  - If the arrived id:attr matches that of  $p_i$ , then  $p_i$ ’s attr must be the greatest (why?), and it becomes the new coordinator. This process then sends an “Elected” message to its neighbor with its id, announcing the election result.

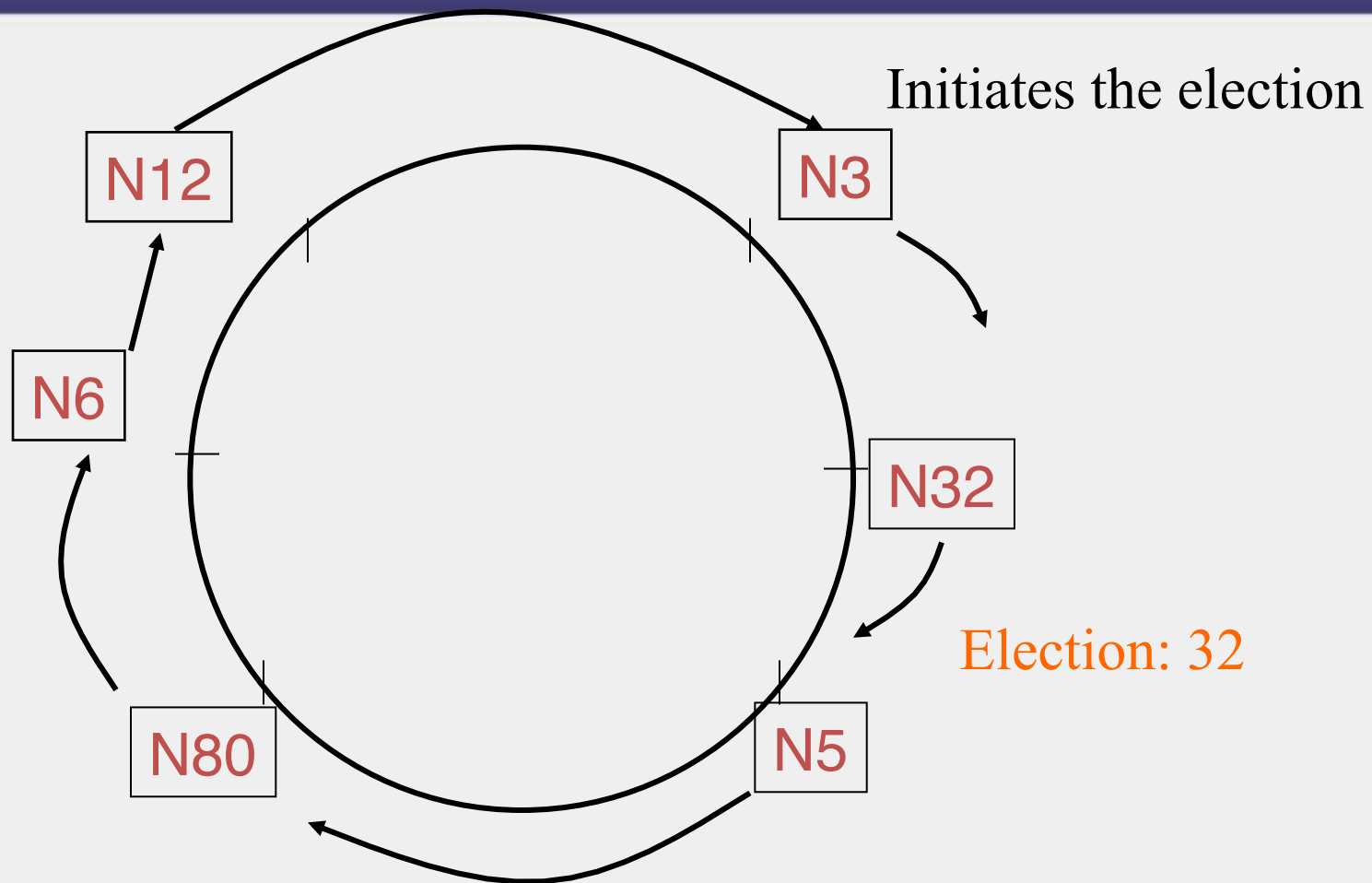
# THE RING ELECTION PROTOCOL (2)

- When a process  $p_i$  receives an “Elected” message, it
  - sets its variable  $elect_i \leftarrow$  id of the message.
  - forwards the message unless it is the new coordinator.

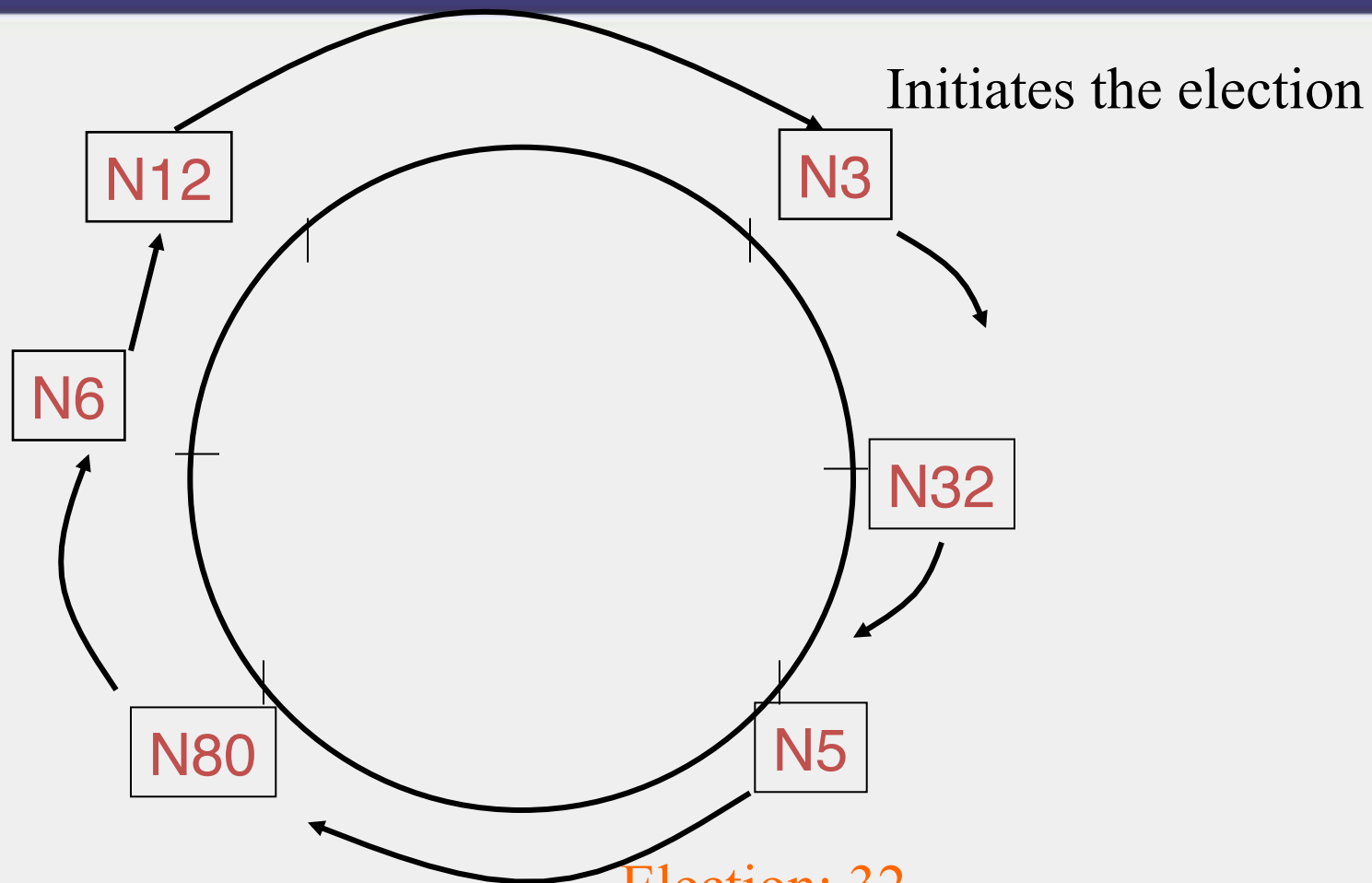
# RING ELECTION: EXAMPLE



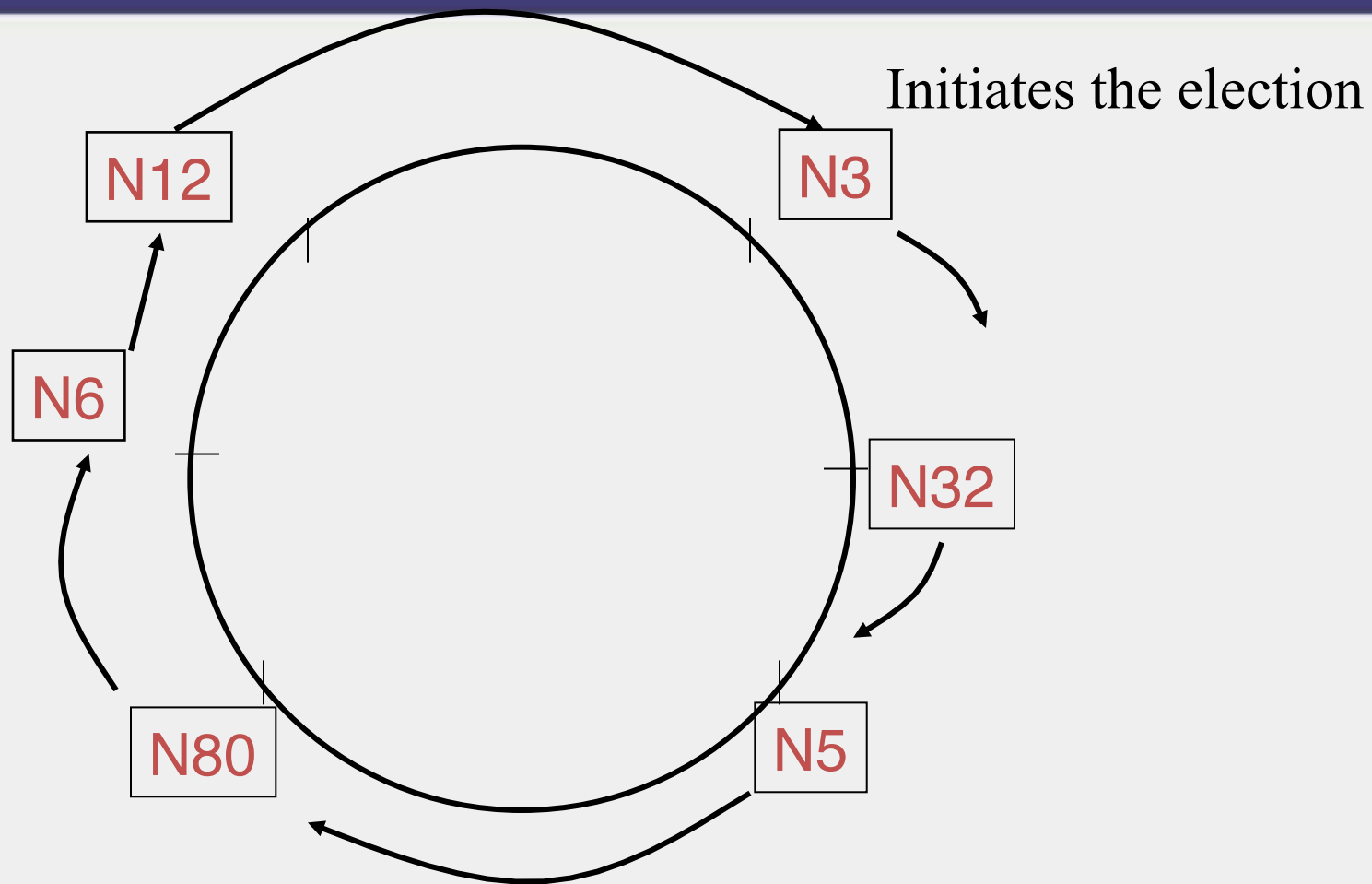
Goal: Elect highest id process as leader



Goal: Elect highest id process as leader



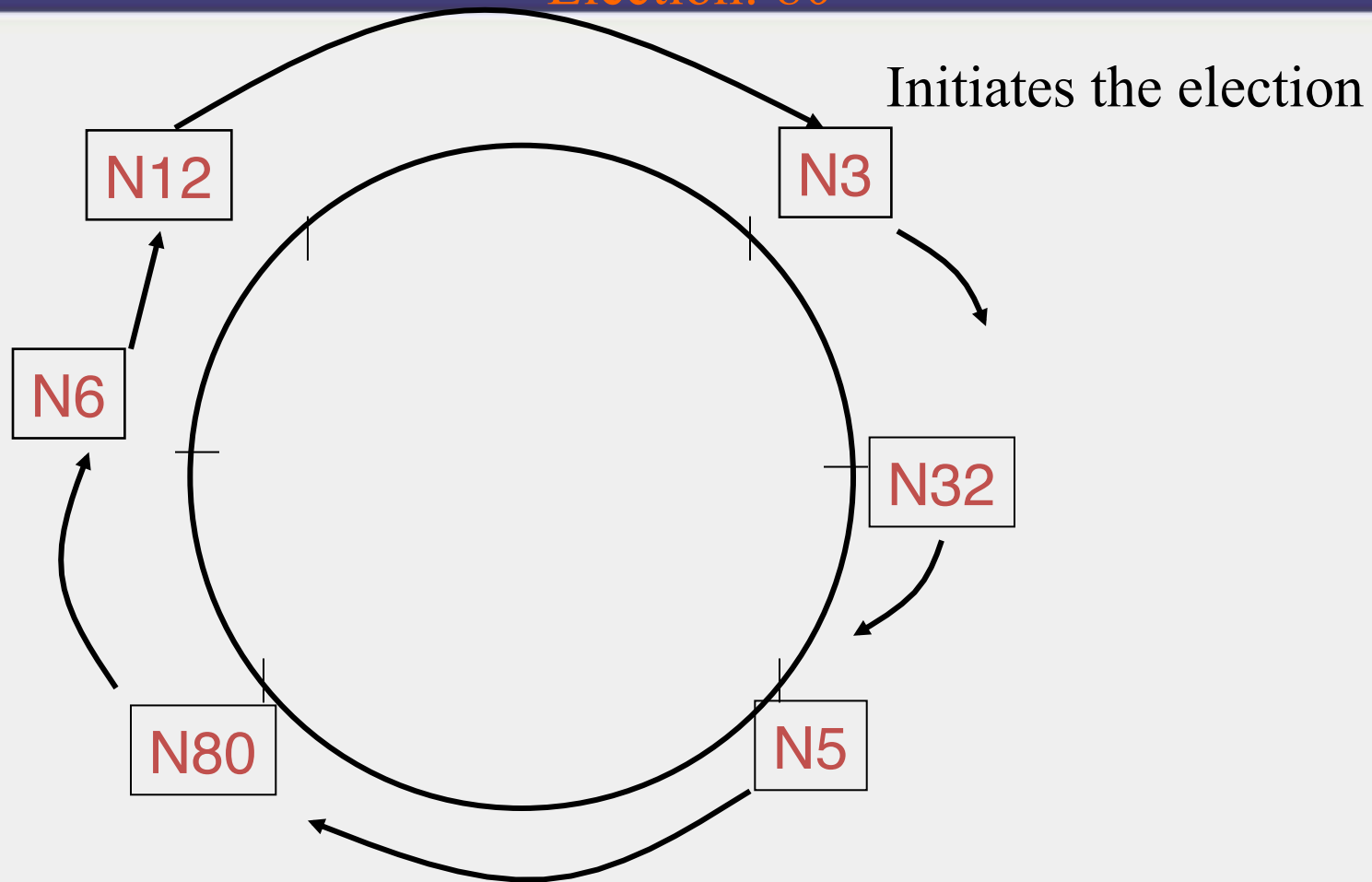
Goal: Elect highest id process as leader Election: 32



Election: 80

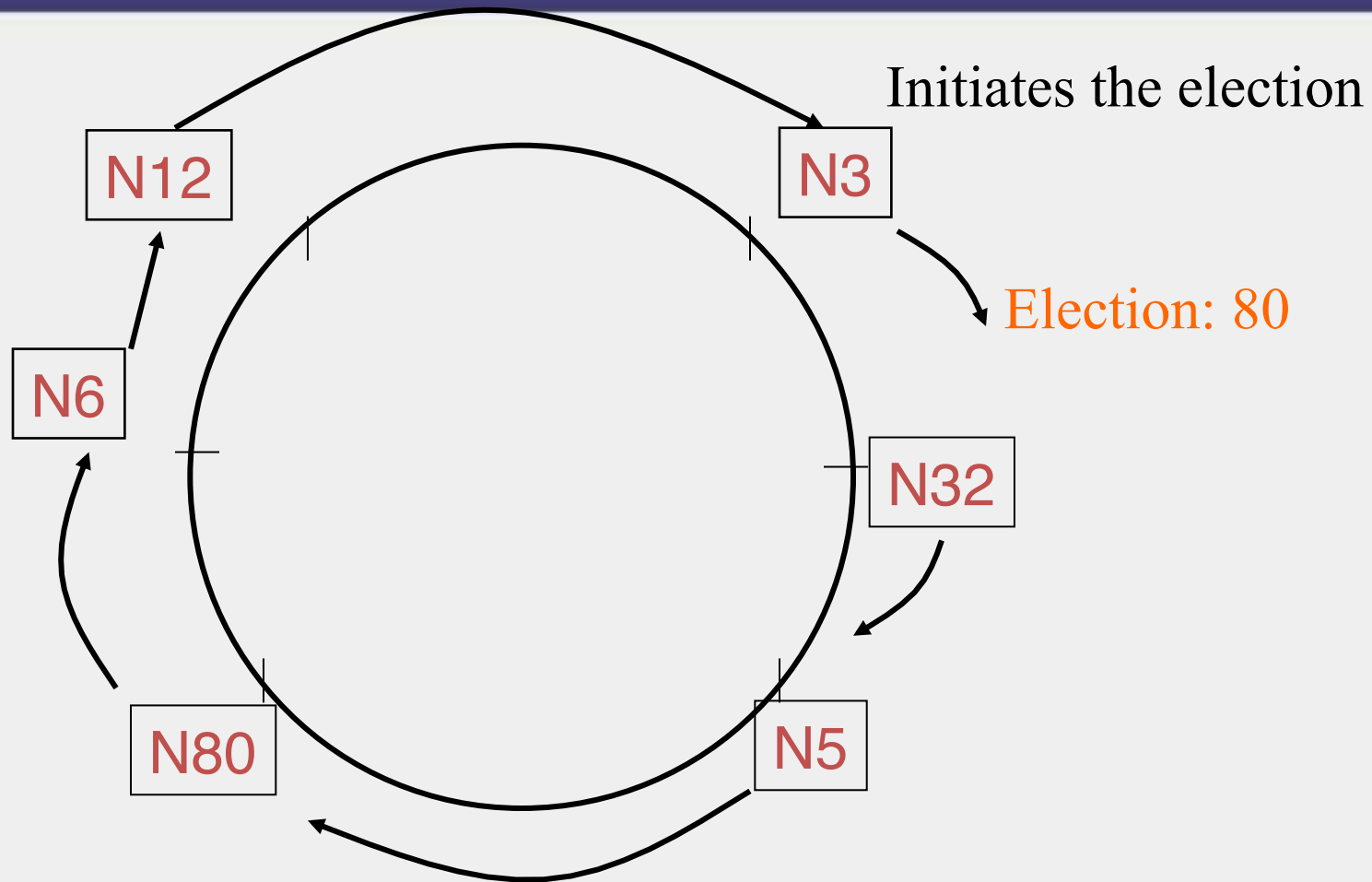
Goal: Elect highest id process as leader

Election: 80

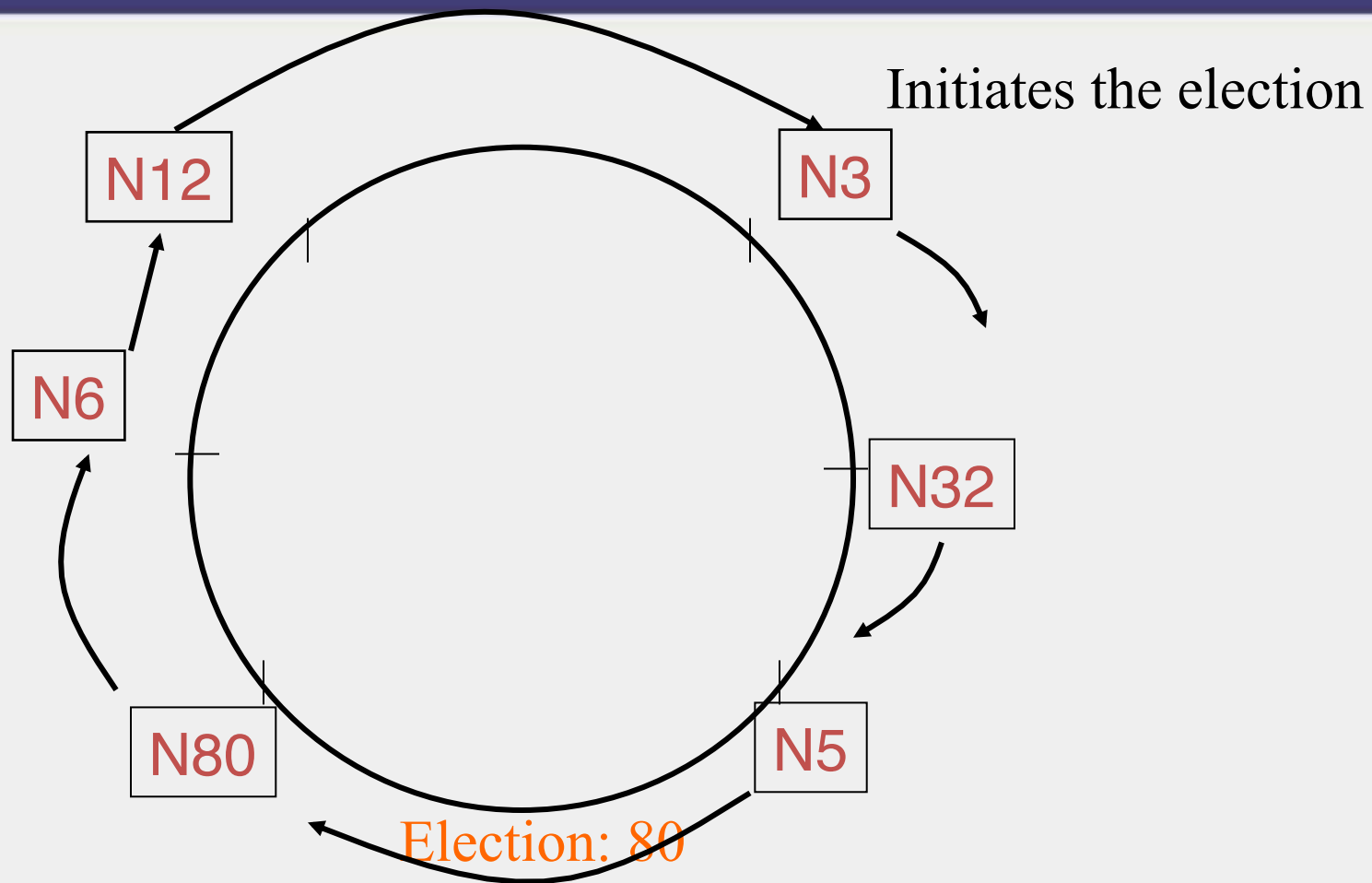


Goal: Elect highest id process as leader

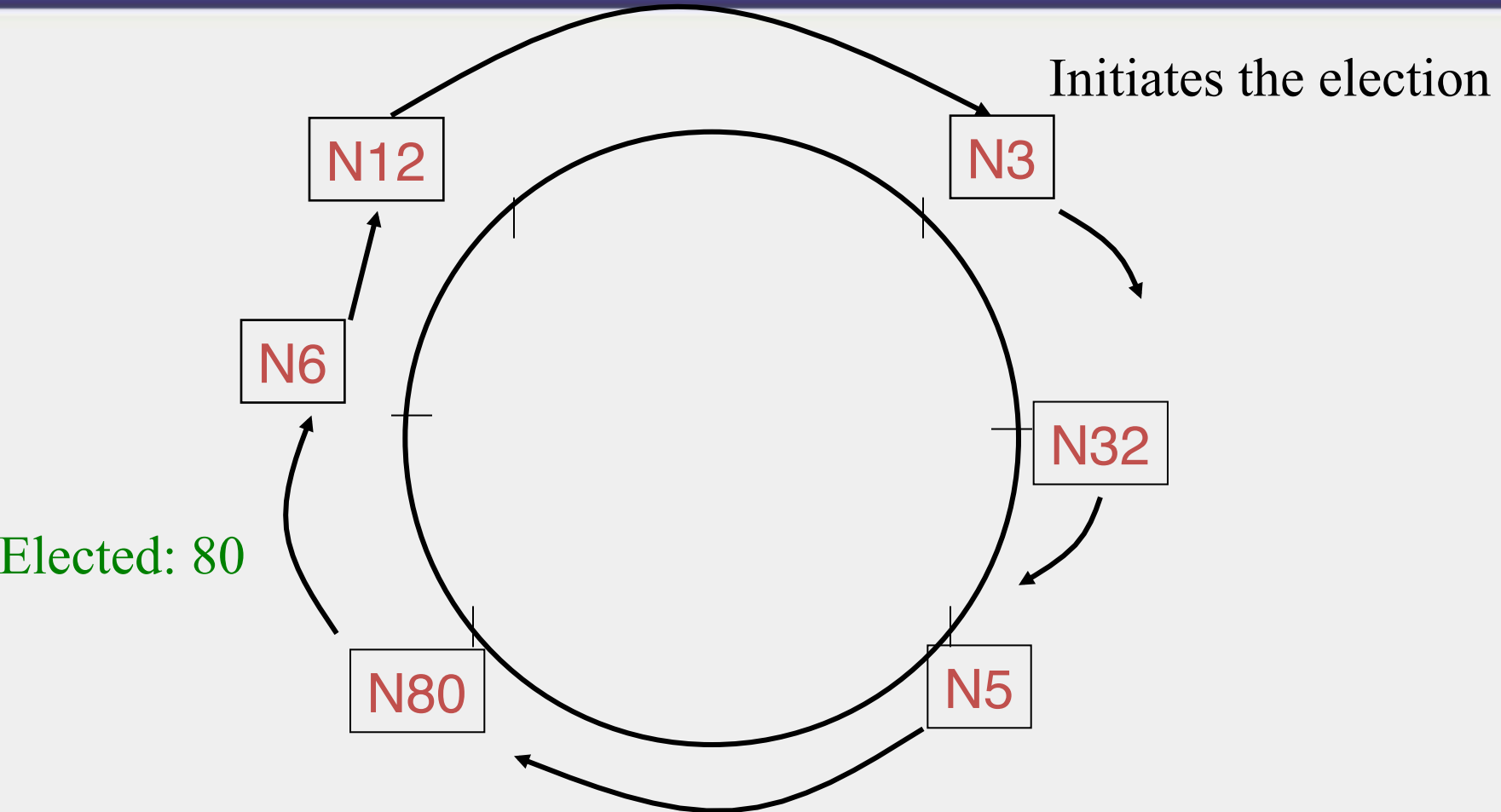




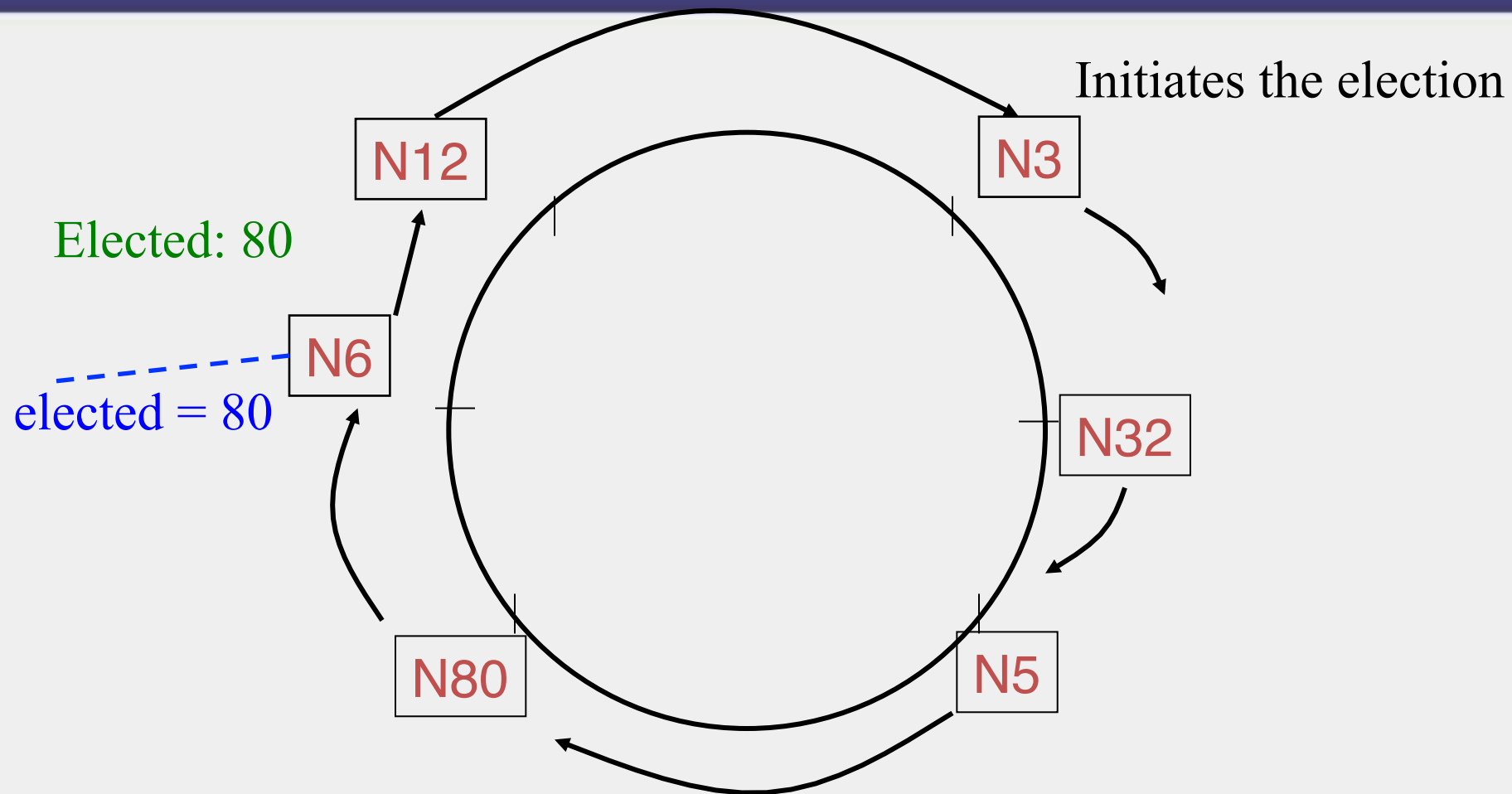
Goal: Elect highest id process as leader



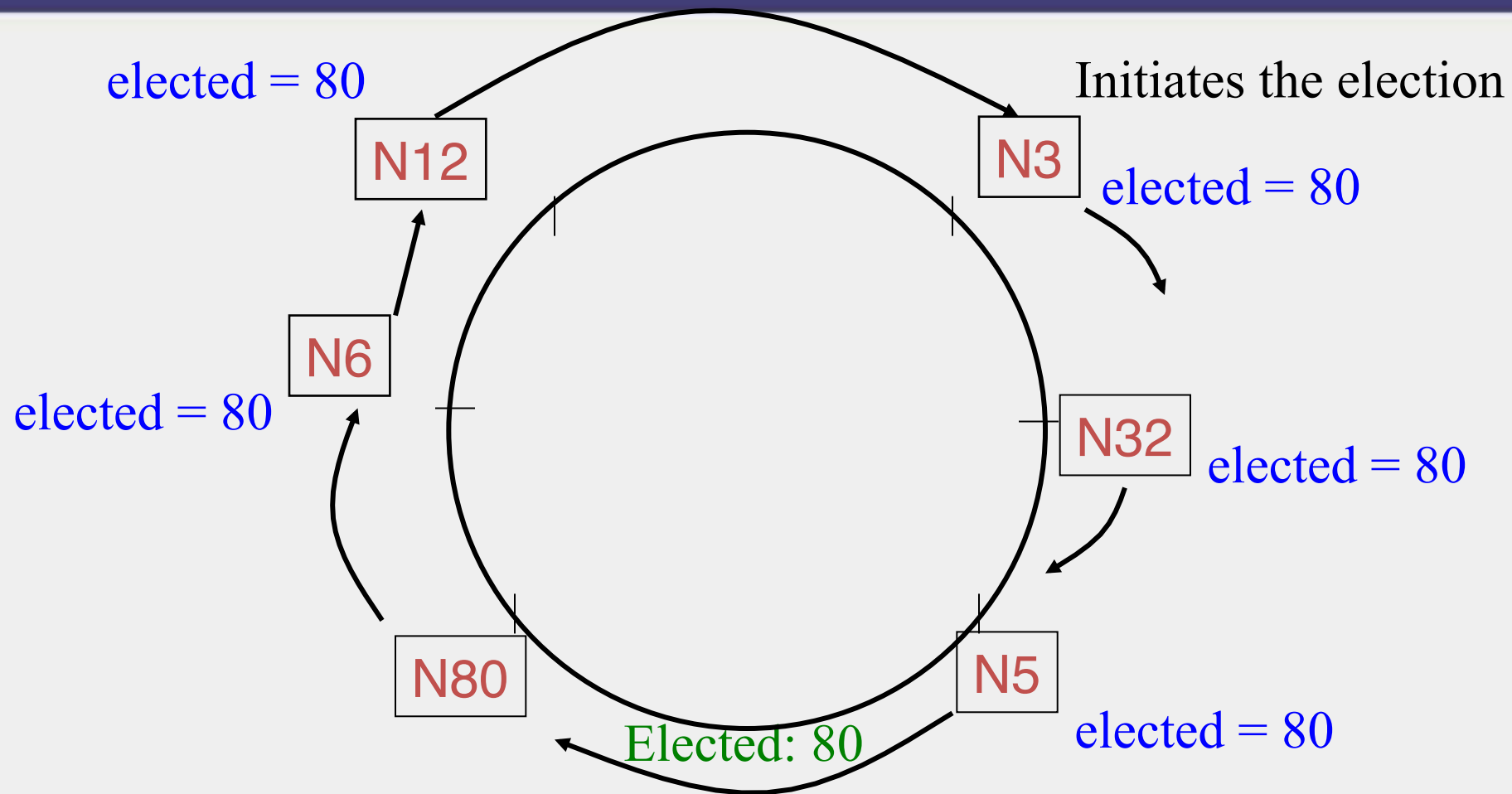
Goal: Elect highest id process as leader



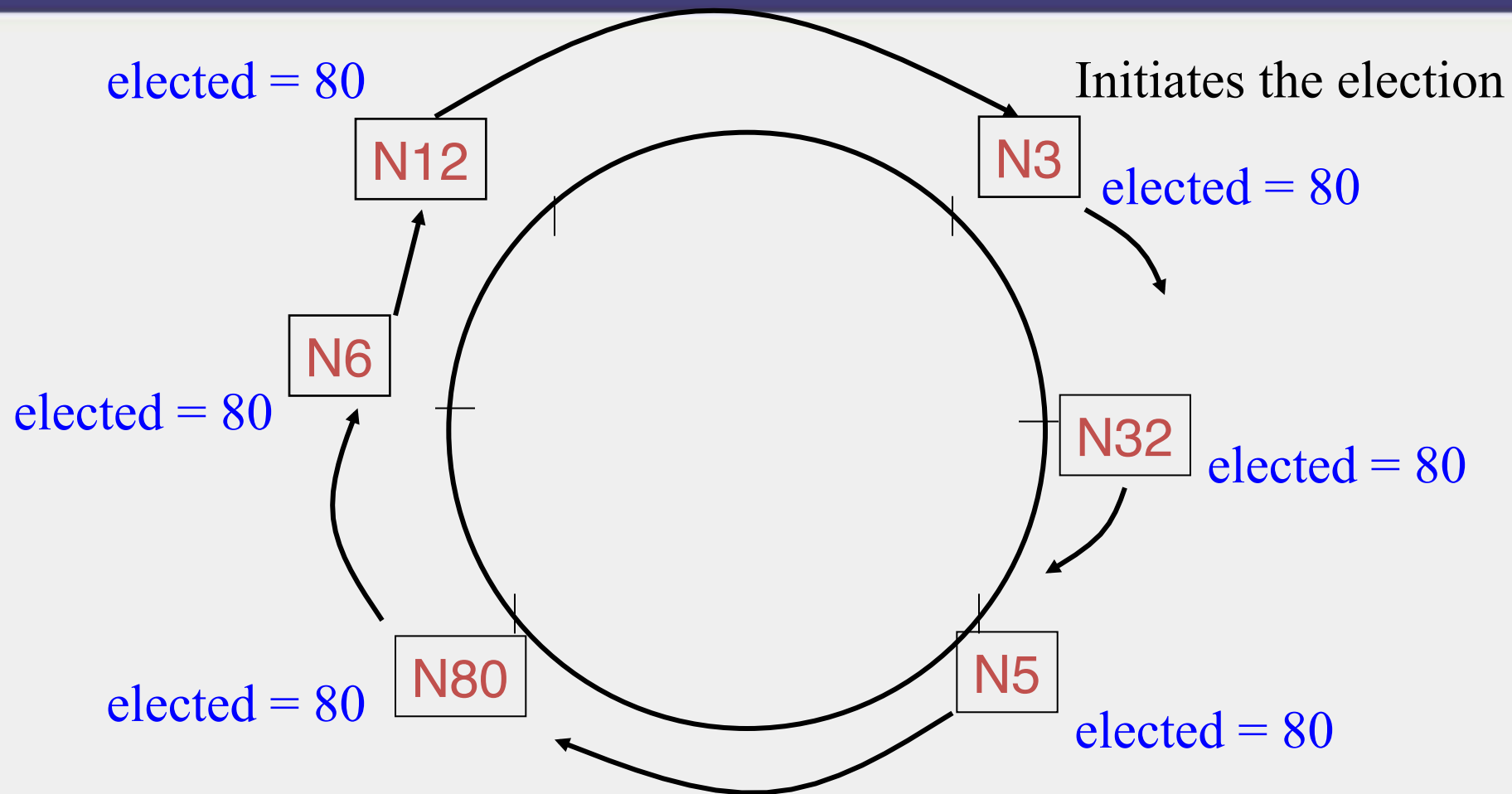
Goal: Elect highest id process as leader



Goal: Elect highest id process as leader



Goal: Elect highest id process as leader

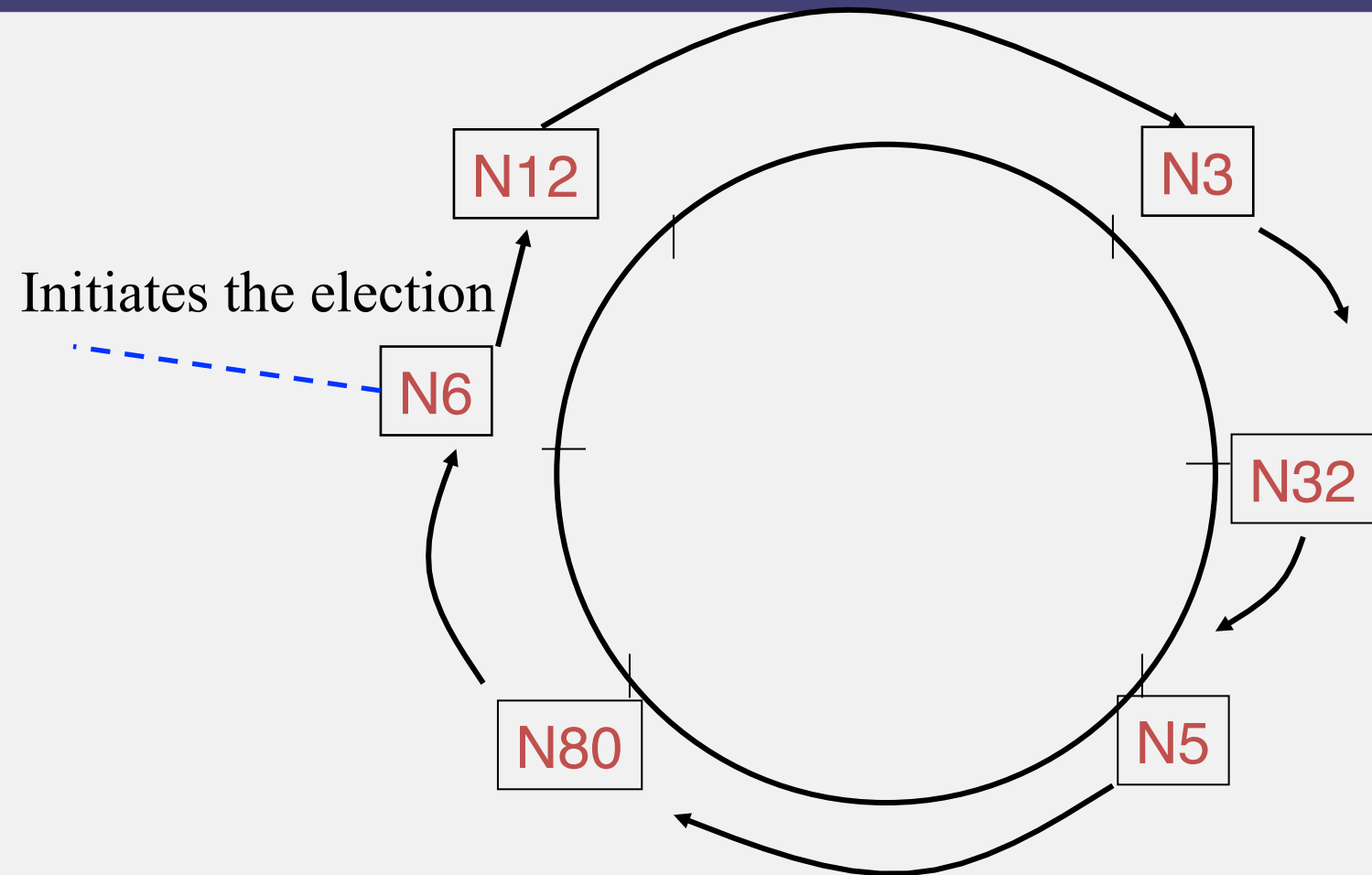


Goal: Elect highest id process as leader

# ANALYSIS

- Let's assume no failures occur during the election protocol itself, and there are  $N$  processes
- How many messages?
- Worst case occurs when the initiator is the ring successor of the would-be leader

# WORST-CASE



Goal: Elect highest id process as leader



# WORST-CASE ANALYSIS

- $(N-1)$  messages for Election message to get from Initiator (N6) to would-be coordinator (N80)
- $N$  messages for Election message to circulate around ring without message being changed
- $N$  messages for Elected message to circulate around the ring
- Message complexity:  $(3N-1)$  messages
- Completion time:  $(3N-1)$  message transmission times
- Thus, if there are no failures, election terminates (liveness) and everyone knows about highest-attribute process as leader (safety)

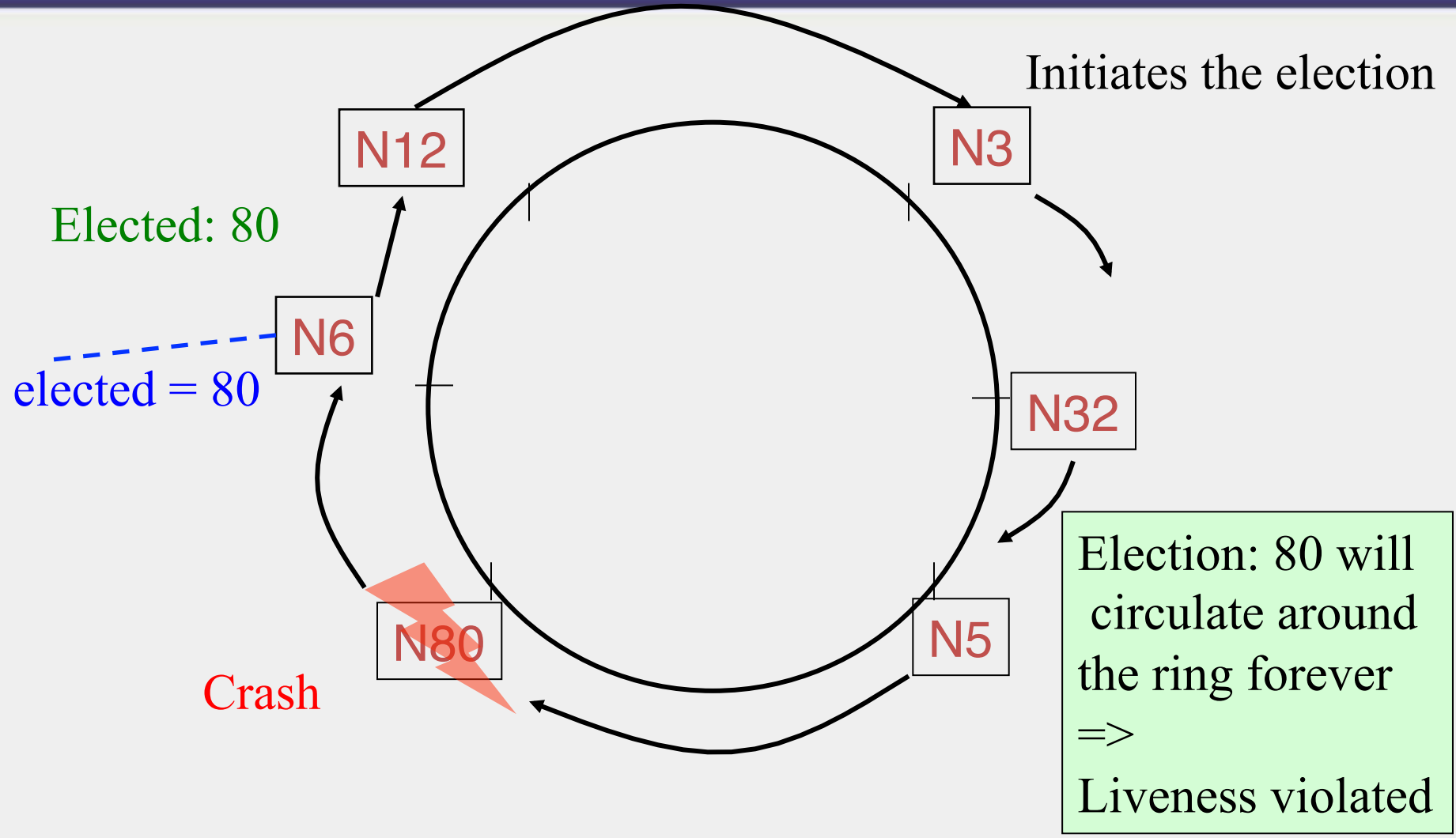
# BEST CASE?

- Initiator is the would-be leader, i.e., N80 is the initiator
- Message complexity:  $2N$  messages
- Completion time:  $2N$  message transmission times

# MULTIPLE INITIATORS?

- Include initiator's id with all messages
- Each process remembers in cache the initiator of each Election/Elected message it receives
- (All the time) Each process suppresses Election/Elected messages of any lower-id initiators
- Updates cache if receives higher-id initiator's Election/Elected message
- Result is that only the highest-id initiator's election run completes
- What about failures?

# EFFECT OF FAILURES



# FIXING FOR FAILURES

- One option: have predecessor (or successor) of would-be leader N80 detect failure and start a new election run
  - May re-initiate election if
    - Receives an Election message but times out waiting for an Elected message
    - Or after receiving the Elected:80 message
  - But what if predecessor also fails?
  - And its predecessor also fails? (and so on)

# FIXING FOR FAILURES (2)

- Second option: use the failure detector
- Any process, after receiving Election:80 message, can detect failure of N80 via its own local failure detector
  - If so, start a new run of leader election
- But failure detectors may not be both complete and accurate
  - Incompleteness in FD  $\Rightarrow$  N80's failure might be missed  $\Rightarrow$  Violation of Safety
  - Inaccuracy in FD  $\Rightarrow$  N80 mistakenly detected as failed
    - $\Rightarrow$  new election runs initiated forever
    - $\Rightarrow$  Violation of Liveness

# WHY IS ELECTION SO HARD?

- Because it is related to the consensus problem!
- If we could solve election, then we could solve consensus!
  - Elect a process, use its id's last bit as the consensus decision
- But since consensus is impossible in asynchronous systems, so is election!
- (later in lecture) Consensus-like protocols used in industry for leader election

# ANOTHER CLASSICAL ALGORITHM: BULLY ALGORITHM

- All processes know other process' ids
- When a process finds the coordinator has failed (via the failure detector):
  - **if** it knows its id is the highest
    - it elects itself as coordinator, then sends a *Coordinator* message to all processes with lower identifiers. Election is completed.
  - **else**
    - it initiates an election by sending an *Election* message
    - (contd.)



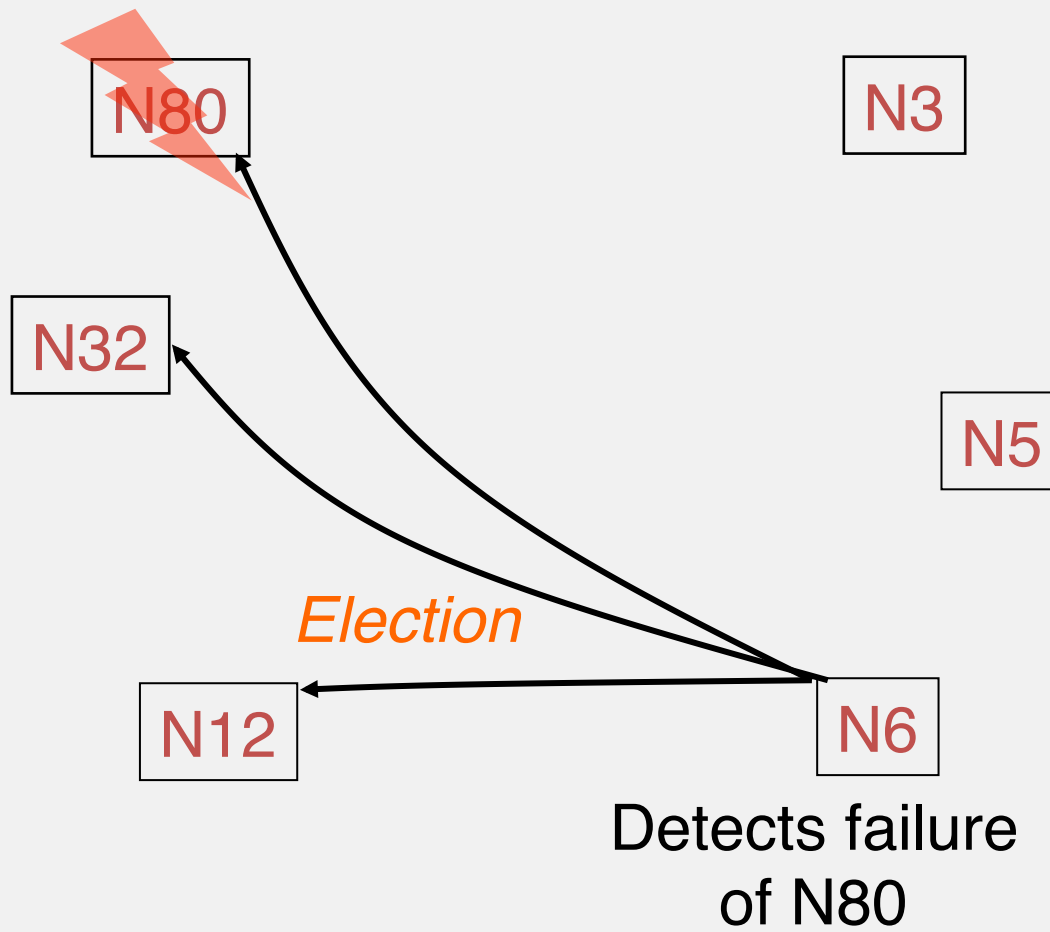
# BULLY ALGORITHM (2)

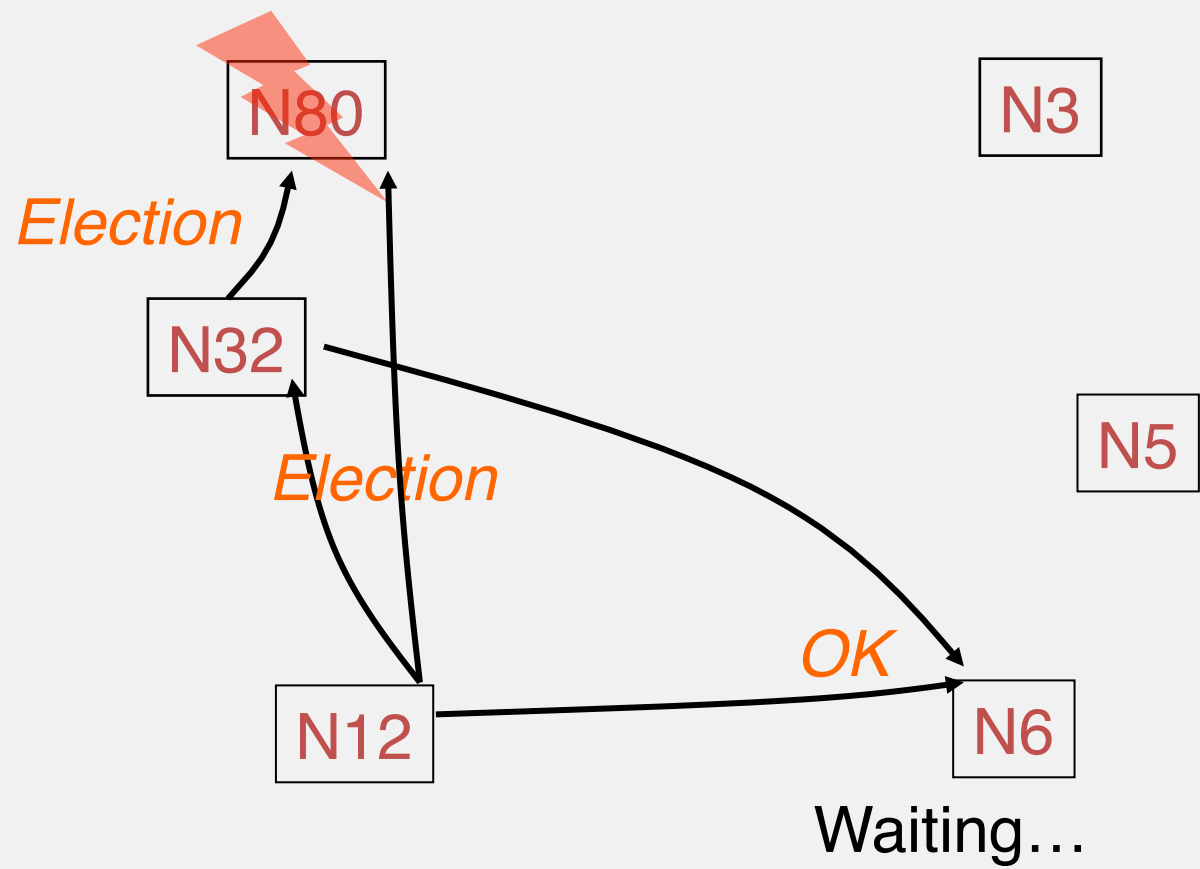
- **else** it initiates an election by sending an *Election* message
  - Sends it to only processes that have a *higher id than itself*.
  - **if** receives no answer within timeout, calls itself leader and sends *Coordinator* message to all lower id processes. Election completed.
  - **if** an answer received however, then there is some non-faulty higher process => so, wait for coordinator message. If none received after another timeout, start a new election run.
- A process that receives an *Election* message replies with *OK* message, and starts its own leader election protocol (unless it has already done so)

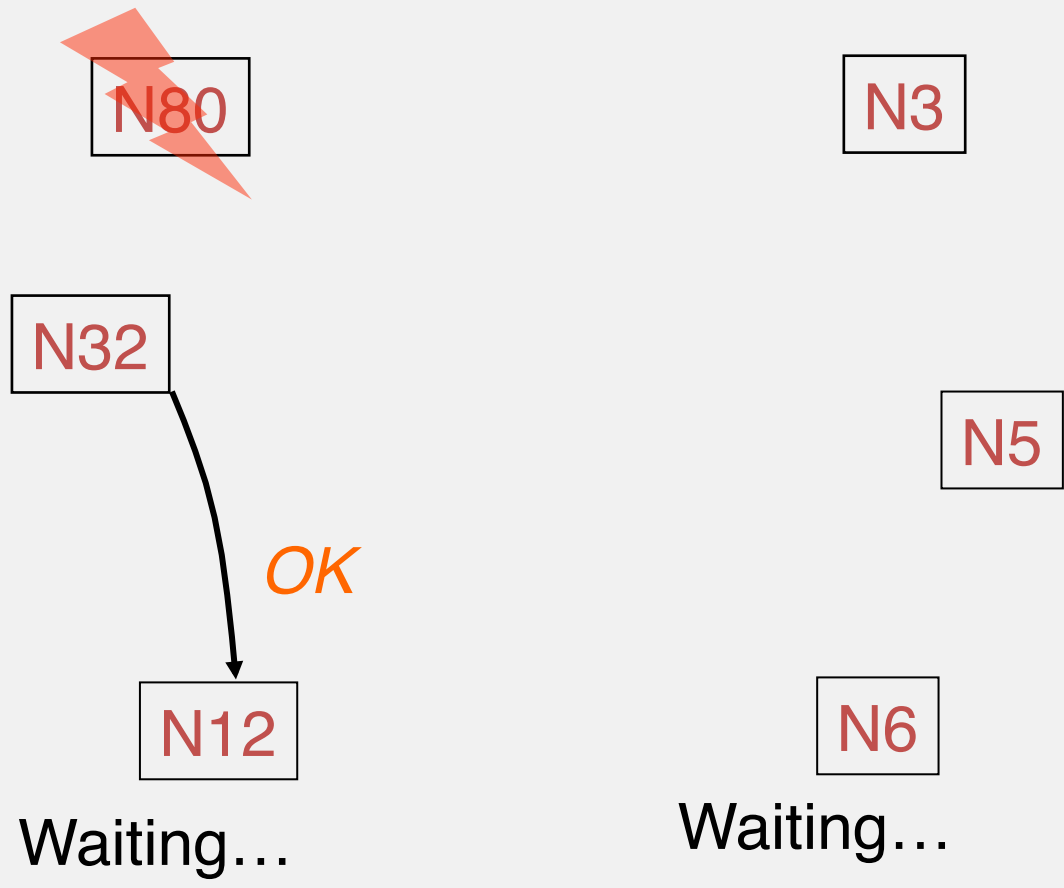
# BULLY ALGORITHM: EXAMPLE

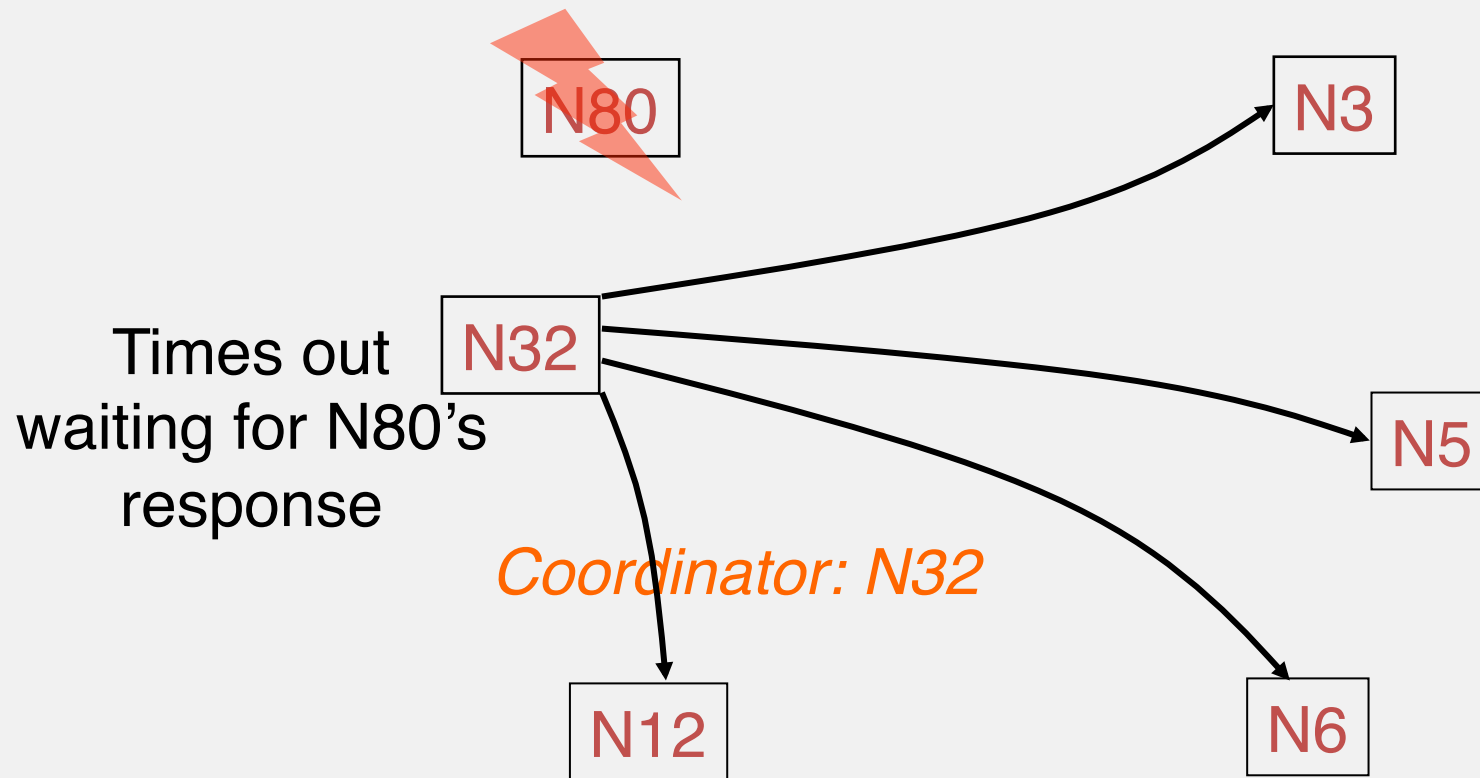


Detects failure  
of N80





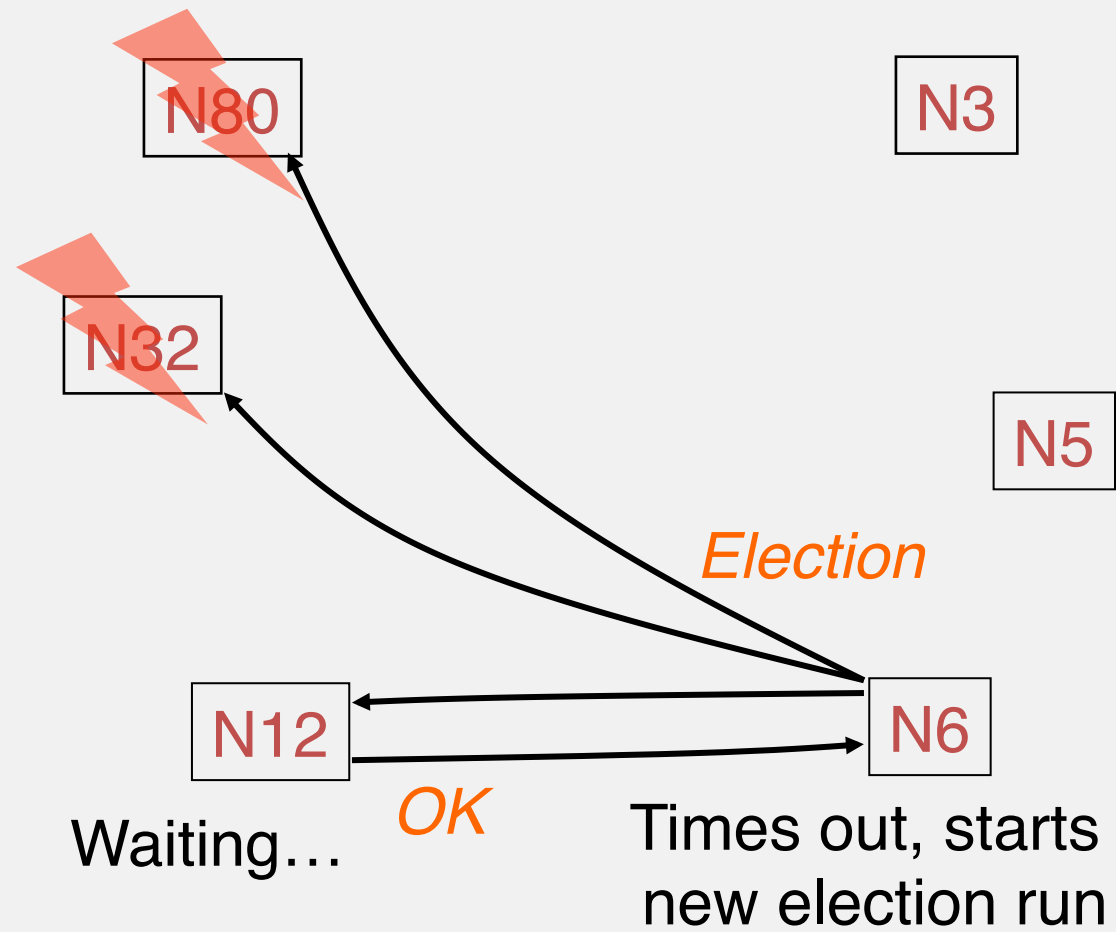




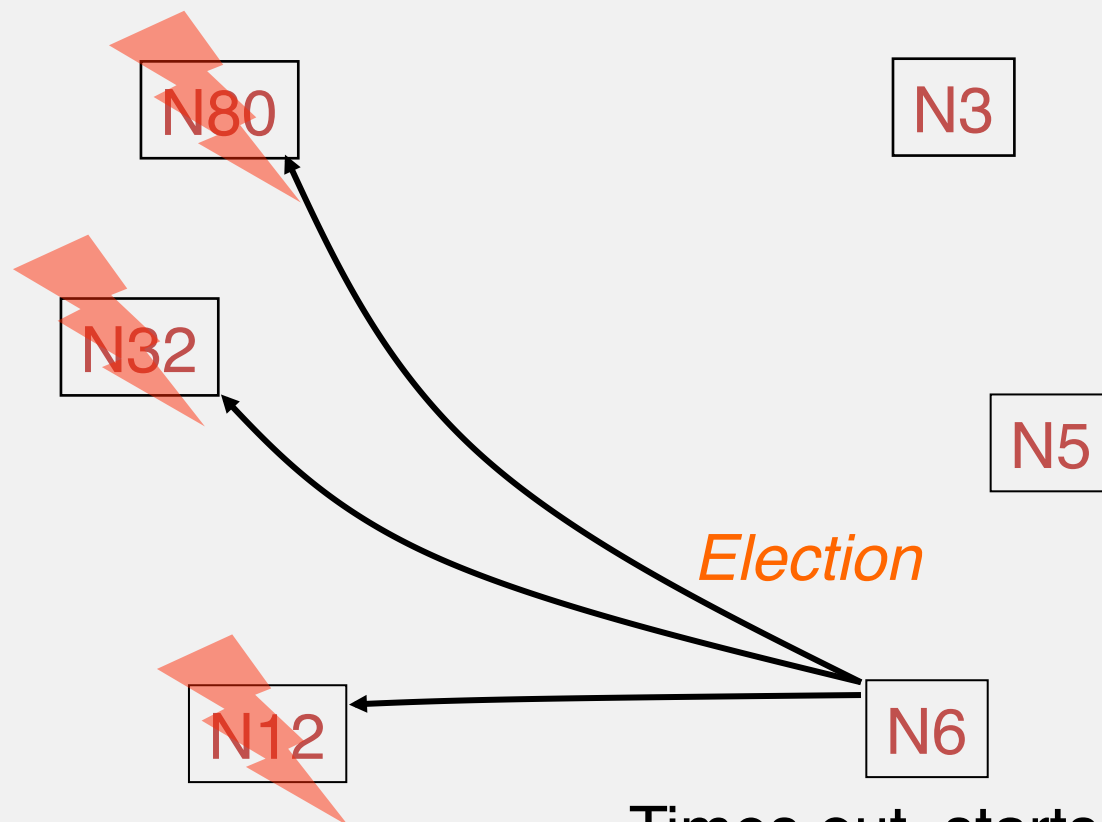
**Election is completed**

# FAILURES DURING ELECTION RUN









Times out, starts  
another new election run

# FAILURES AND TIMEOUTS

- If failures stop, eventually will elect a leader
- How do you set the timeouts?
- Based on Worst-case time to complete election
  - 5 message transmission times if there are no failures during the run:
    1. Election from lowest id server in group
    2. Answer to lowest id server from 2<sup>nd</sup> highest id process
    3. Election from 2nd highest id server to highest id
    4. Timeout for answers @ 2nd highest id server
    5. Coordinator from 2<sup>nd</sup> highest id server

# ANALYSIS

- **Worst-case** completion time: 5 message transmission times
  - When the process with the lowest id in the system detects the failure.
    - $(N-1)$  processes altogether begin elections, each sending messages to processes with higher ids.
    - $i$ -th highest id process sends  $(i-1)$  election messages
  - Number of Election messages  
 $= N-1 + N-2 + \dots + 1 = (N-1)*N/2 = O(N^2)$
- **Best-case**
  - Second-highest id detects leader failure
  - Sends  $(N-2)$  Coordinator messages
  - Completion time: 1 message transmission time

# IMPOSSIBILITY?

- Since timeouts built into protocol, in asynchronous system model:
  - Protocol may never terminate  $\Rightarrow$  Liveness not guaranteed
- But satisfies liveness in synchronous system model where
  - Worst-case one-way latency can be calculated = worst-case processing time + worst-case message latency

# CAN USE CONSENSUS TO SOLVE ELECTION

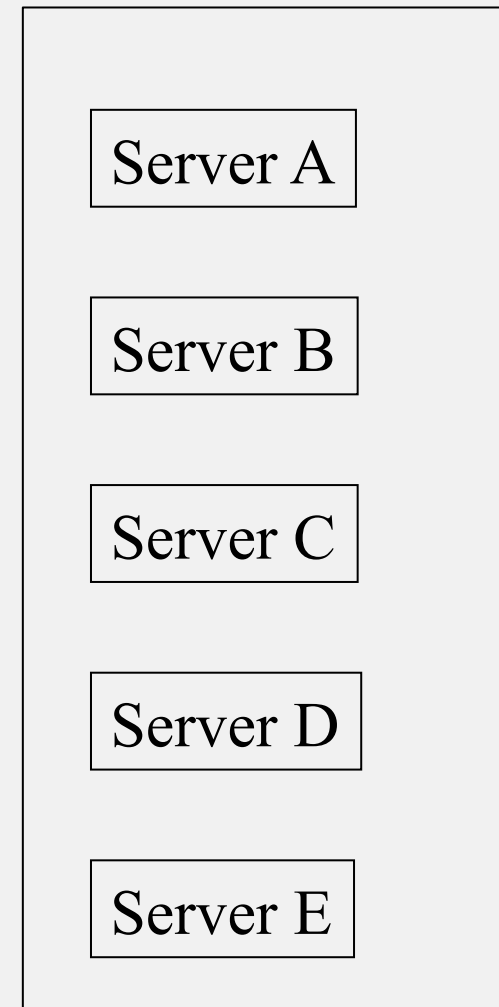
- One approach
  - Each process proposes a value
  - Everyone in group reaches consensus on some process  $P_i$ 's value
  - That lucky  $P_i$  is the new leader!

# ELECTION IN INDUSTRY

- Several systems in industry use Paxos-like approaches for election
  - Paxos is a consensus protocol (safe, but eventually live): elsewhere in this course
- Google's Chubby system
- Apache Zookeeper

# ELECTION IN GOOGLE CHUBBY

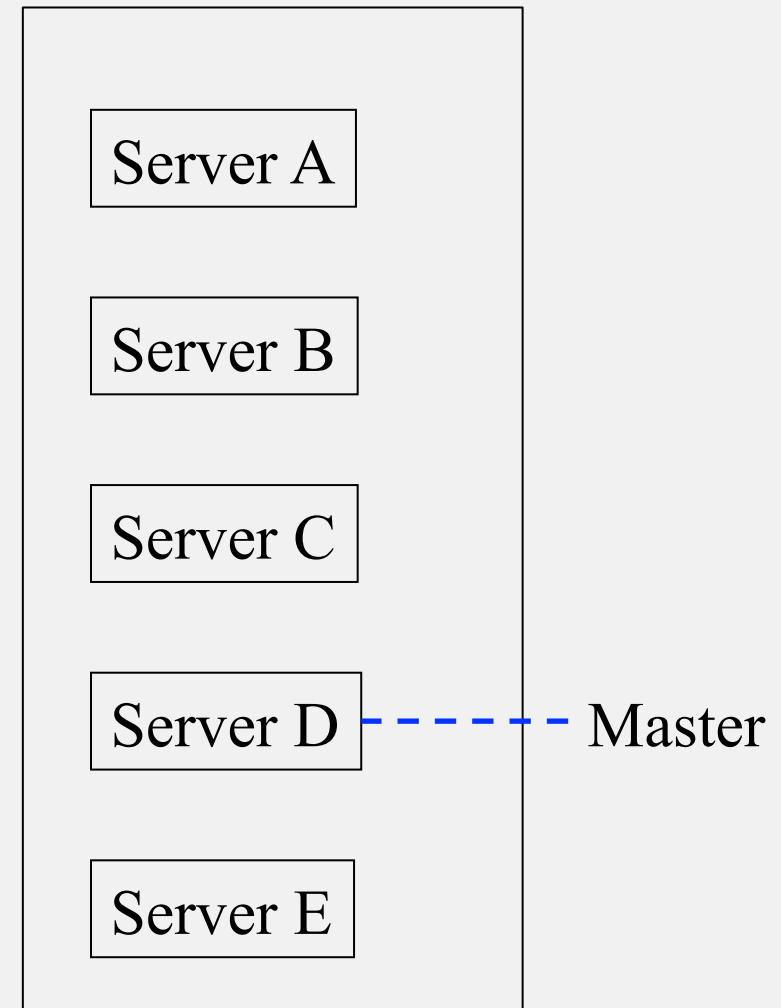
- A system for locking
- Essential part of Google's stack
  - Many of Google's internal systems rely on Chubby
  - BigTable, Megastore, etc.
- Group of replicas
  - Need to have a master server elected at all times



Reference: <http://research.google.com/archive/chubby.html>

# ELECTION IN GOOGLE CHUBBY (2)

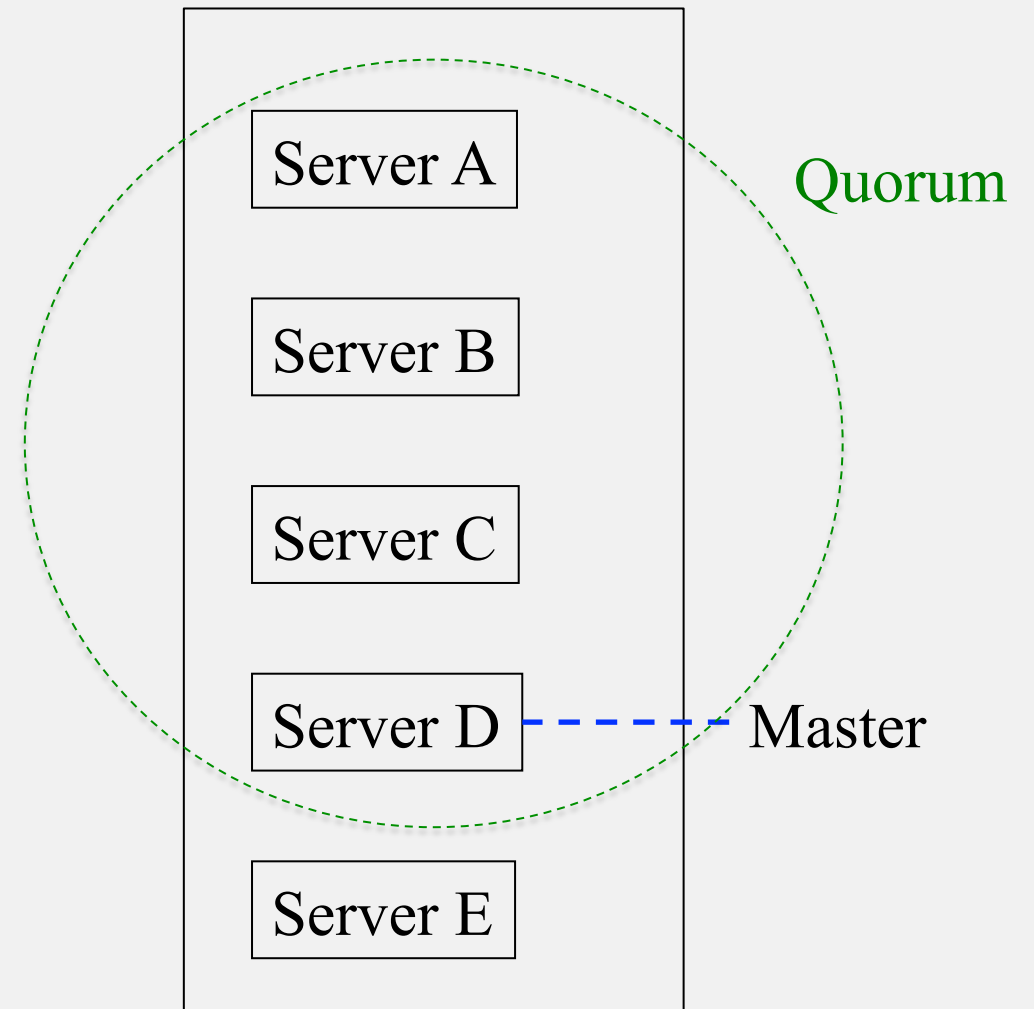
- Group of replicas
  - Need to have a master (i.e., leader)
- Election protocol
  - Potential leader tries to get votes from other servers
  - Each server votes for at most one leader
  - Server with *majority* of votes becomes new leader, informs everyone





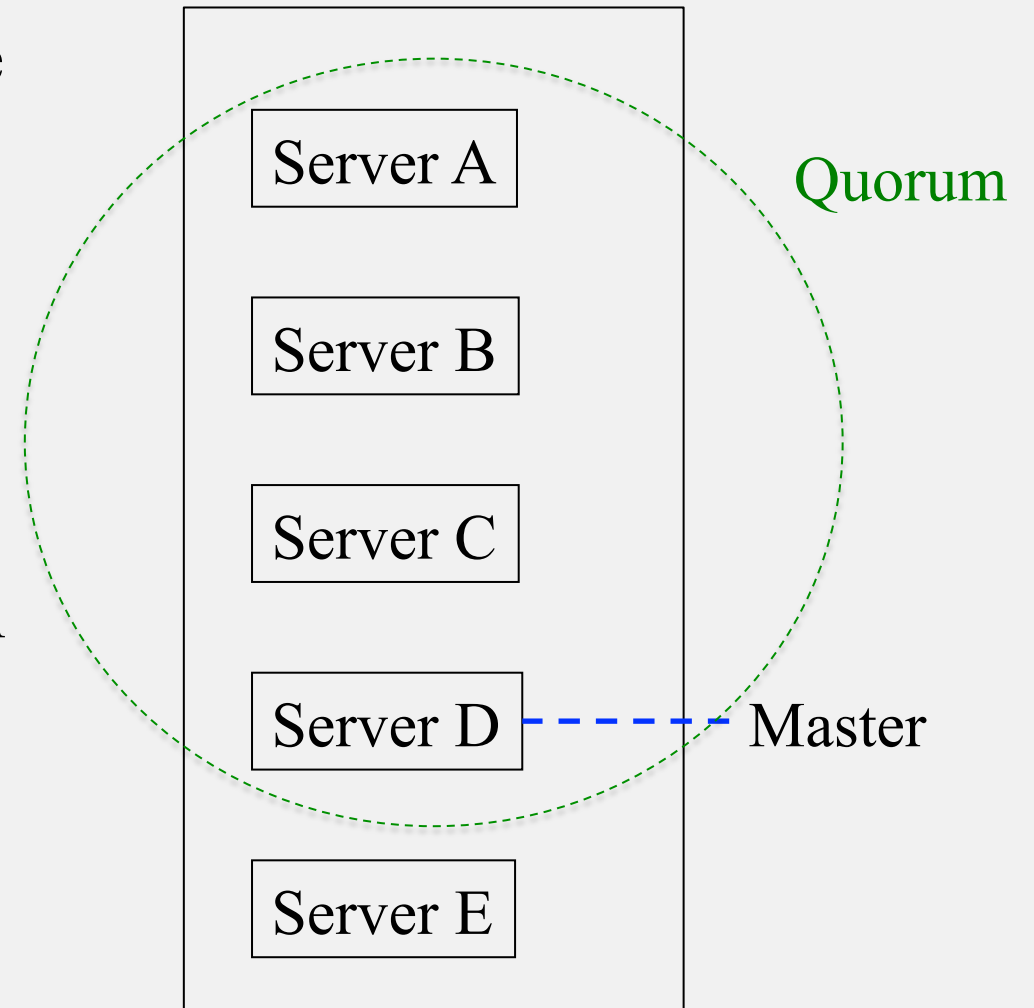
# ELECTION IN GOOGLE CHUBBY (3)

- Why **safe**?
  - Essentially, each potential leader tries to reach a *quorum*
  - Since any two quorums intersect, and each server votes at most once, cannot have two leaders elected simultaneously
- Why **live**?
  - Only eventually live! Failures may keep happening so that no leader is ever elected
  - In practice: elections take a few seconds.  
Worst-case noticed by Google: 30 s



# ELECTION IN GOOGLE CHUBBY (4)

- After election finishes, other servers promise not to run election again for “a while”
  - “While” = time duration called “Master lease”
  - Set to a few seconds
- Master lease can be renewed by the master as long as it continues to win a majority each time
- Lease technique ensures automatic re-election on master failure



# ELECTION IN ZOOKEEPER

- Centralized service for maintaining configuration information
- Uses a variant of Paxos called Zab (Zookeeper Atomic Broadcast)
- Needs to keep a leader elected at all times
- <http://zookeeper.apache.org/>

# ELECTION: SUMMARY

- Leader election an important component of many cloud computing systems
- Classical leader election protocols
  - Ring-based
  - Bully
- But failure-prone
  - Paxos-like protocols used by Google Chubby, Apache Zookeeper