CSE 421/521 - Operating Systems
Fall 2018

Lecture - XXII

Distributed Systems - I

Tevfik Koşar

University at Buffalo

November 20th, 2018

Motivation

- **Distributed system** is collection of loosely coupled processors that
  - do not share memory
  - interconnected by a communications network

- Reasons for distributed systems
  - Resource sharing
    - sharing and printing files at remote sites
    - processing information in a distributed database
    - accessing remote files
    - using remote specialized hardware devices
  - Computation speedup – **load sharing**
  - Reliability – detect and recover from site failure, function transfer, reintegrate failed site

Distributed-Operating Systems

- Users not aware of multiplicity of machines
  - Access to remote resources similar to access to local resources
- Data Migration – transfer data by transferring entire file, or transferring only those portions of the file necessary for the immediate task
- Computation Migration – transfer the computation, rather than the data, across the system
  - Process Migration
  - VM migration

Distributed-Operating Systems (Cont.)

- Process Migration – execute an entire process, or parts of it, at different sites
  - Load balancing – distribute processes across network to even the workload
  - Computation speedup – subprocesses can run concurrently on different sites
  - Hardware preference – process execution may require specialized processor
  - Software preference – required software may be available at only a particular site
  - Data access – run process remotely, rather than transfer all data locally

Distributed File Systems

- Distributed file system (**DFS**) – a distributed implementation of the classical time-sharing model of a file system, where multiple users share files and storage resources over a network

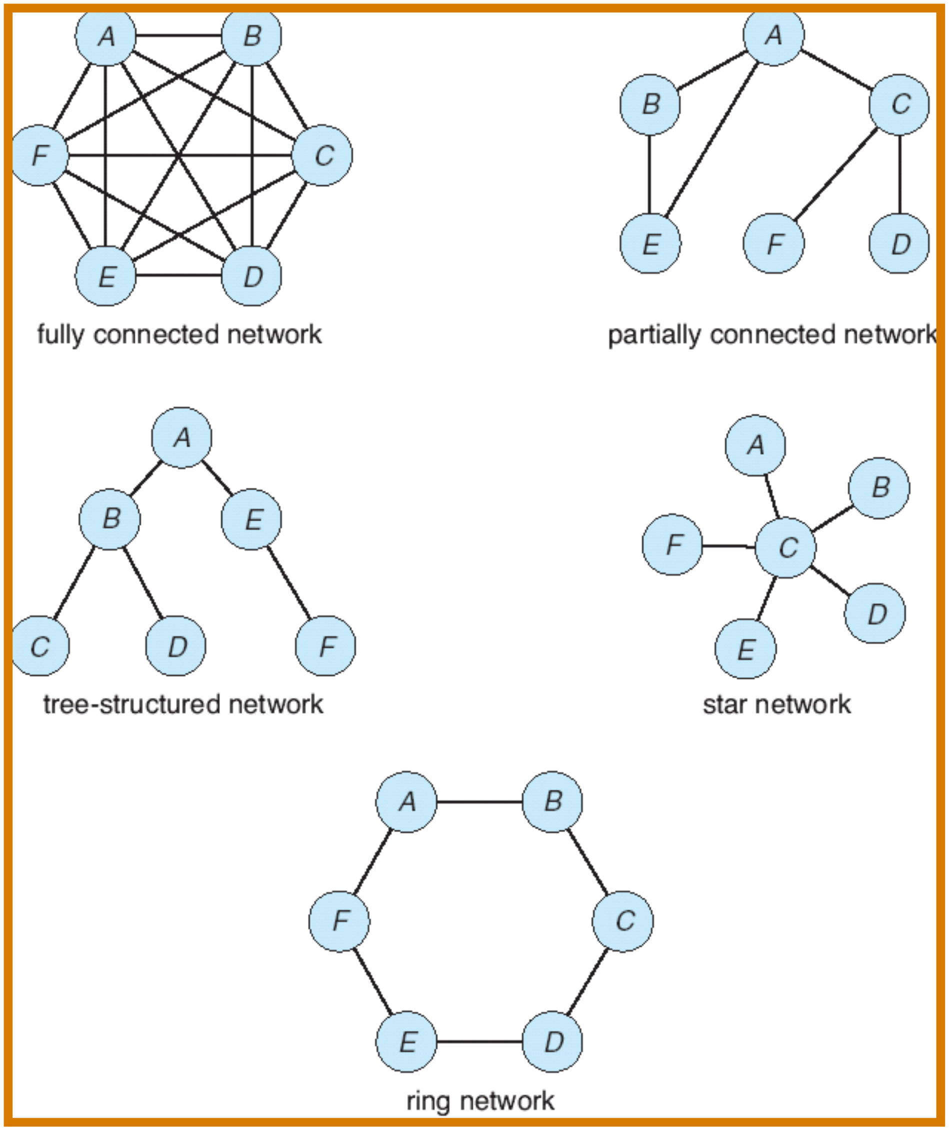
- A DFS manages set of dispersed storage devices

- Overall storage space managed by a DFS is composed of different, remotely located, smaller storage spaces

- There is usually a correspondence between constituent storage spaces and sets of files

- i.e., NFS, AFS, GFS

Distributed Network Topology



A Large-scale Distributed System: Google Search Infrastructure

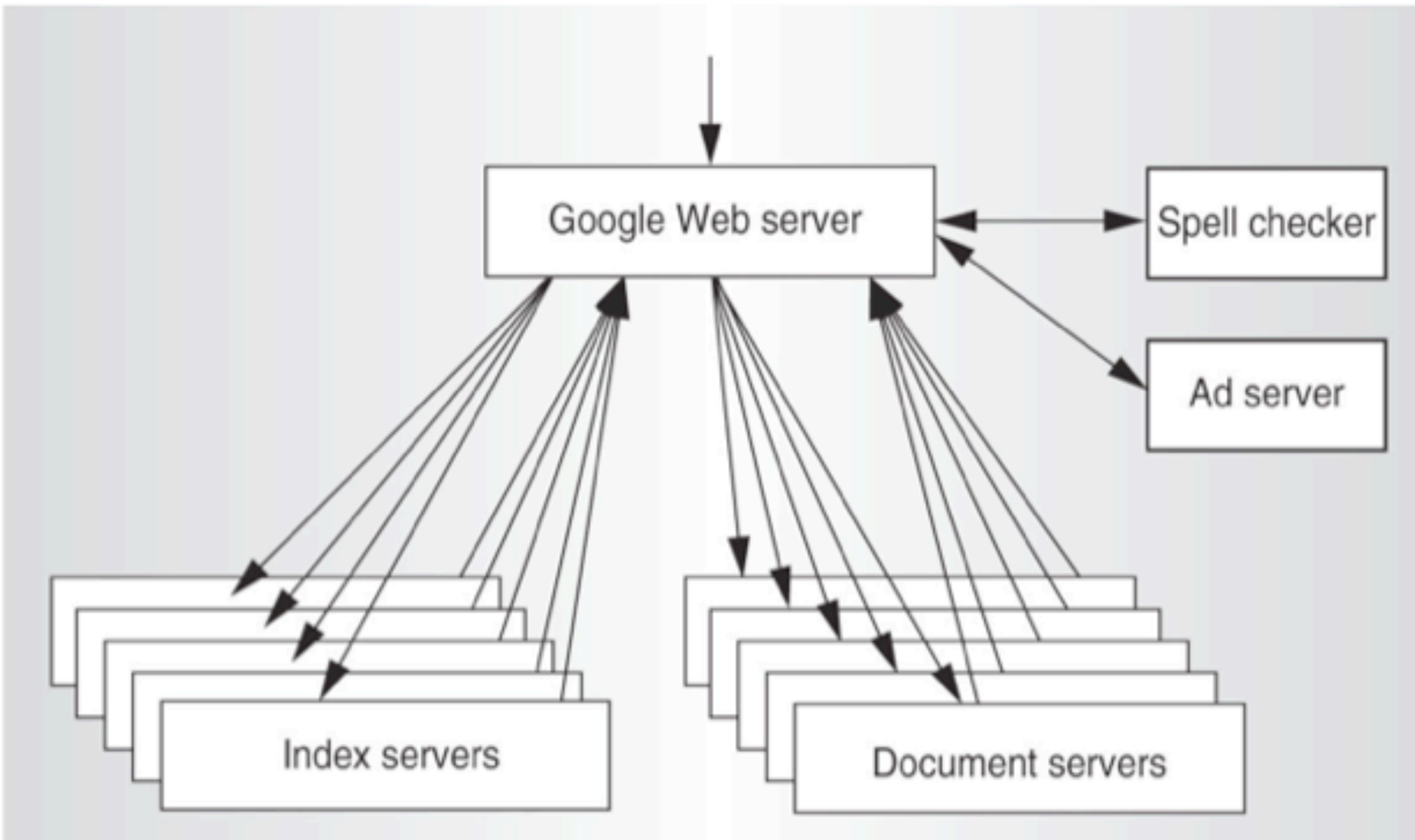
- It's likely that Google has several million machines
  - But let's be conservative – 1,000,000 machines
  - A rack holds 176 CPUs (88 1U dual-processor boards), so that's about 6,000 racks
  - A rack requires about 50 square feet (given datacenter cooling capabilities), so that's about 300,000 square feet of machine room space (more than 6 football fields of real estate – although of course Google divides its machines among dozens of datacenters all over the world)
  - A rack requires about 10kw to power, and about the same to cool, so that's about 120,000 kw of power, or nearly 100,000,000 kwh per month ($10 million at $0.10/kwh)
    - Equivalent to about 20% of Seattle City Light's generating capacity

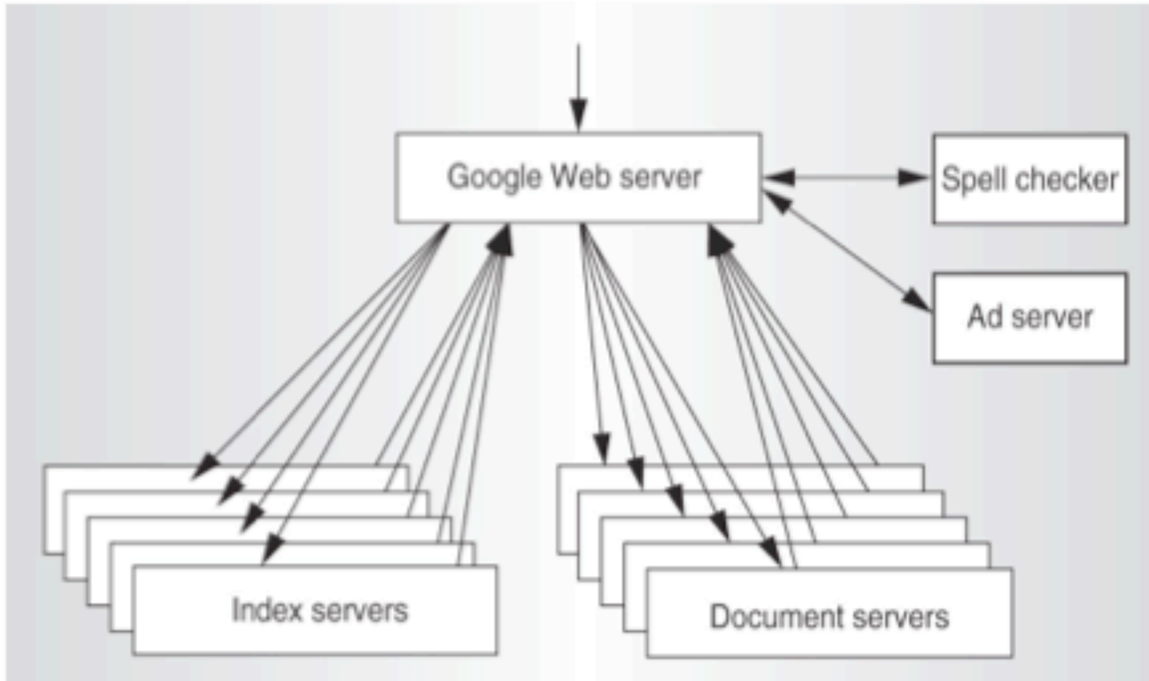
- There are multiple clusters (of thousands of computers each) all over the world

- *Many hundreds of machines are involved in a <u>single</u> Google search request* (remember, the web is 400+TB)

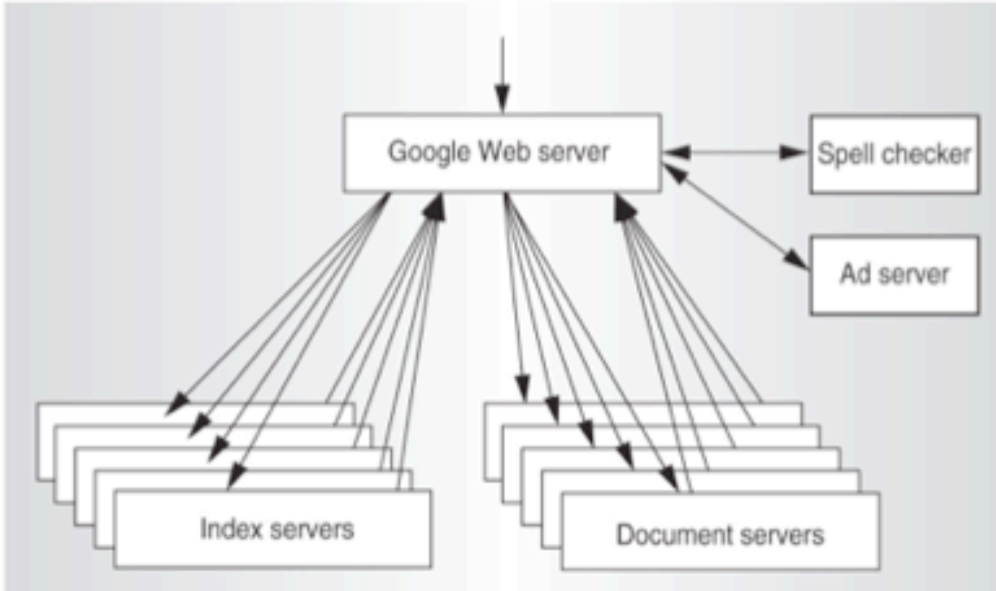
1. DNS routes your search request to a nearby cluster

- A cluster consists of Google Web Servers, Index Servers, Doc Servers, and various other servers (ads, spell checking, etc.)
  - These are cheap standalone computers, rack-mounted, connected by commodity networking gear



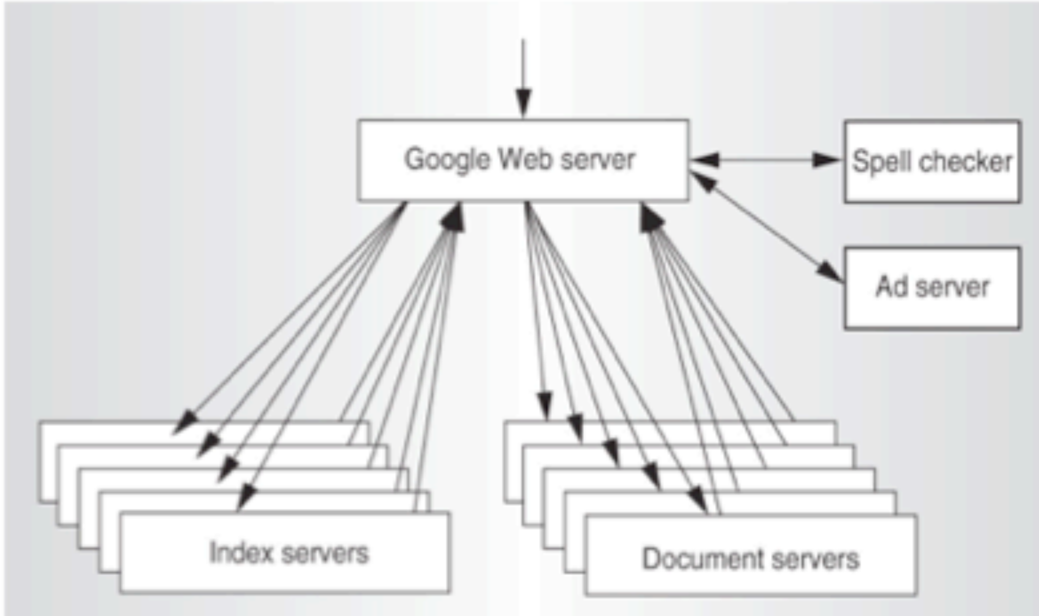


2. Within the cluster, load-balancing routes your search to a lightly-loaded Google Web Server (GWS), which will coordinate the search and response



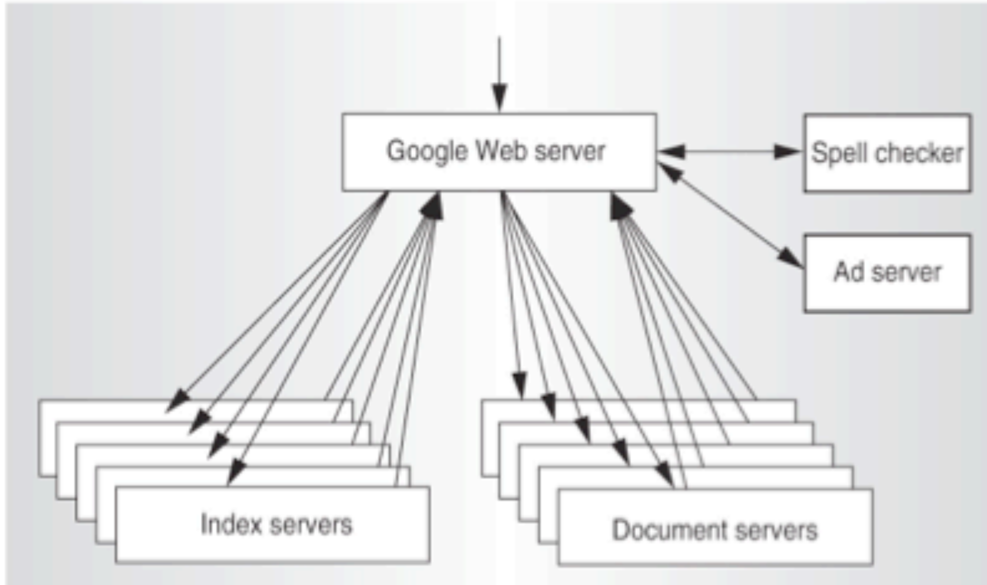
- The index is partitioned into "shards." Each shard indexes a subset of the docs (web pages). Each shard is replicated, and can be searched by multiple computers – "index servers"

3. The GWS routes your search to one index server associated with each shard, through another load-balancer
4. When the dust has settled, the result is an ID for every doc satisfying your search, rank-ordered by relevance



- The docs, too, are partitioned into "shards" – the partitioning is a hash on the doc ID. Each shard contains the full text of a subset of the docs. Each shard can be searched by multiple computers – "doc servers"

5. The GWS sends appropriate doc IDs to one doc server associated with each relevant shard
6. When the dust has settled, the result is a URL, a title, and a summary for every relevant doc



7. Meanwhile, the ad server has done its job, the spell checker has done its job, etc.
8. The GWS builds an HTTP response to your search and ships it off

- Many hundreds of computers have enabled you to search 400+TB of web in ~100 ms.

Google: The Big Picture

- Enormous volumes of distributed data
- Extreme parallelism
- The cheapest imaginable components
  - Failures occur all the time
  - You could not afford to prevent this in hardware
- Software makes it
  - Fault-Tolerant
  - Highly Available
  - Recoverable
  - Consistent
  - Scalable
  - Predictable
  - Secure

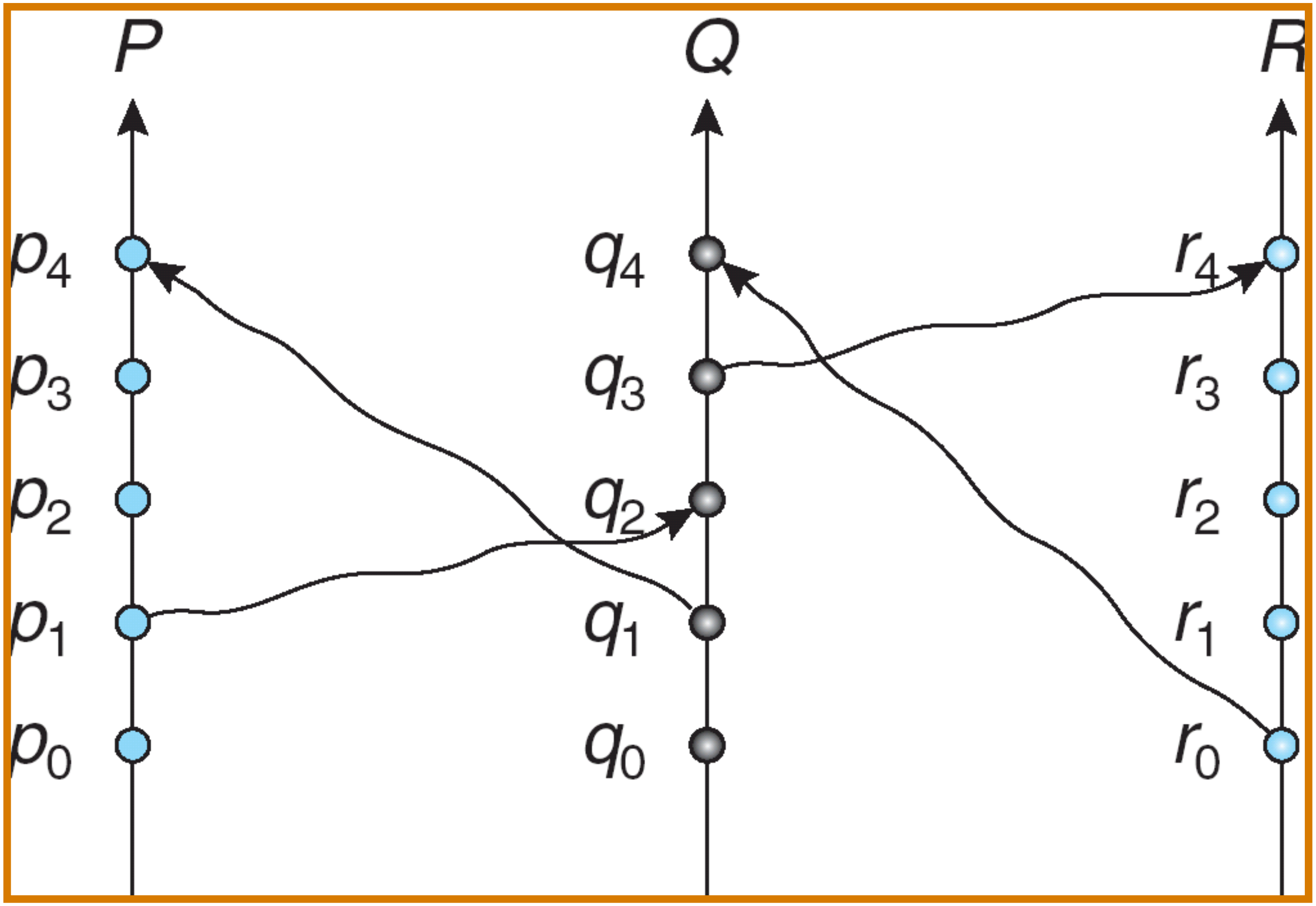
Distributed Coordination

- Ordering events and achieving synchronization in centralized systems is easier.
  - We can use common clock and memory
- What about distributed systems?
  - No common clock or memory
  - *happened-before* relationship provides partial ordering
  - How to provide total ordering?

Event Ordering

- Happened-before relation (denoted by $\rightarrow$)
  - If A and B are events in the same process (assuming sequential processes), and A was executed before B, then $A \rightarrow B$
  - If A is the event of sending a message by one process and B is the event of receiving that message by another process, then $A \rightarrow B$
  - If $A \rightarrow B$ and B $\rightarrow C$ then $A \rightarrow C$
  - If two events A and B are not related by the $\rightarrow$ relation, then these events are executed concurrently.

Relative Time for Three Concurrent Processes

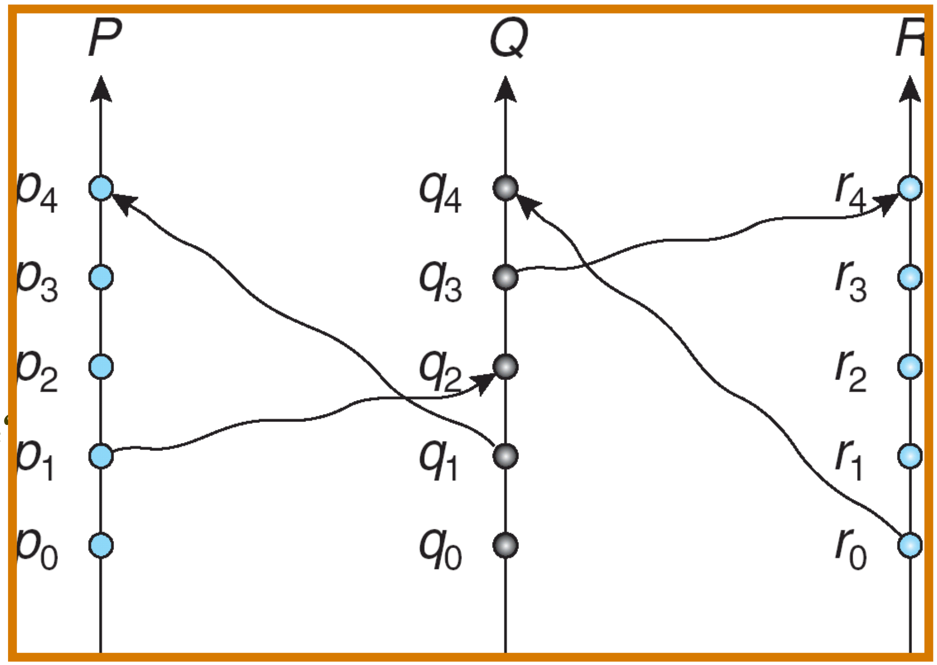


Which events are concurrent and which ones are ordered?

Exercise

Which of the following event orderings are true?

(a) p0 --> p3 :
(b) p1 --> q3 :
(c) q0 --> p3 :
(d) r0 --> p4 :
(e) p0 --> r4 :



Which of the following statements are true?

(a) p2 and q2 are concurrent processes.
(b) q1 and r1 are concurrent processes.
(c) p0 and q3 are concurrent processes.
(d) r0 and p0 are concurrent processes.
(e) r0 and p4 are concurrent processes.

Implementation of $\rightarrow$

- Associate a timestamp with each system event
  - Require that for every pair of events A and B, if A $\rightarrow$ B, then the timestamp of A is less than the timestamp of B
- Within each process Pi, define a **logical clock**
  - The logical clock can be implemented as a simple counter that is incremented between any two successive events executed within a process
    - Logical clock is **monotonically increasing**
- A process advances its logical clock when it receives a message whose timestamp is greater than the current value of its logical clock
  - Assume A sends a message to B, $LC_1(A)=200$, $LC_2(B)=195$ --> $LC_2(B)=201$
- If the timestamps of two events A and B are the same, then the events are concurrent
  - We may use the process identity numbers to break ties and to create a total ordering

Distributed Mutual Exclusion (DME)

- Assumptions
  - The system consists of *n* processes; each process P_i resides at a different processor
  - Each process has a critical section that requires mutual exclusion
- Requirement
  - If P_i is executing in its critical section, then no other process P_j is executing in its critical section
- We present different approaches to ensure the mutual exclusion of processes in their critical sections

1. Centralized Approach

- One of the processes in the system is chosen to coordinate the entry to the critical section
- A process that wants to enter its critical section sends a request message to the coordinator
- The coordinator decides which process can enter the critical section next, and it sends that process a reply message
- When the process receives a reply message from the coordinator, it enters its critical section
- After exiting its critical section, the process sends a release message to the coordinator and proceeds with its execution
- This scheme requires three messages per critical-section entry: ***request***, ***reply***, ***release***
- Single point of failure! -> would need a new coordinator to be elected..

2. Fully Distributed Approach

- When process P_i wants to enter its critical section, it generates a new timestamp, *TS*, and sends the message *request* (P_i, *TS*) to all processes in the system
- When process P_j receives a *request* message, it may reply immediately or it may defer sending a reply back
- When process P_i receives a *reply* message from all other processes in the system, it can enter its critical section
- After exiting its critical section, the process sends *reply* messages to all its deferred requests

Fully Distributed Approach (Cont.)

- The decision whether process P_j replies immediately to a *request*(P_i, *TS*) message or defers its reply is based on three factors:
  - If P_j is in its critical section, then it defers its reply to P_i
  - If P_j does *not* want to enter its critical section, then it sends a *reply* immediately to P_i
  - If P_j wants to enter its critical section but has not yet entered it, then it compares its own request timestamp with the timestamp *TS*
    - If its own request timestamp is greater than *TS*, then it sends a *reply* immediately to P_i (P_i asked first)
    - Otherwise, the reply is deferred

  - Example: P1 sends a request to P2 and P3 (timestamp=10)
    P3 sends a request to P1 and P2 (timestamp=4)

Undesirable Consequences

- The processes need to know the identity of all other processes in the system, which makes the dynamic addition and removal of processes more complex

- If one of the processes fails, then the entire scheme collapses
  - This can be dealt with by continuously monitoring the state of all the processes in the system, and notifying all processes if a process fails

==> More suitable for small, stable set of coordinating processes

3. Token-Passing Approach

- Circulate a token among processes in system
  - **Token** is special type of message
  - Possession of token entitles holder to enter critical section
- Processes *logically* organized in a **ring structure**
- Unidirectional ring guarantees freedom from starvation
- Two types of failures
  - Lost token – election must be called
  - Failed processes – new logical ring established

Election Algorithms

- Determine where a new copy of the coordinator should be restarted
- Assume that a unique priority number is associated with each active process in the system, and assume that the priority number of process P_i is i
- The coordinator is always the process with the highest priority number. When a coordinator fails, the algorithm must elect that active process with the largest priority number
- Two algorithms, the bully algorithm and a ring algorithm, can be used to elect a new coordinator in case of failures

1. Bully Algorithm

- Applicable to systems where every process can send a message to every other process in the system
- If process P_i sends a request that is not answered by the coordinator within a time interval *T*, assume that the coordinator has failed; P_i tries to elect itself as the new coordinator
- P_i sends an election message to every process with a higher priority number, P_i then waits for any of these processes to answer within *T*

Bully Algorithm (Cont.)

- If no response within *T*, assume that all processes with numbers greater than i have failed; P_i elects itself the new coordinator

- If answer is received, P_i begins time interval T', waiting to receive a message that a process with a higher priority number has been elected

- If no message is sent within T', assume the process with a higher number has failed; P_i should restart the algorithm

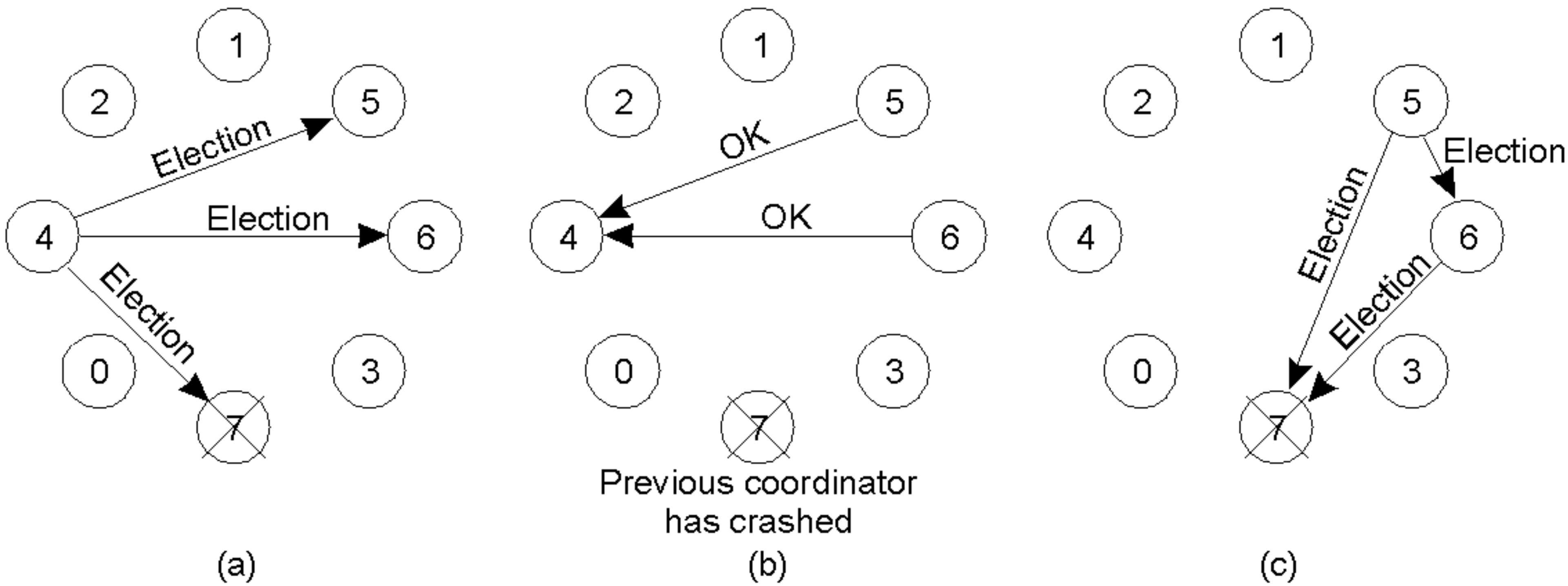
Bully Algorithm (Cont.)

- If P_i is not the coordinator, then, at any time during execution, P_i may receive one of the following two messages from process P_j
  - P_j is the new coordinator ($j > i$). P_i, in turn, records this information
  - P_j started an election ($j < i$). P_i, sends a response to P_j and begins its own election algorithm, provided that *Pi* has not already initiated such an election

- After a failed process recovers, it immediately begins execution of the same algorithm

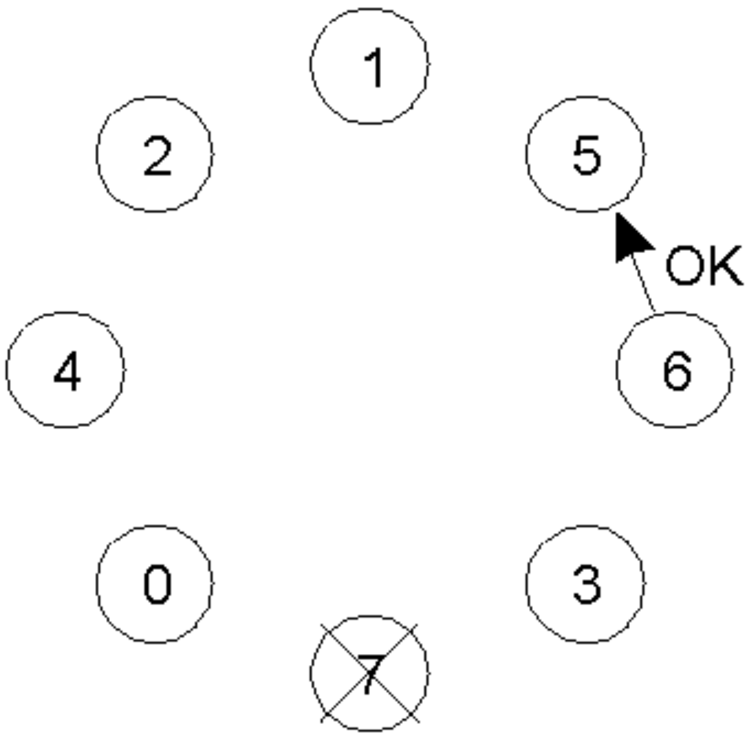
- If there are no active processes with higher numbers, the recovered process forces all processes with lower number to let it become the coordinator process, even if there is a currently active coordinator with a lower number

The Bully Algorithm (Example)

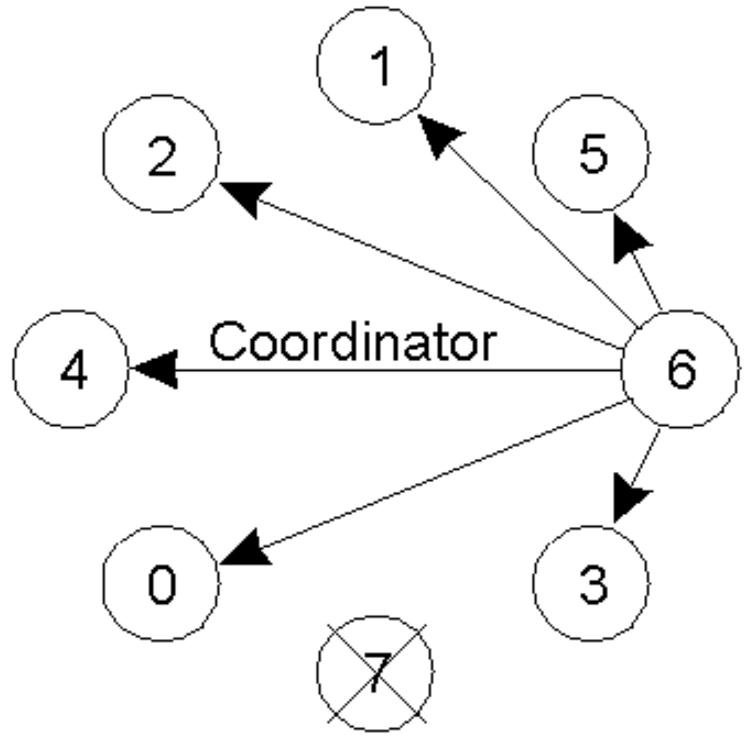


- Process 4 holds an election
- Process 5 and 6 respond, telling 4 to stop
- Now 5 and 6 each hold an election

The Bully Algorithm (Example)



(d)



(e)

Bully Algorithm Analysis

- Best case
  - The node with second highest identifier detects failure
  - Total messages = N-2
    - One message for each of the other processes indicating the process with the second highest identifier is the new coordinator.
- Worst case
  - The node with lowest identifier detects failure. This causes N-1 processes to initiate the election algorithm each sending messages to processes with higher identifiers.
  - Total messages = $O(N^2)$

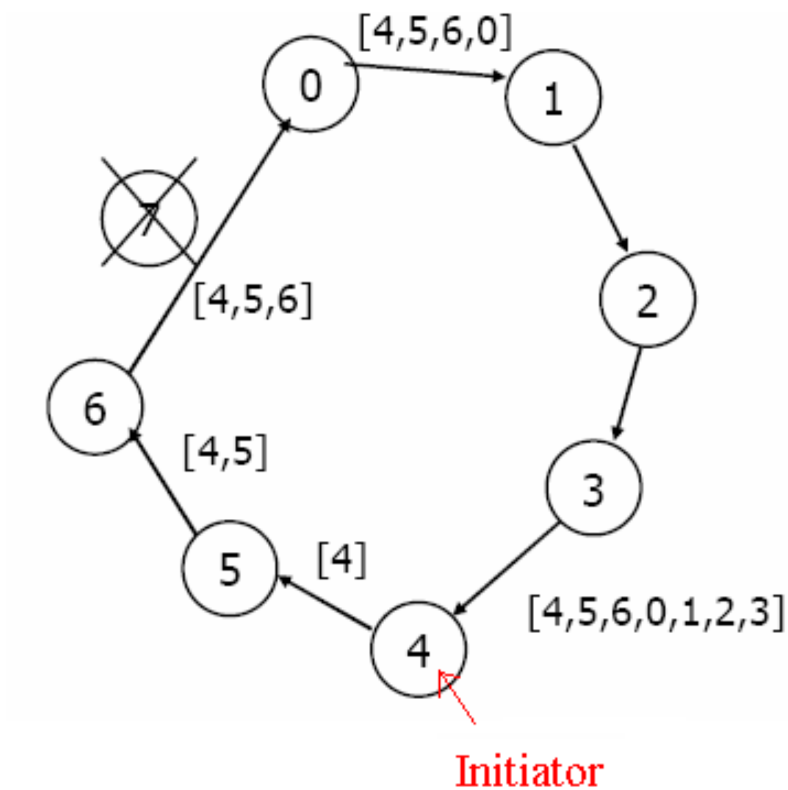
2. Ring Algorithm

- Applicable to systems organized as a ring (logically or physically)
- Assumes that the links are unidirectional, and that processes send their messages to their right neighbors
- Each process maintains an active list, consisting of all the priority numbers of all active processes in the system when the algorithm ends
- If process Pi detects a coordinator failure, it creates a new active list that is initially empty. It then sends a message elect(i) to its right neighbor, and adds the number i to its active list

Ring Algorithm (Cont.)

- If P_i receives a message elect(j) from the process on the left, it must respond in one of three ways:
  - ✦ If this is the first *elect* message it has seen or sent, P_i creates a new active list with the numbers i and j
    - It then sends the message *elect(i)*, followed by the message *elect(j)*
  - ✦ If $i \neq j$, the message dos not contain P_i
    - P_i adds j to its active list and forward message to the right
  - ✦ If $i = j$, then P_i receives the message *elect(i)*
    - The active list for P_i contains all the active processes in the system
    - P_i can now determine the largest number in the active list to identify the new coordinator process

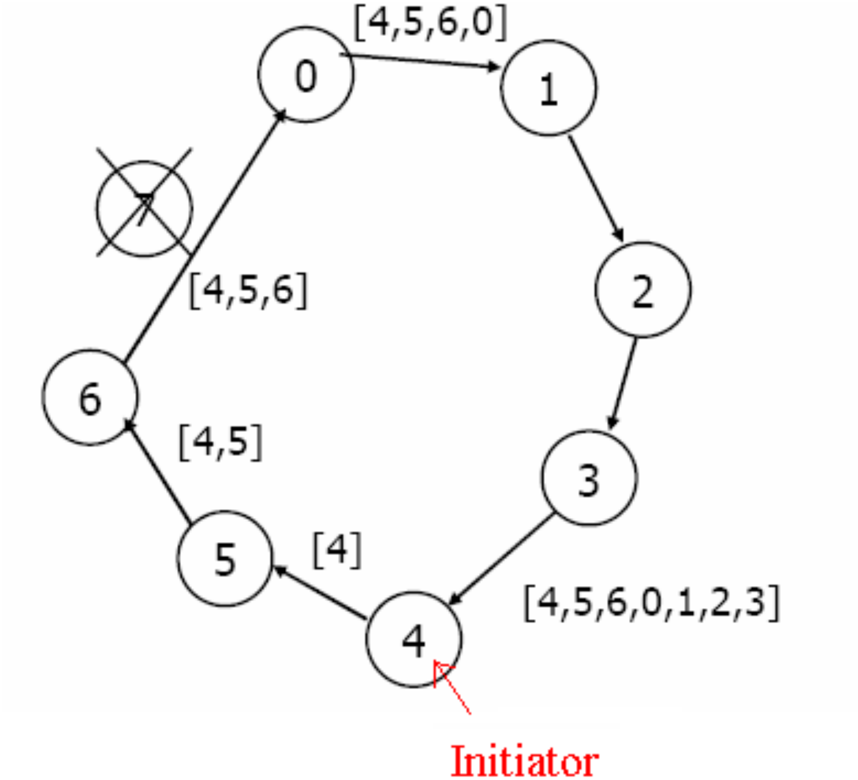
Ring Algorithm (Example)



Initiation:

1. Process 4 sends an ELECTION message to its successor (or next alive process) with its ID

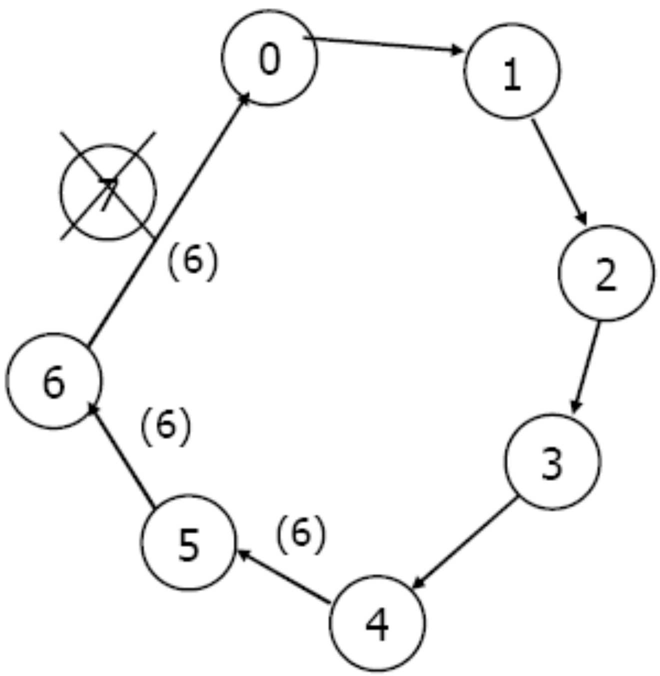
Ring Algorithm (Example)



Initiation:

2. Each process adds its own ID and forwards the ELECTION message

Ring Algorithm (Example)



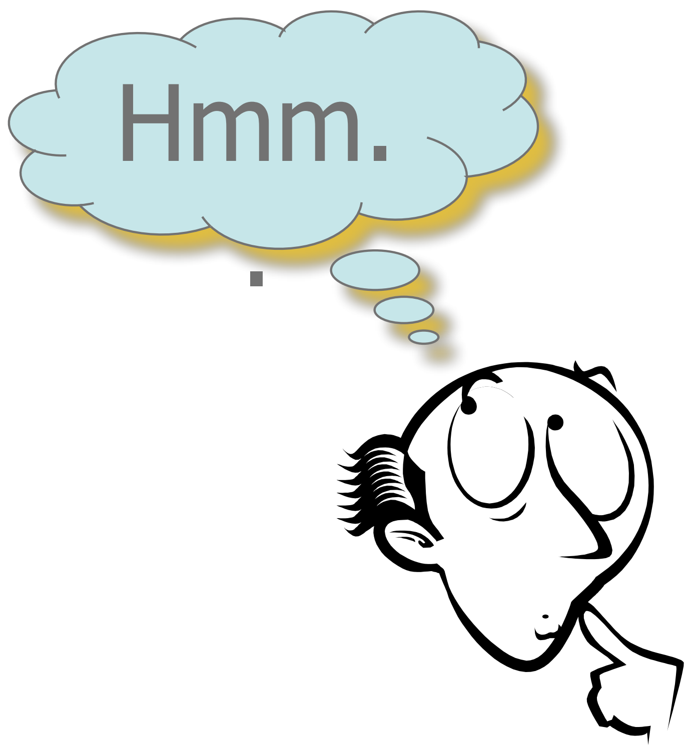
Leader Election:

3. Message comes back to initiator, here the initiator is 4.
4. Initiator announces the winner by sending another message around the ring

Ring Algorithm Analysis

- At best 2(N-1) messages are passed
  - One round for the ELECTION message
  - One round for the COORDINATOR
  - Assumes that only a single process starts an election.
- Multiple elections cause an increase in messages but no real harm done.

Any Questions?



Acknowledgements

- "Operating Systems Concepts" book and supplementary material by A. Silberschatz, P. Galvin and G. Gagne

- "Operating Systems: Internals and Design Principles" book and supplementary material by W. Stallings

- "Modern Operating Systems" book and supplementary material by A. Tanenbaum

- R. Doursat and M. Yuksel from UNR; E. Lazowska from UW