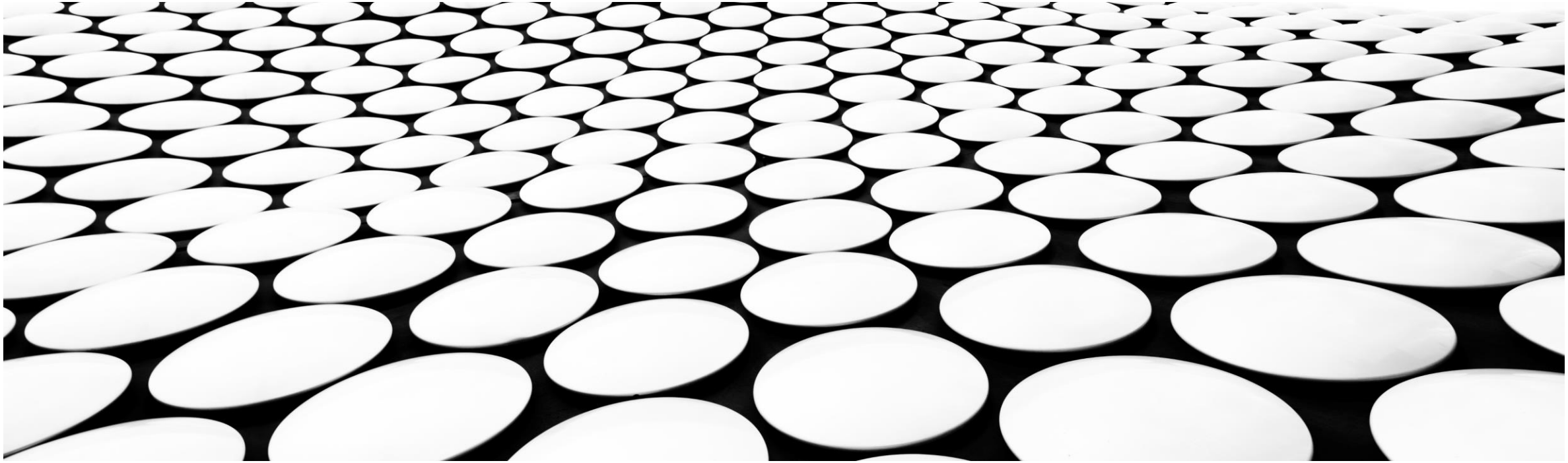
Cooperation and coordination in distributed systems

COEN-317: Distributed Systems
Robert Bruce
Department of Computer Science and Engineering
Santa Clara University



Synchronization versus coordination

Process synchronization: "...one process waits for another to complete its operation" [1].

Data synchronization: "...ensure that two sets of data are the same" [1].

Coordination: "...manage the interactions and dependencies between activities in a distributed system" [1].

[1] p. 297, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Network Time Protocol (NTP)

Network Time Protocol (NTP) is used for synchronizing the clocks in a network [1].

NTP computes offset and delay between an NTP client and NTP server to correct for network latency errors.

[1] p. 297, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Factoring network latency into the Network Time Protocol

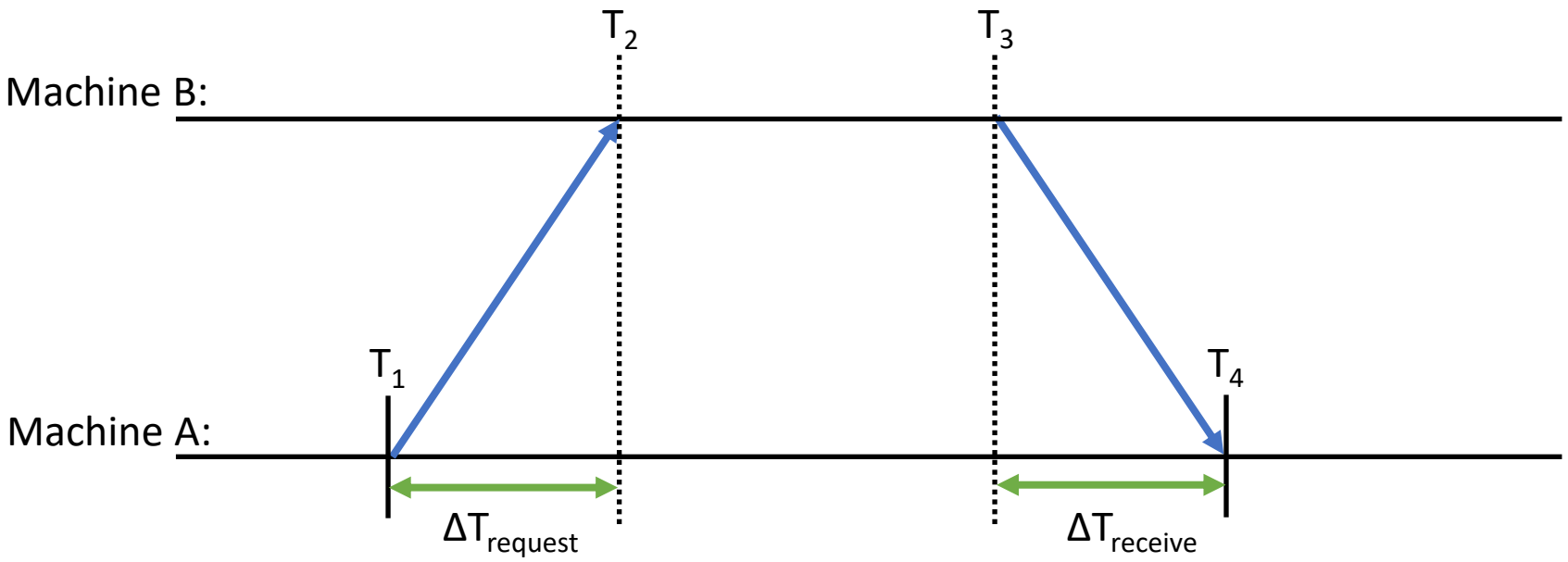


Figure source: p. 305, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Factoring network latency into the Network Time Protocol

$$\theta = \frac{(T_2 - T_1) + (T_3 - T_4)}{2}$$

Where theta (θ) denotes offset in machine A relative to machine B.

$$\Delta = \frac{(T_4 - T_1) - (T_3 - T_2)}{2}$$

Where delta (Δ) denotes round-trip delay from machine A relative to machine B.

Figure source: p. 305, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Coordination through election algorithms

Leadership election algorithms:
- Provide a means for choosing a coordinator in distributed systems [1].
- Provide resiliency in distributed systems when the existing coordinator fails (discussed in future lecture)

Bully leadership election algorithm:
- Processes are hard-coded and identified by a non-negative integer [2].
- This non-negative integer is not unique.
- The highest numbered process is elected as coordinator [2].

Ring leadership election algorithm:
- Processes are connected in a circular (ring) network [3].
- Candidate processes elect themselves to neighbor processes in a circular propagation pattern [3].
- A candidate process in the ring network becomes coordinator when it receives its own election message (i.e. one revolution of the election process) [3].

Wireless network leadership election algorithm:
- Difficult to achieve in an ad-hoc network due to topology: the network is dynamic (it can change at any time) [4].
- Focus on electing the best leader rather than a random leader [4].

[1] p. 329, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.
[2] p. 330, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.
[3] p. 332, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.
[4] p. 333, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Bully leadership election algorithm

Bully leadership election algorithm:

1. "P_k sends an ELECTION message to all processes with higher identifiers: P_{k+1}, P_{k+2}, ... P_{N-1}" [1].
2. "If no one responds, P_k wins the election and becomes coordinator" [1].
3. "If one of the higher-ups answers, it takes over and P_k's job is done" [1].

[1] p. 329, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.
[2] p. 330, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.
[3] p. 332, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.
[4] p. 333, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Bully leadership election algorithm

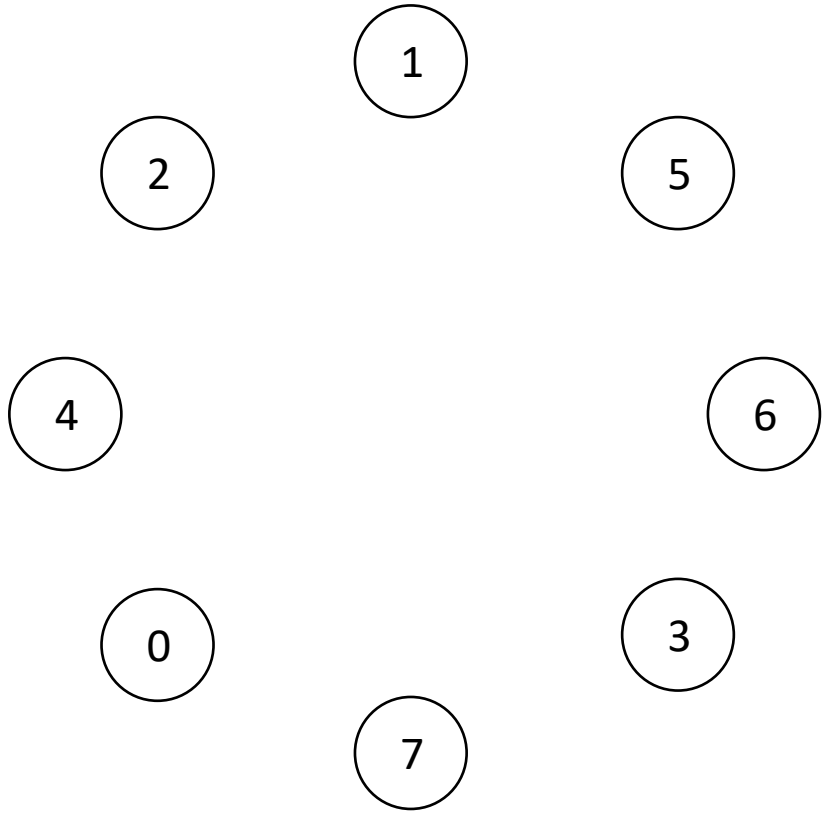


Figure source: p. 331, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Bully leadership election algorithm

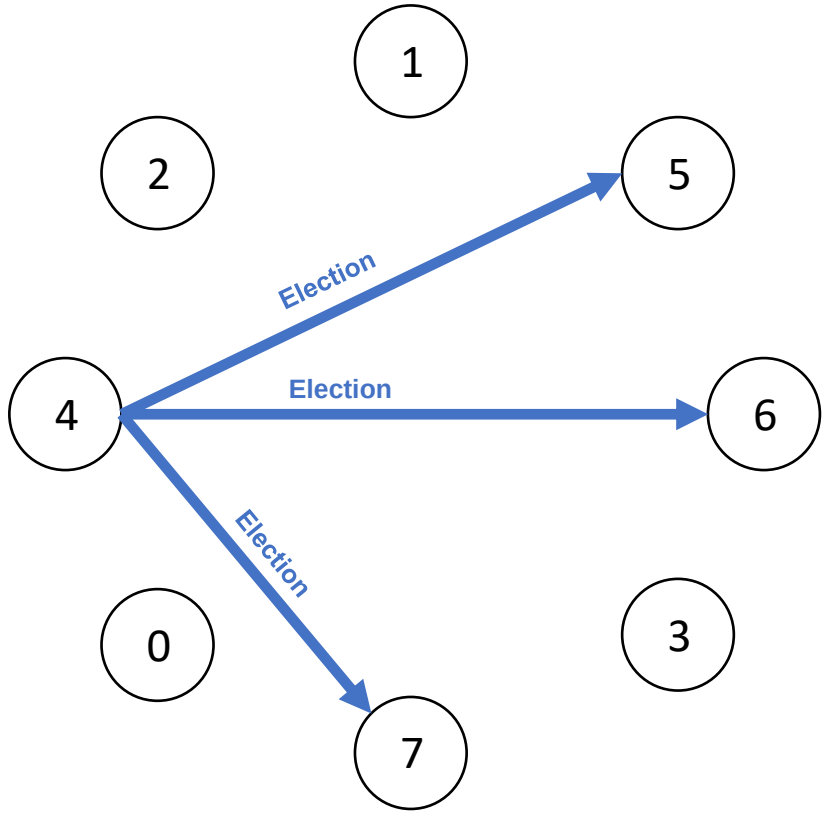


Figure source: p. 331, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Bully leadership election algorithm

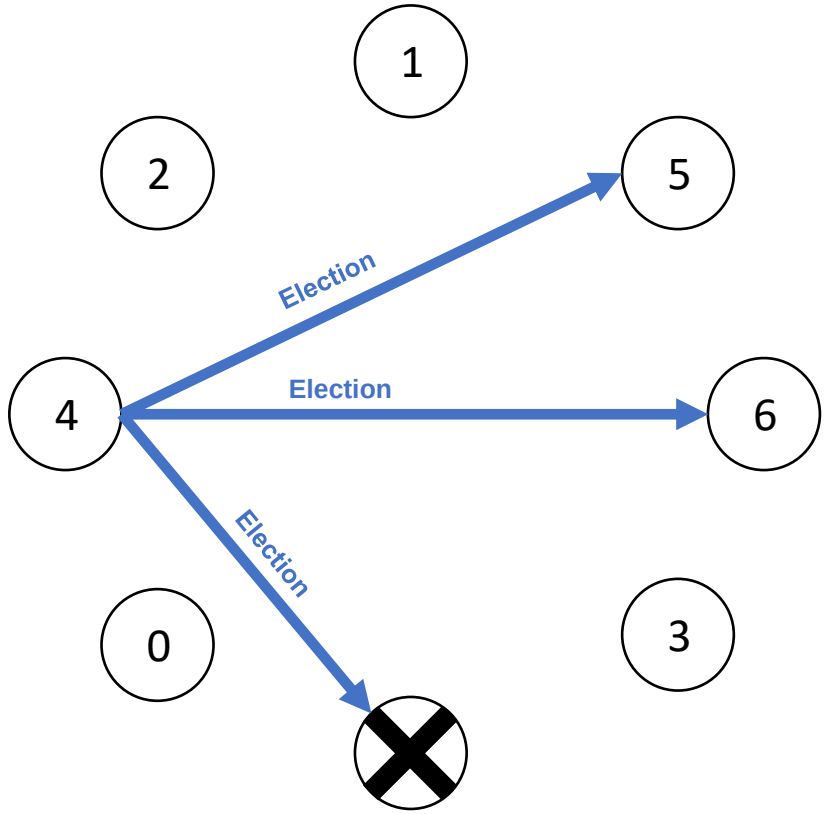
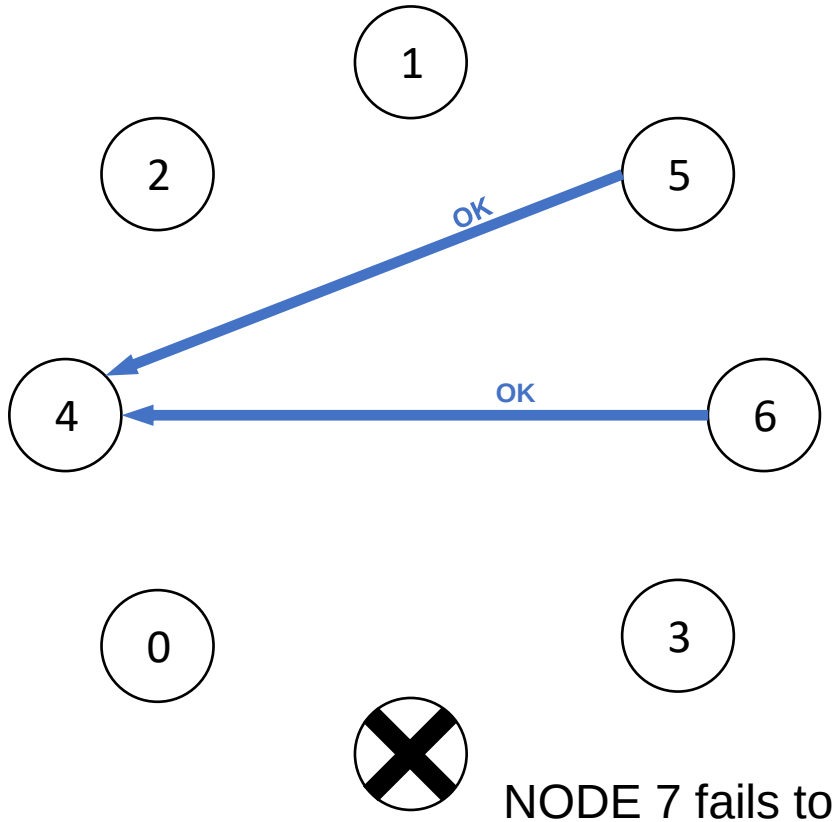


Figure source: p. 331, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Bully leadership election algorithm



NODE 7 fails to respond to NODE 4.

Figure source: p. 331, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Bully leadership election algorithm

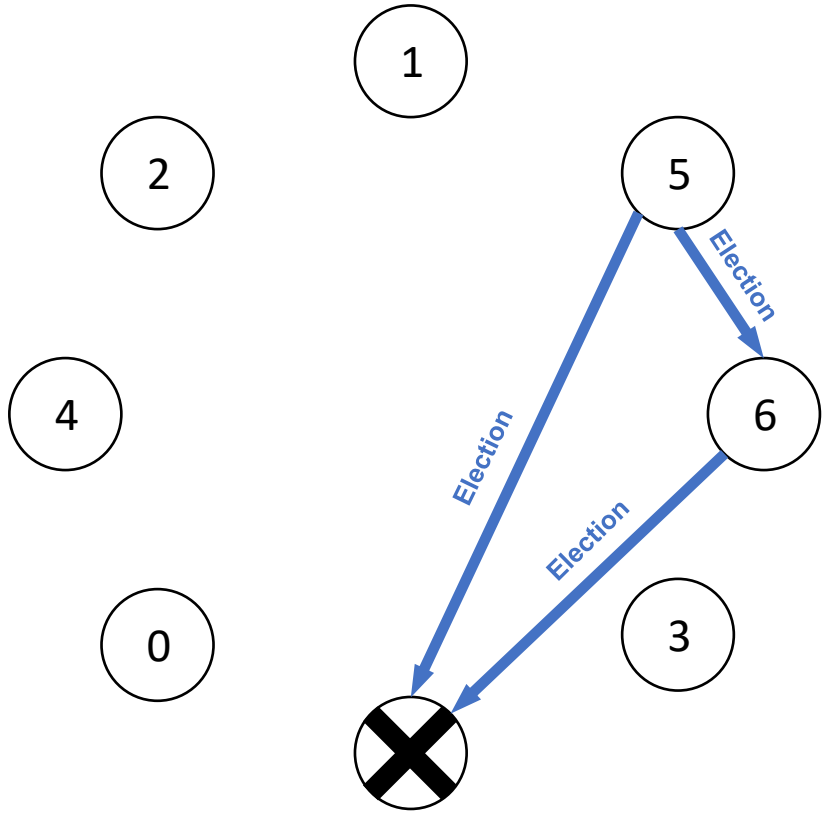


Figure source: p. 331, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Bully leadership election algorithm

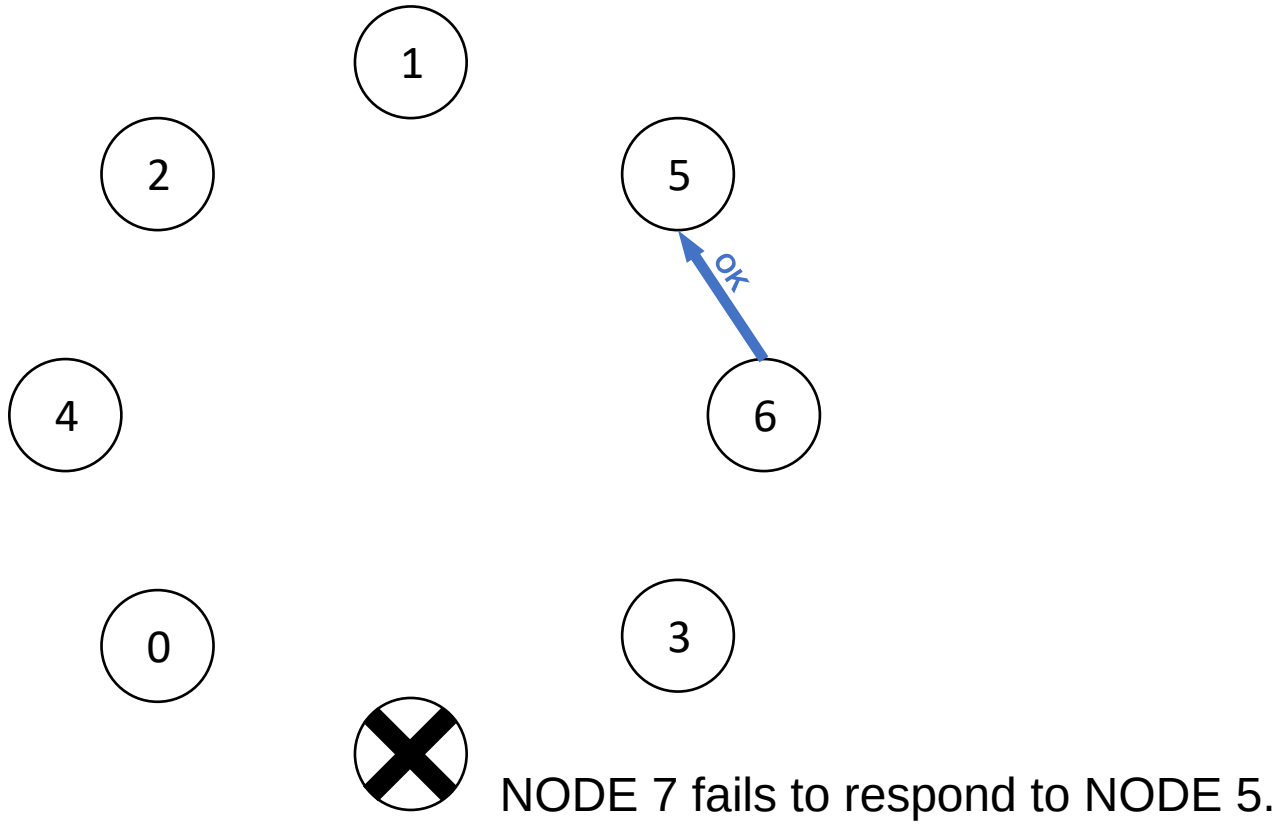


Figure source: p. 331, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Bully leadership election algorithm

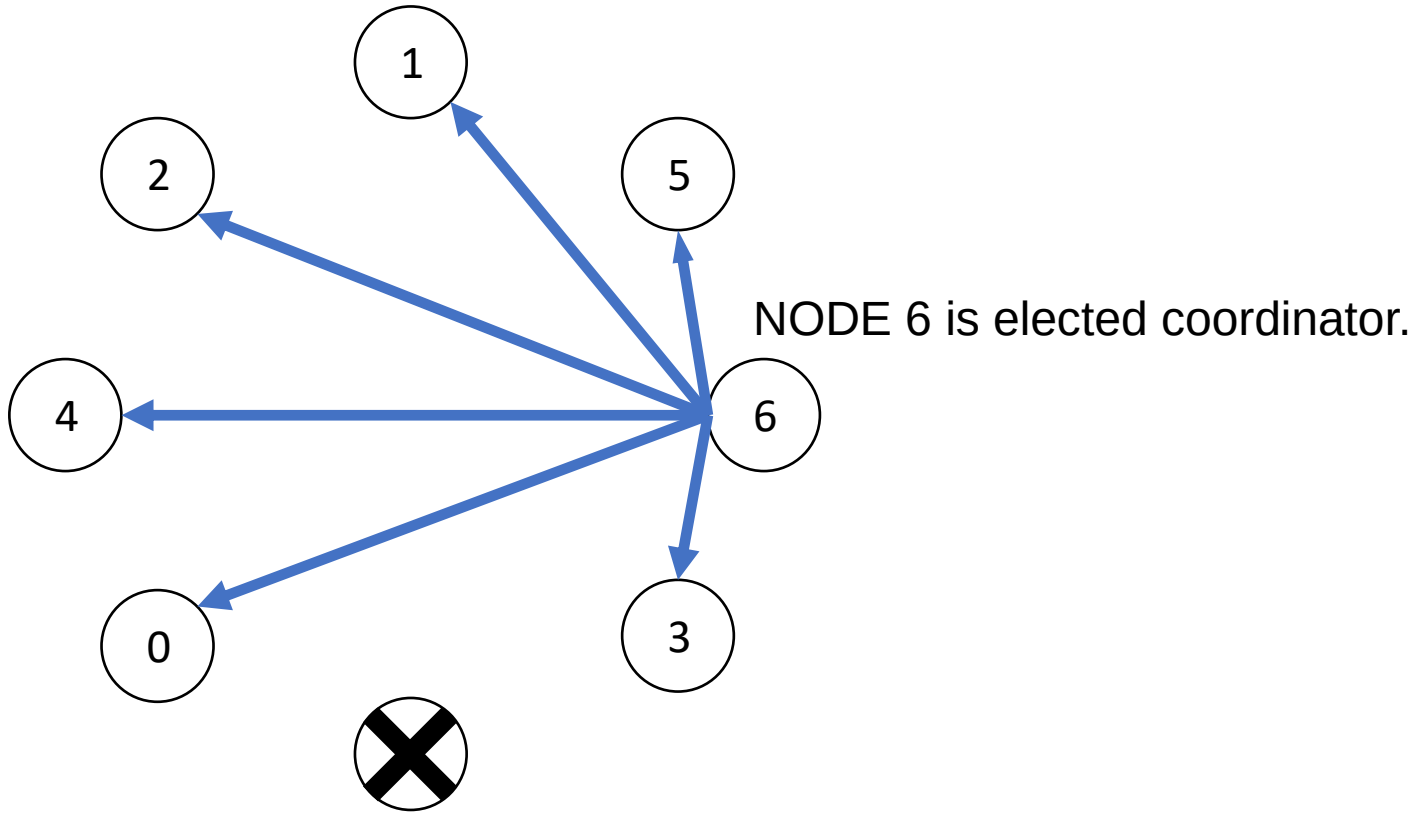


Figure source: p. 331, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Ring leadership election algorithm

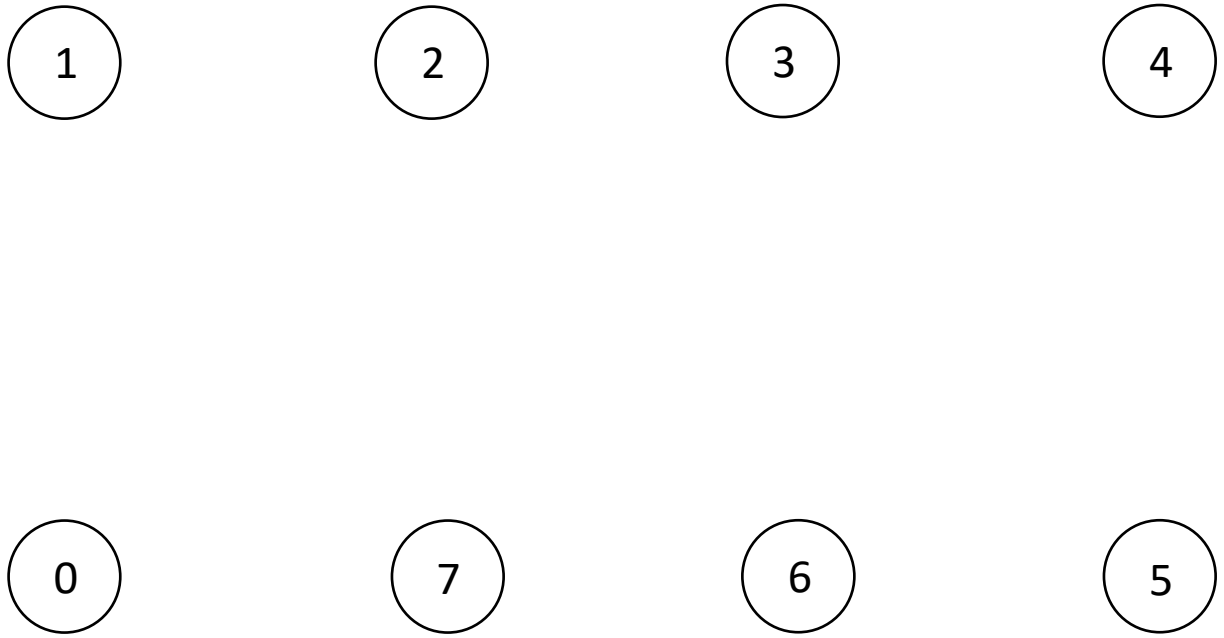
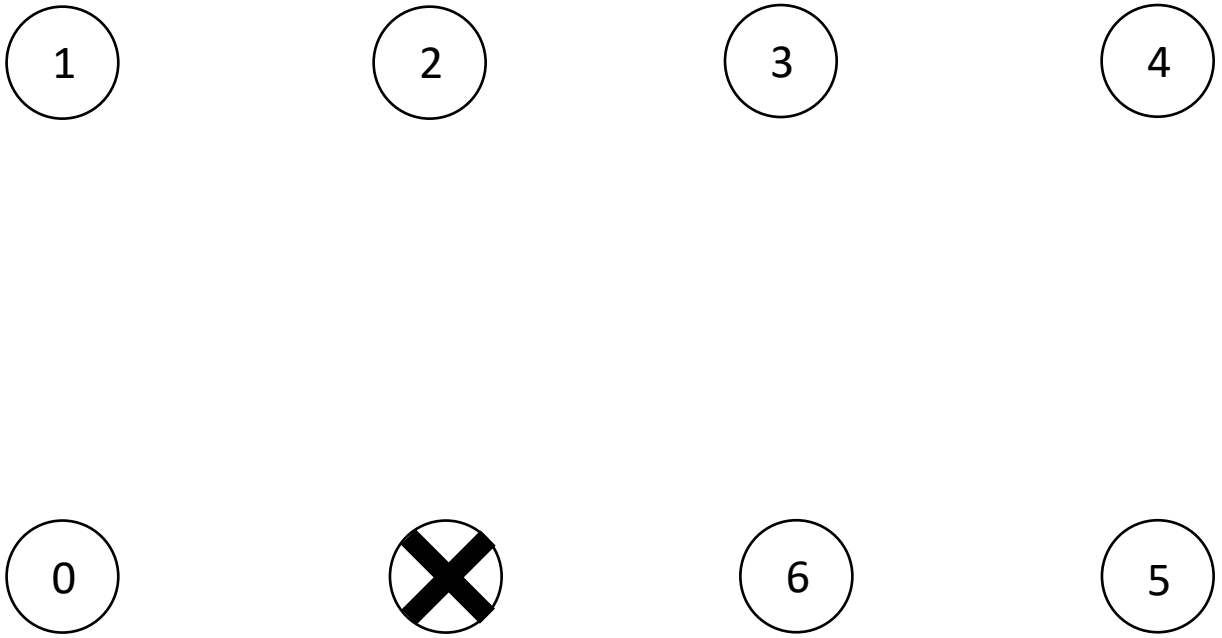


Figure source: p. 332, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Ring leadership election algorithm



NODE 7 is inaccessible (offline).

Figure source: p. 332, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Ring leadership election algorithm

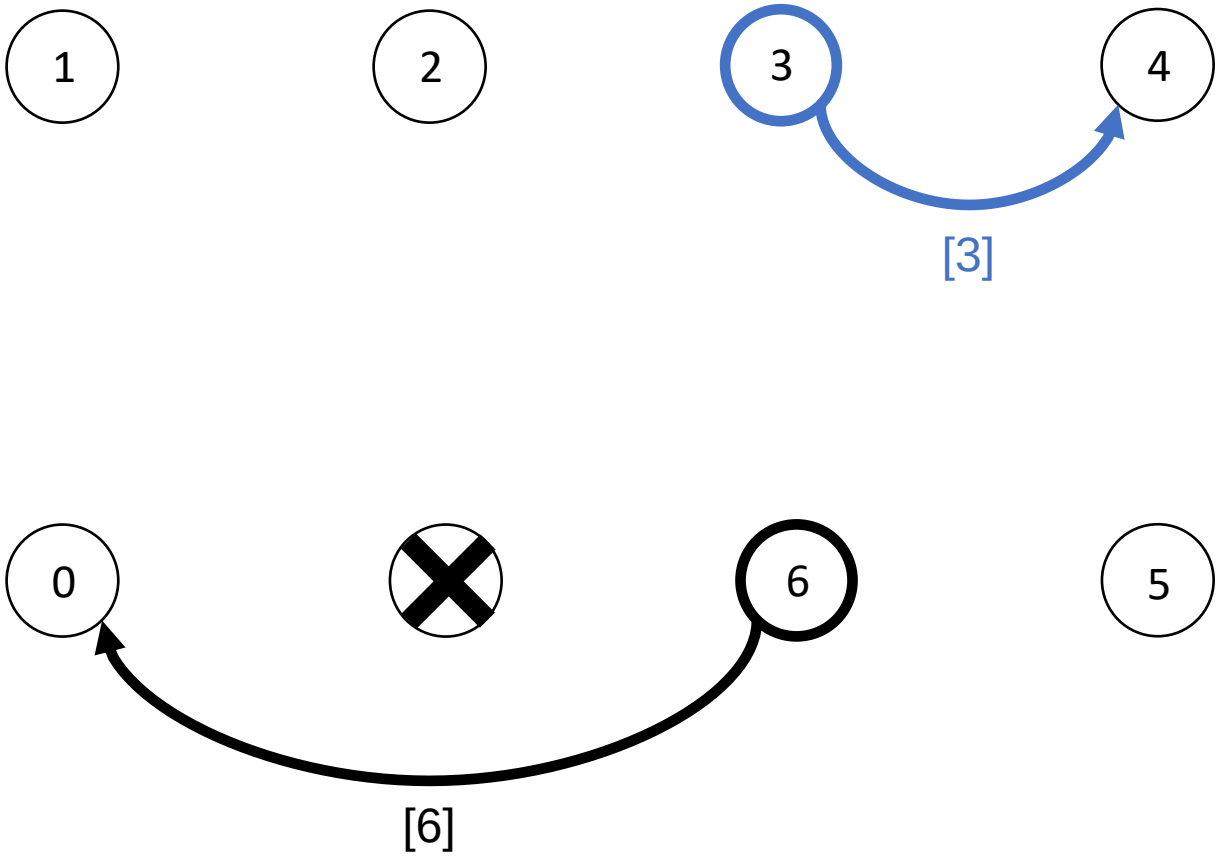


Figure source: p. 332, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Ring leadership election algorithm

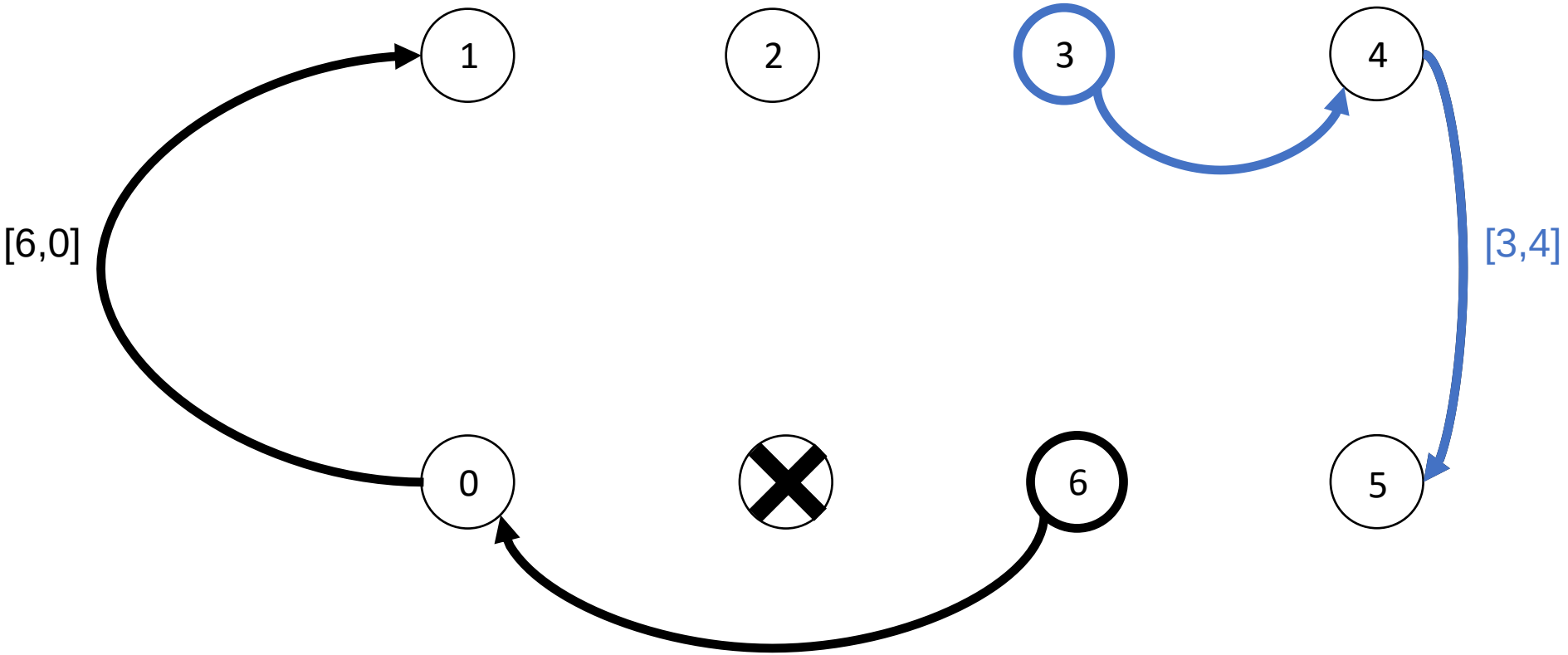


Figure source: p. 332, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Ring leadership election algorithm

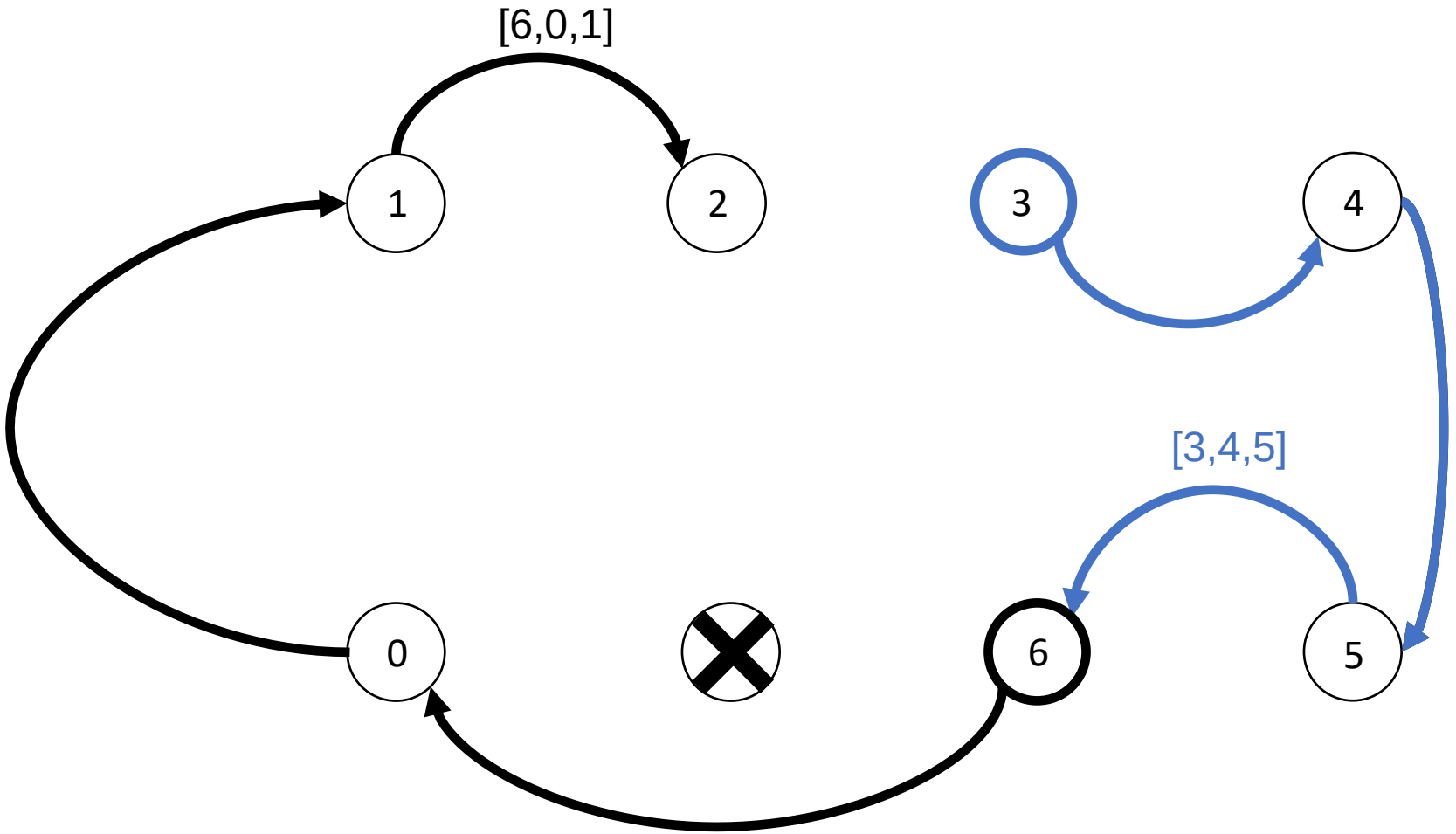


Figure source: p. 332, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Ring leadership election algorithm

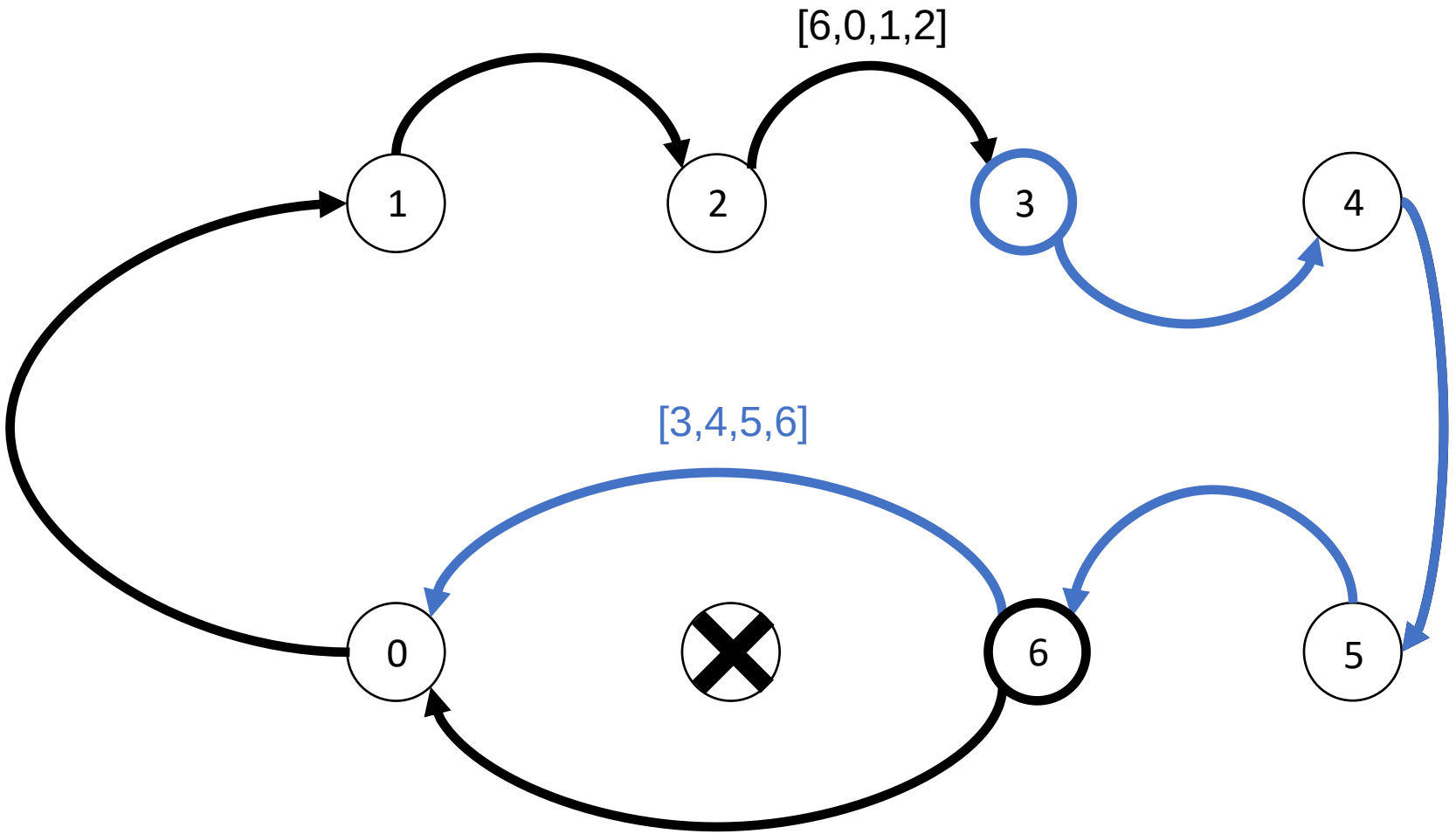


Figure source: p. 332, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Ring leadership election algorithm

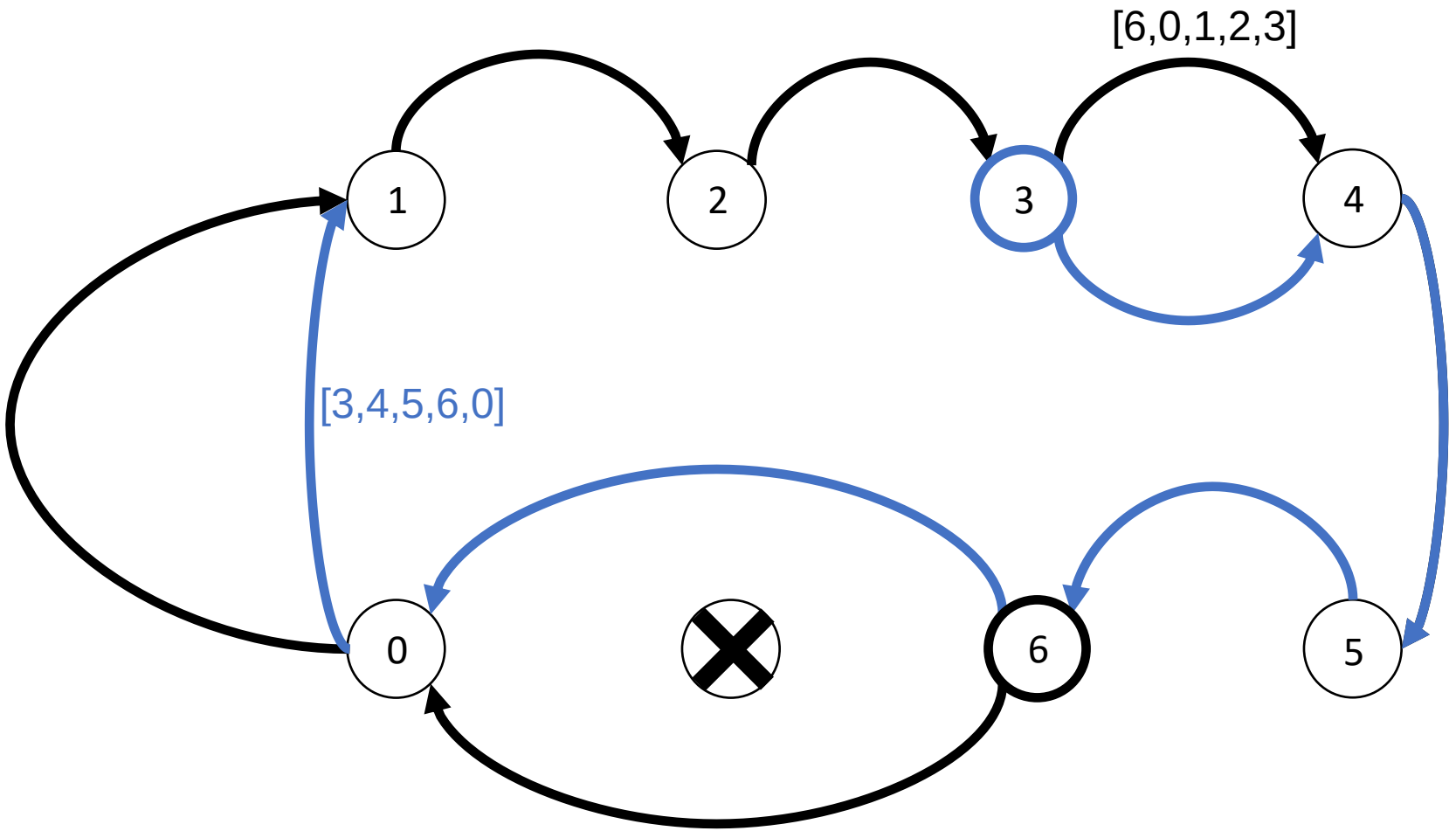


Figure source: p. 332, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Ring leadership election algorithm

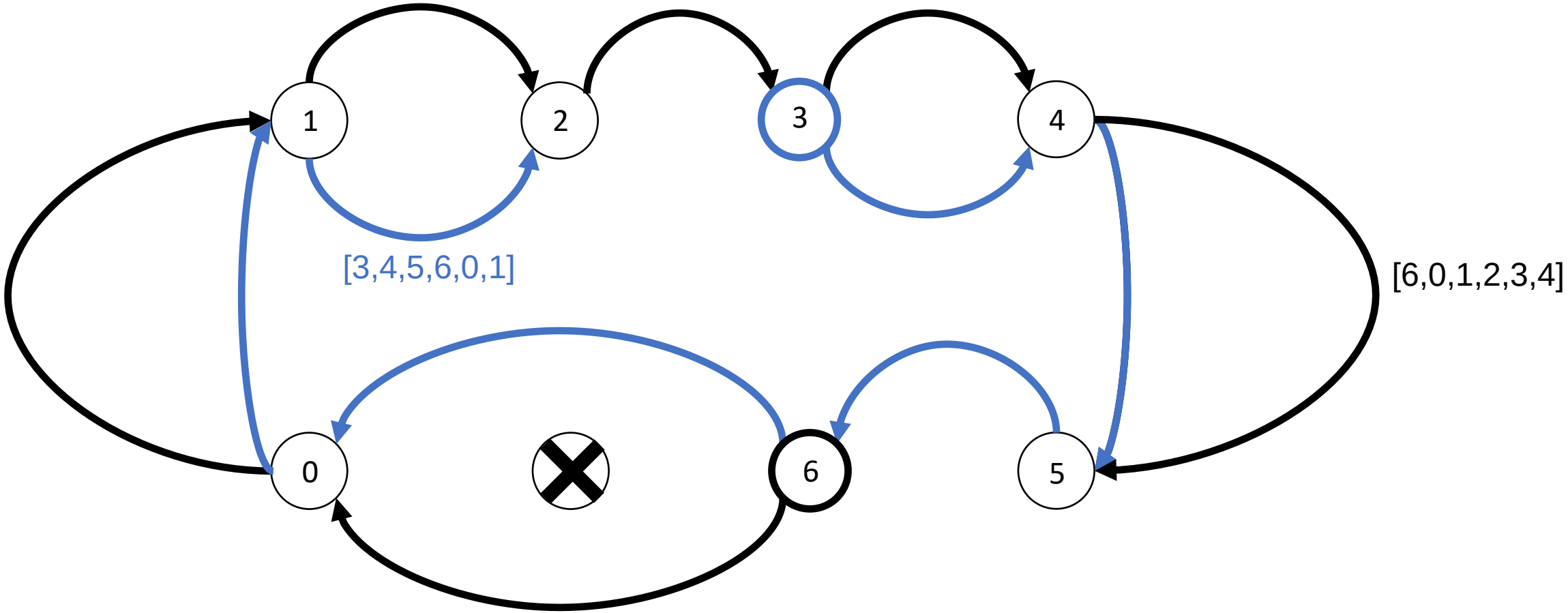


Figure source: p. 332, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Ring leadership election algorithm

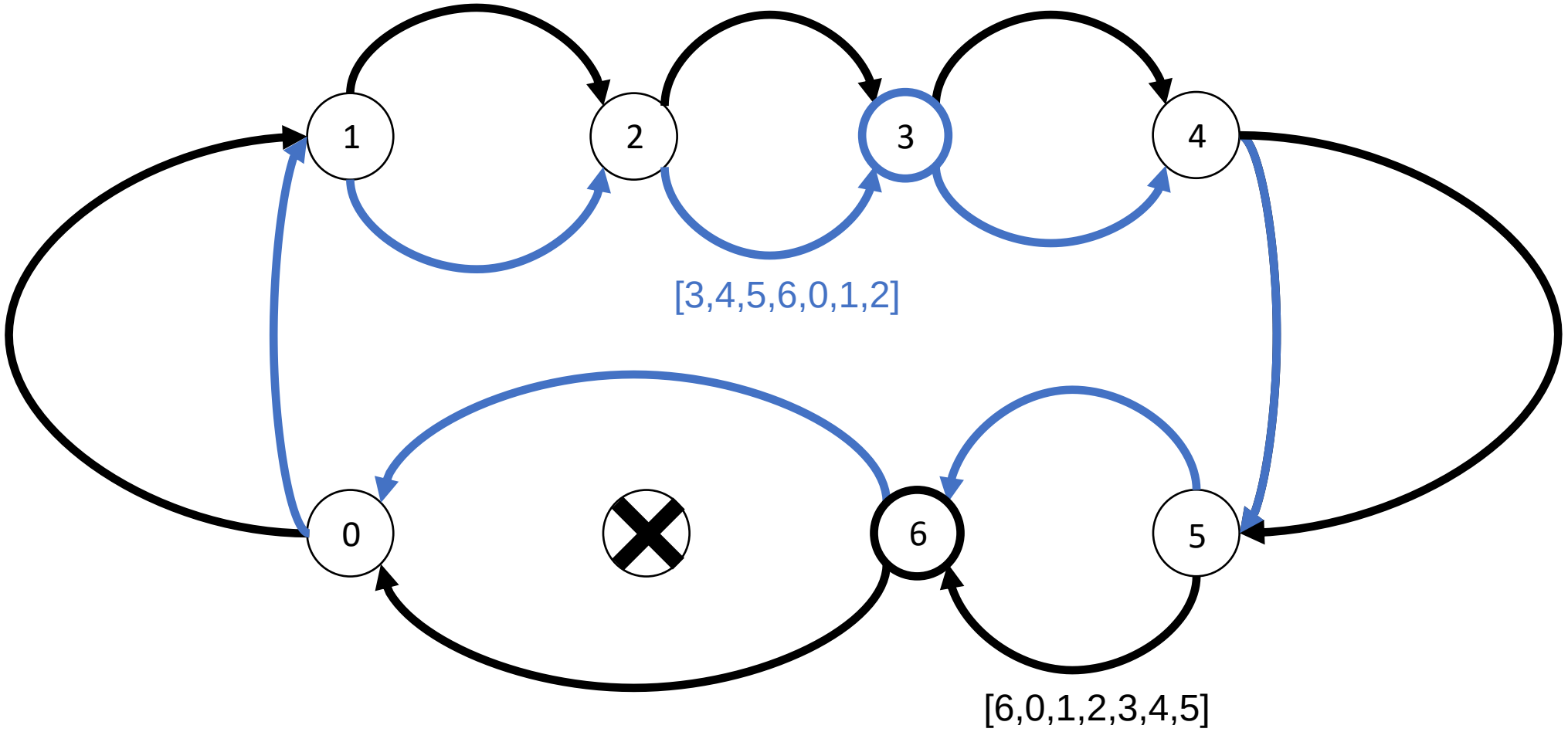


Figure source: p. 332, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Ring leadership election algorithm

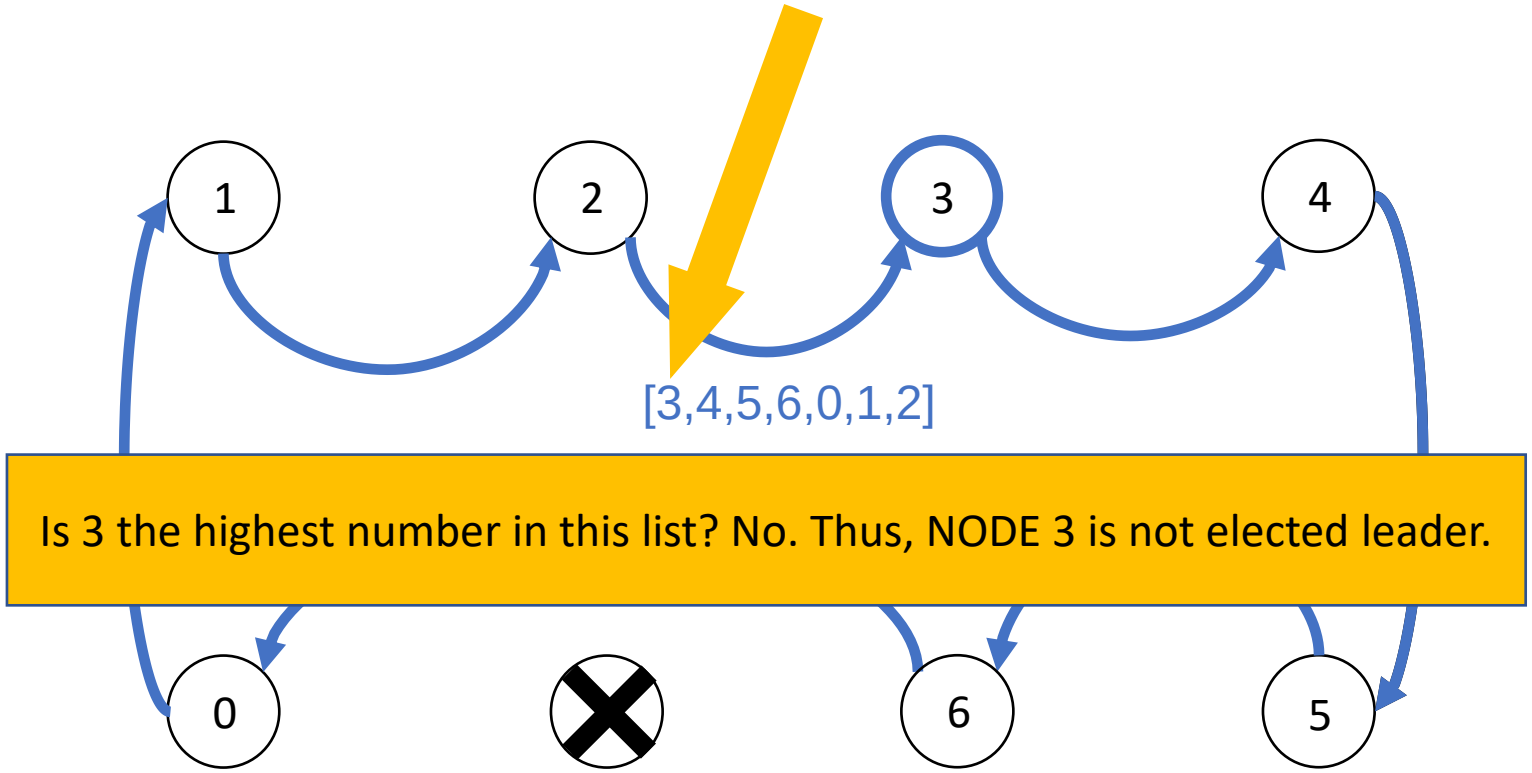


Figure source: p. 332, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Ring leadership election algorithm

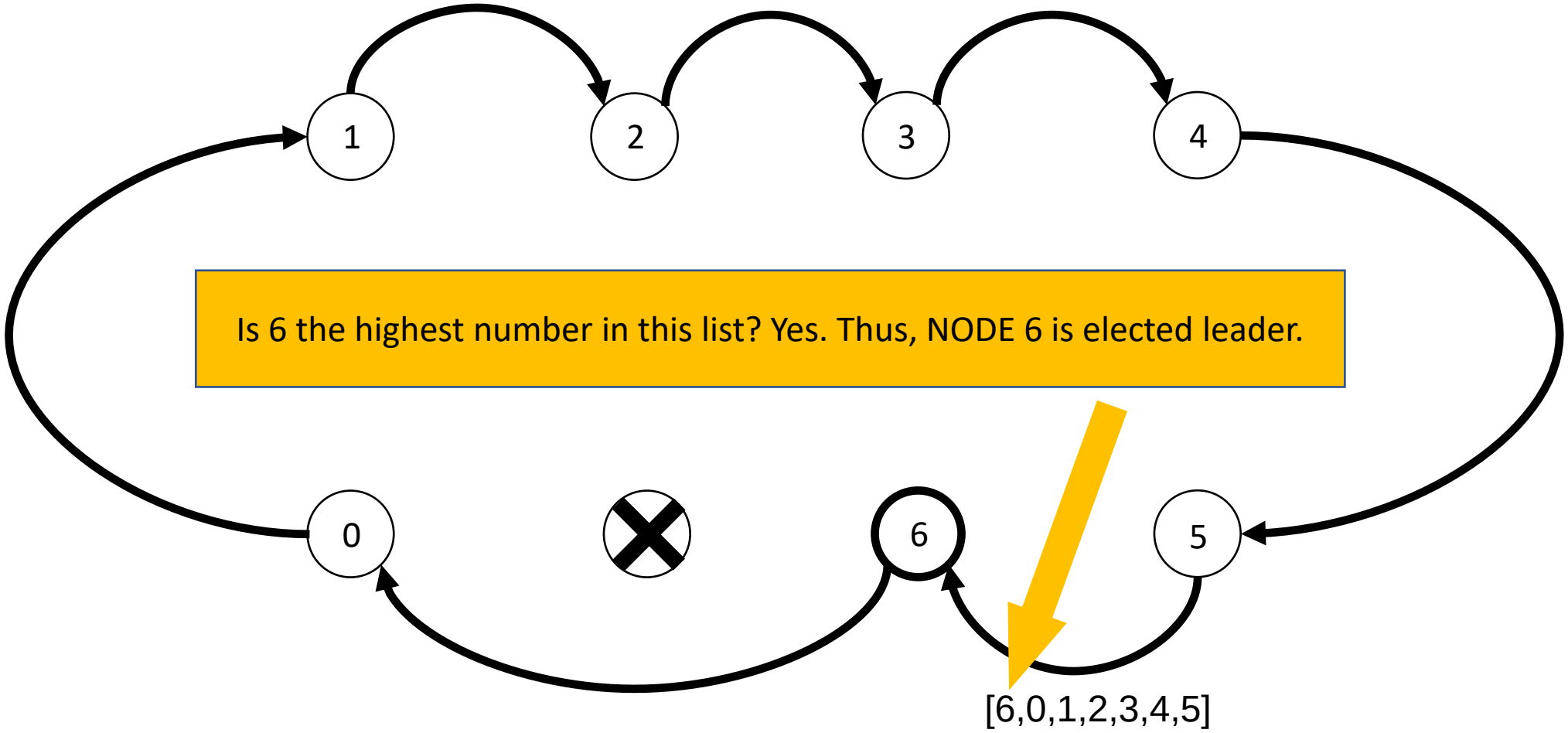


Figure source: p. 332, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Ring leadership election algorithm

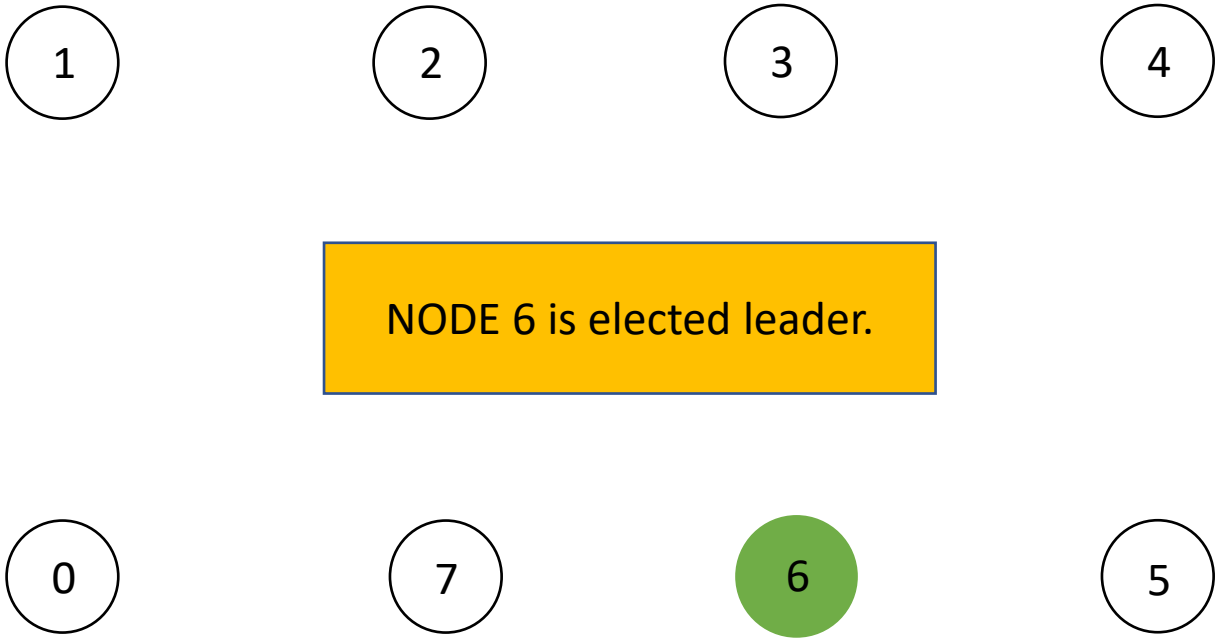


Figure source: p. 332, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Leadership election algorithm in a wireless network

Procedure:

1. Sender node propagates election message to neighbor nodes (recipient) [1].
2. If the recipient node has not received an election message already, the recipient denotes the sender node as parent. The recipient node then sends the election message to all its immediate neighbor nodes except for the parent node [1].
3. If the recipient node has already received an election message, it merely acknowledges the message but does not make the sender node a parent [1].
4. If the recipient node receives two or more election messages concurrently, the recipient node chooses one sender as the parent node. The process for choosing is decided by an ID that is also propagated in the voting messages [2].

[1] p. 333, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.
[2] p. 335, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Leadership election algorithm in a wireless network

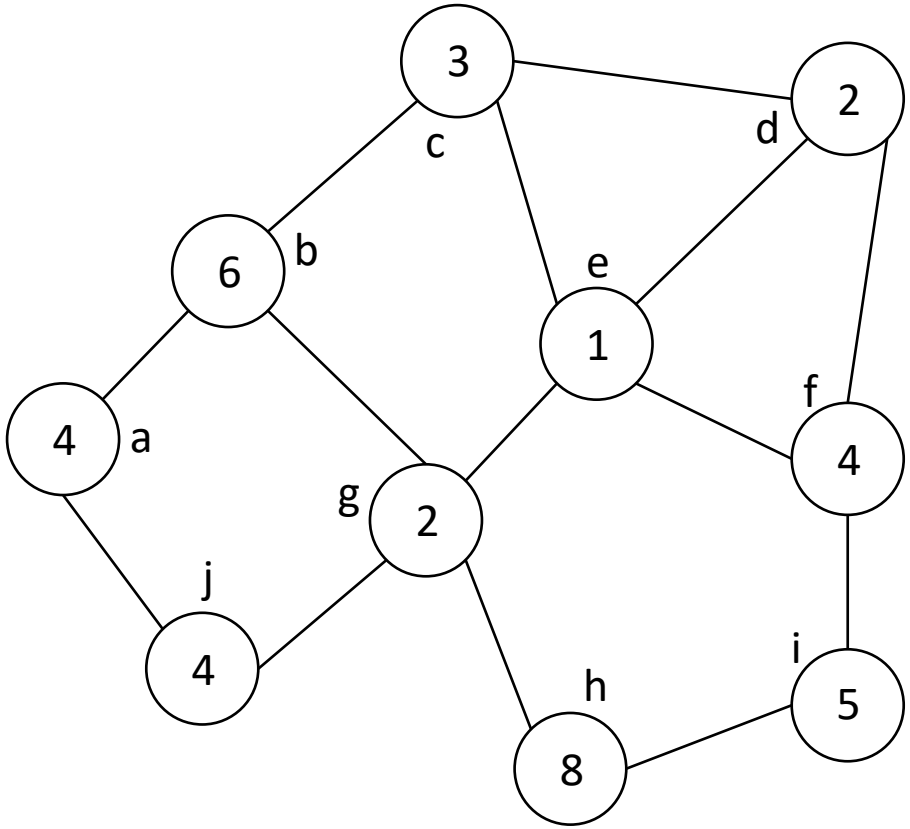


Figure source: p. 334, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Leadership election algorithm in a wireless network

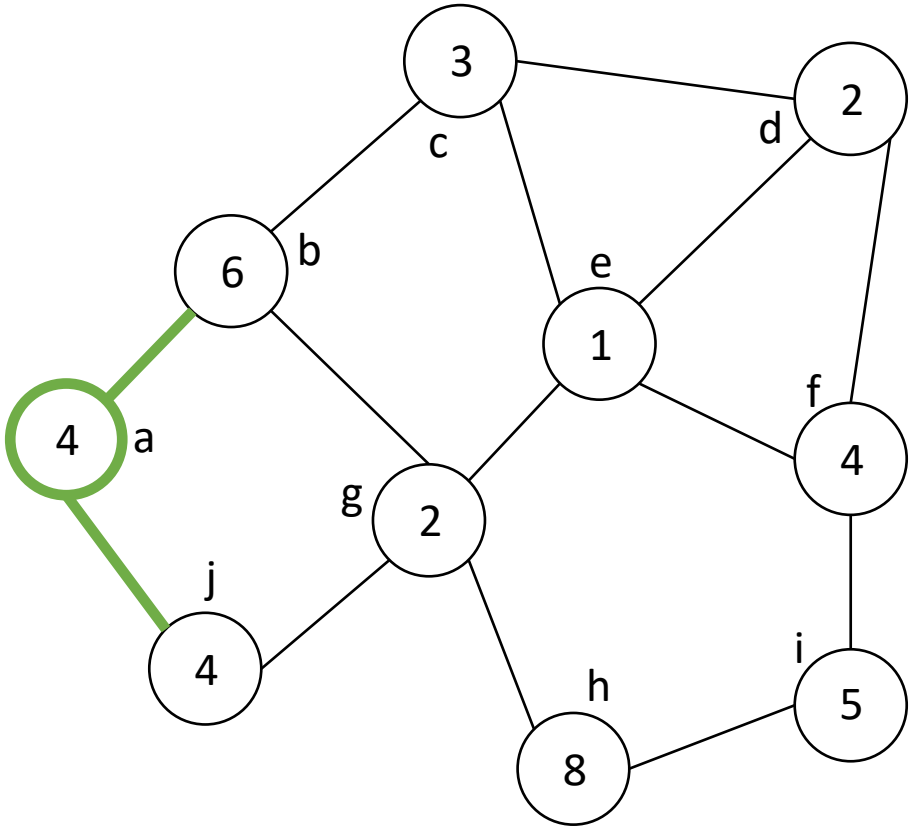


Figure source: p. 334, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Leadership election algorithm in a wireless network

NODE "b" versus NODE "j":

NODE "g" received election message from NODE "b" first.

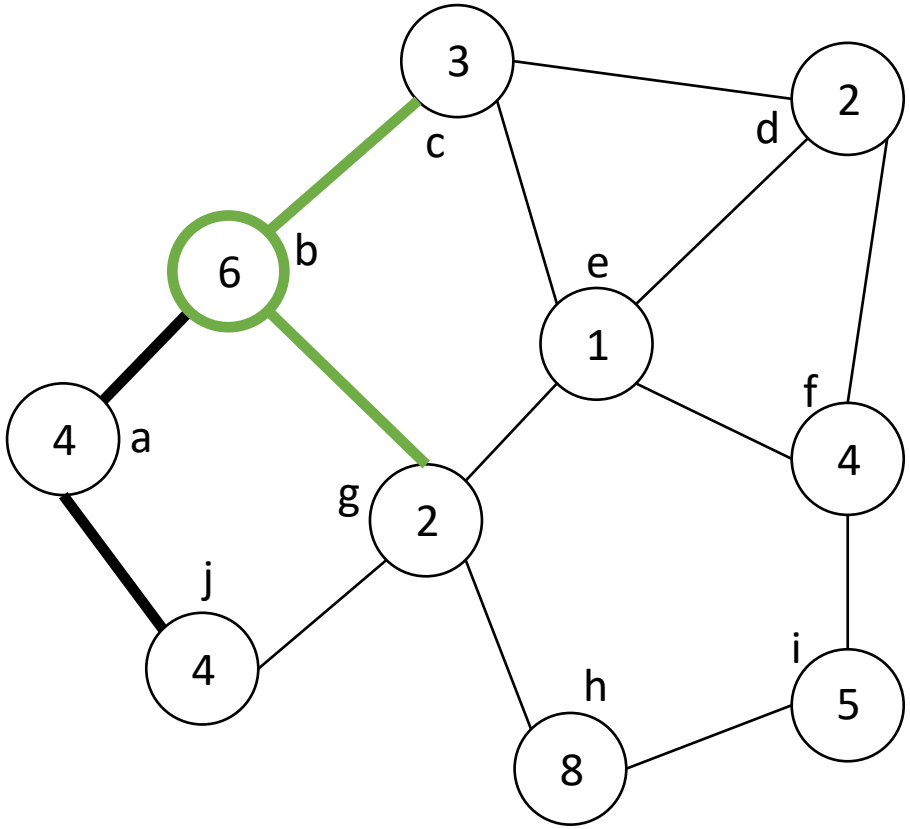


Figure source: p. 334, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Leadership election algorithm in a wireless network

NODE "c" versus NODE "g":

NODE "e" received election message from NODE "g" first.

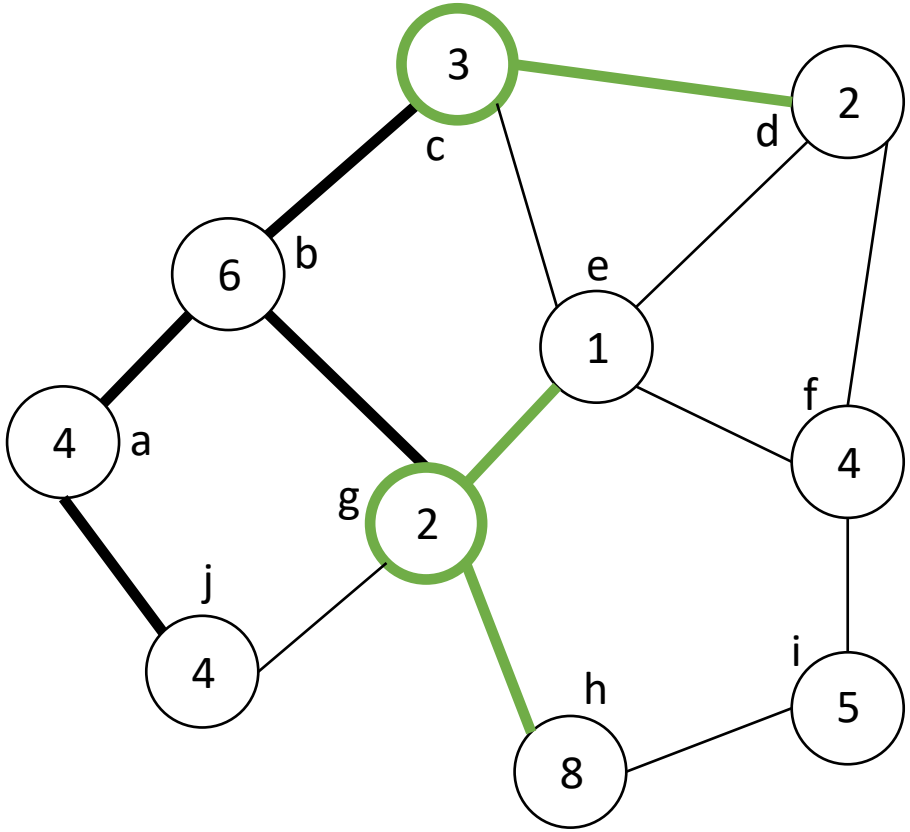


Figure source: p. 334, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Leadership election algorithm in a wireless network

NODE "e" versus NODE "d":

NODE "f" received election message from NODE "e" first.

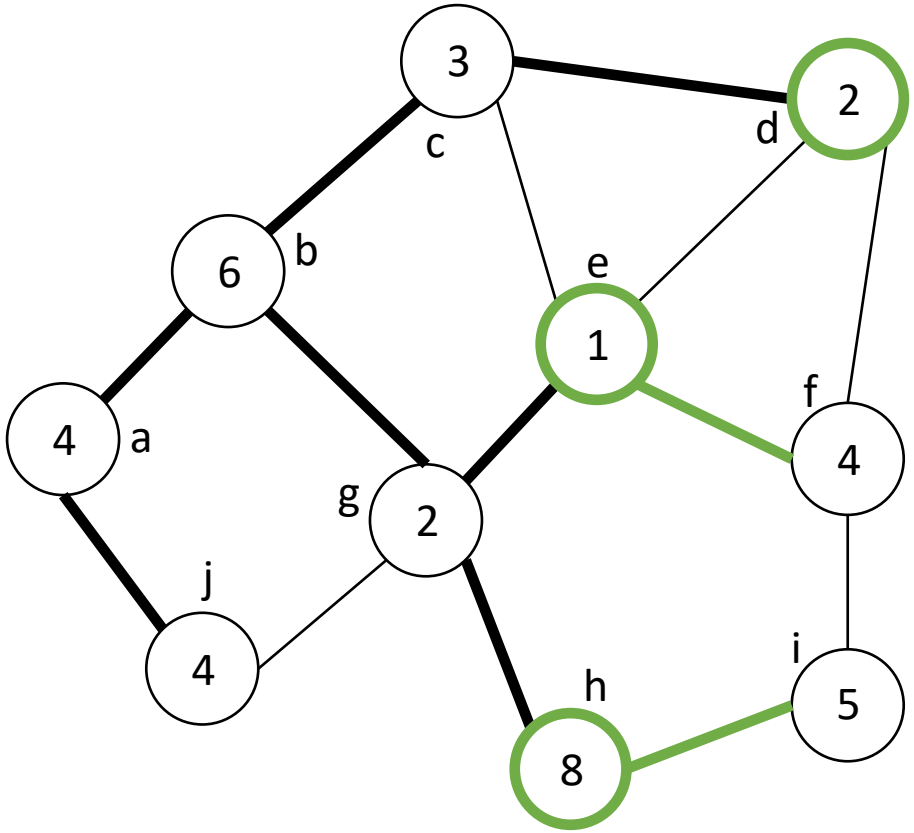


Figure source: p. 334, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Leadership election algorithm in a wireless network

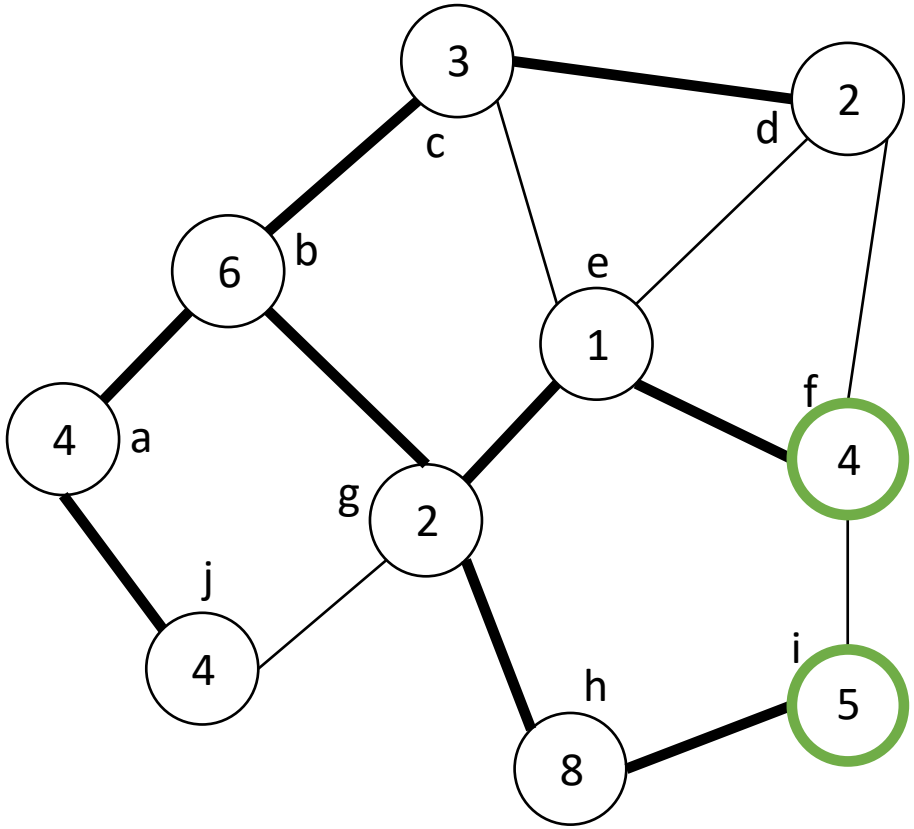


Figure source: p. 334, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Leadership election algorithm in a wireless network

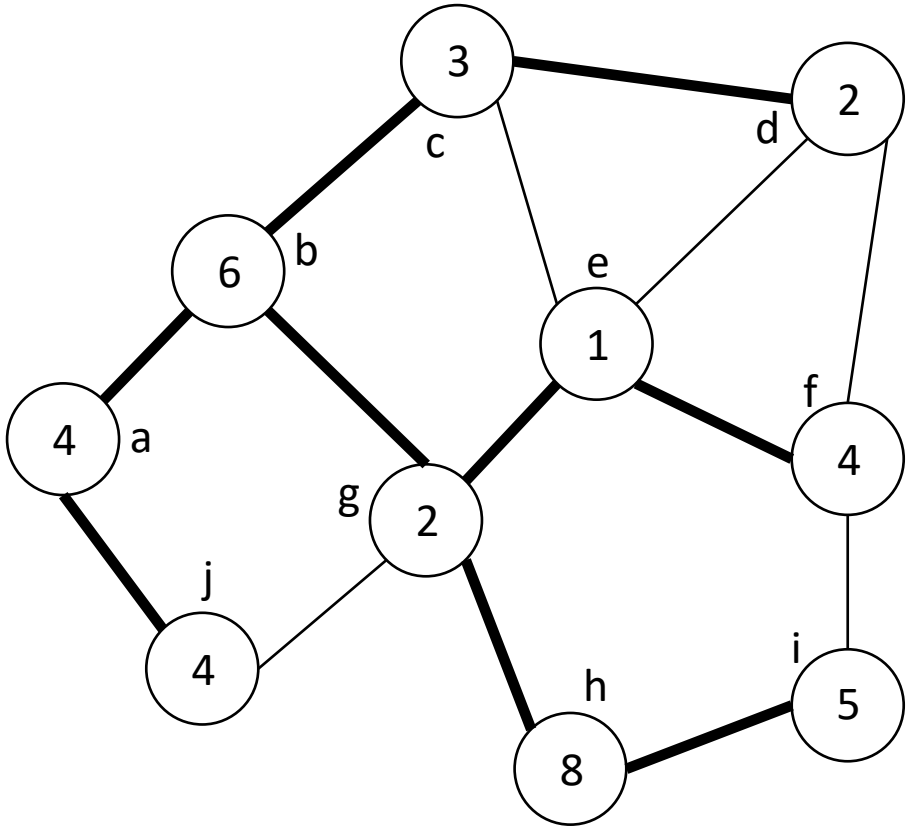


Figure source: p. 334, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

Leadership election algorithm in a wireless network

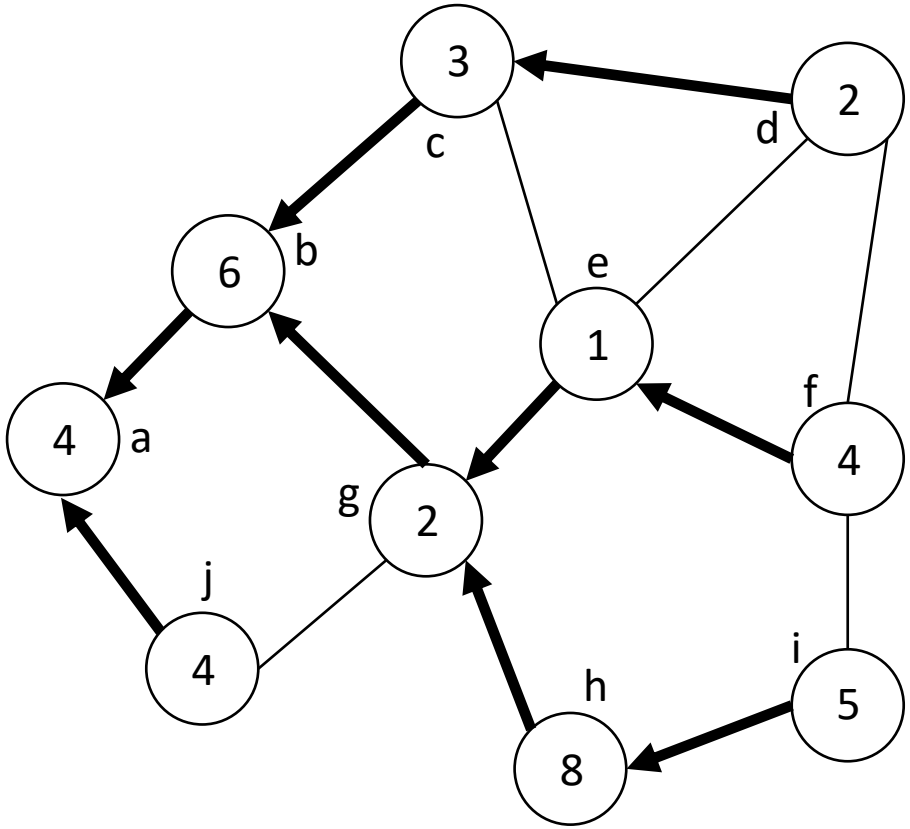


Figure source: p. 334, *Distributed Systems* (3rd edition) by Maarten van Steen and Andrew S. Tanenbaum.

For further reading

Official NTP Documentation:
https://www.ntp.org/documentation.html

Research paper on the ring election algorithm:
E. Chang and R. Roberts, "An improved algorithm for decentralized extrema-finding in circular configurations of processes", *Communications of the ACM*, vol. 22, no. 5, May 1979, pp. 281-283. DOI:10.1145/359104.359108