Chapter 3
Leader Election

Distributed Systems

SS 2019

Fabian Kuhn

Leader Election

General goal: Elect some node as a leader

Leader Election Problem:

Each node eventually whether it is a leader or not, subject to the constraint that there is exactly one leader

- *implicit leader election:* the non-leader do not need to know the name of the leader (a.k.a. test-and-set)
- *explicit leader election:* each node knows the name of the leader

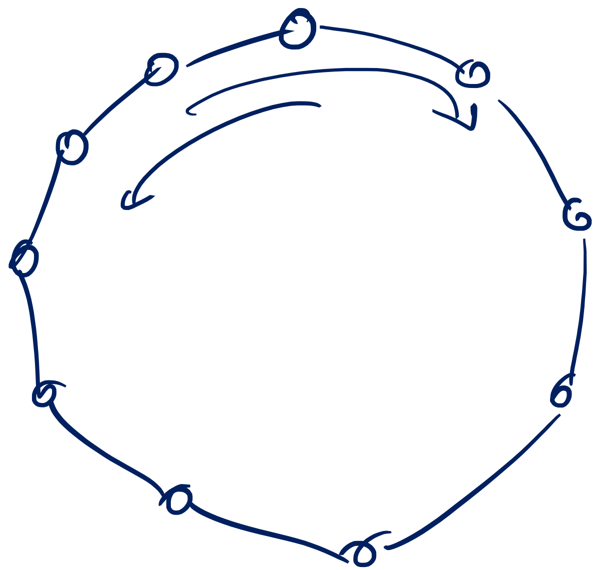
More formally:

- 3 states: undecided, leader, non-leader
- Initially, every node is in the undecided state
- When leaving the undecided state, a node goes into a final state
  - Final state: leader or non-leader
  - Implies termination…

Ring Network

For this lecture, we assume a ring topology

- Many important challenges already reveal on ring networks



n nodes

Anonymous Systems / Uniform Algorithms

Definition: A distributed system is called **anonymous** if the nodes ***do not have unique identifiers***.

- That is, initially all nodes are indistinguishable from each other

Definition: A distributed algorithms is called **uniform** if the **number of nodes n is not known** to the algorithm (i.e., to the nodes)
If **n is known**, the algorithm is called **non-uniform**.

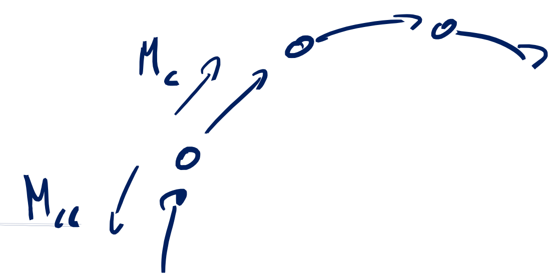
Leader Election in Anonymous Rings

- Is it possible to elect a leader in an anonymous ring?
  - Say if communication is synchronous and the ~~system~~ algorithm is non-~~anonymous~~ uniform?

Lemma: After k rounds of any deterministic algorithm on an anonymous ring, every node is in the same state S_k.

Anonymous → every node is in the same initial state S_0

Lemma follows by induction on rounds



after round i → state S_i

all nodes send same msg.

" " recv. " "

all nodes move to same new state S_{i+1}

Leader Election in Anonymous Rings

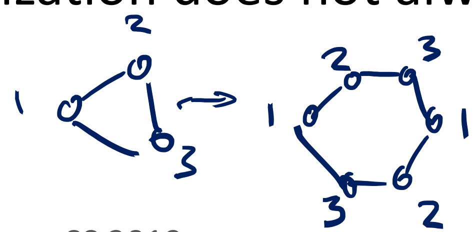
Theorem: Deterministic leader election in anonymous rings is impossible.

Proof:

- All nodes are always in the same state (previous lemma)
  → at the end either one or all nodes are in the leader state

Remarks:

- Holds for synchronous algorithms and thus also for asynchronous ones
- Holds for non-uniform algorithms and thus also for uniform ones
- Sense of direction does not help
  - Sense of direction: distinguish clockwise from counter-clockwise direction
- Randomization might help (can be used to break the symmetry)
- Randomization does not always help (for non-uniform alg.)



Leader Election in Asynchronous Rings

- For simplicity: assume sense of direction

Algorithm 1 (Clockwise leader election):

Each node v executes the following code:

1. Node v keeps stores largest known ID in m_v
2. Initialize $m_v \coloneqq \mathrm{ID}(v)$ and send $\mathrm{ID}(v)$ to clockwise neighbor
3. **if** v receives message with $\mathrm{ID}(w) > m_v$ **then**
4. v forwards $\mathrm{ID}(w)$ to clockwise neighbor and sets $m_v \coloneqq \mathrm{ID}(w)$
5. v decides not to be the leader if it has not done so already
6. **else** if v receives message with $ID(v)$ **then**
7. v decides to be the leader

Clockwise Leader Election: Analysis

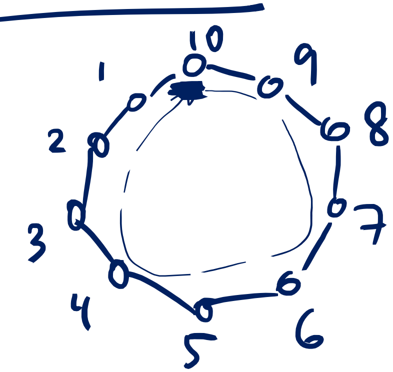
Theorem: The clockwise leader election algorithm correctly solves the leader election problem in $O(n)$ time with message complexity $O(n^2)$.

correctness: largest ID will make completely around the ring
every other ID will not make it around the ring

time compl.: largest ID gets back to its node after
n time steps

msg. compl.: $O(n^2)$ trivial

bad execution



Clockwise Leader Election: Analysis

Theorem: The clockwise leader election algorithm correctly solves the leader election problem in $O(n)$ time with message complexity $O(n^2)$.

Clockwise Leader Election: Analysis

Theorem: The clockwise leader election algorithm correctly solves the leader election problem in $O(n)$ time with message complexity $O(n^2)$.

Remarks:

- Time complexity is optimal, message complexity maybe not?
- Algorithm distinguishes clockwise and counter-clockwise neighbors
  - This is not really necessary

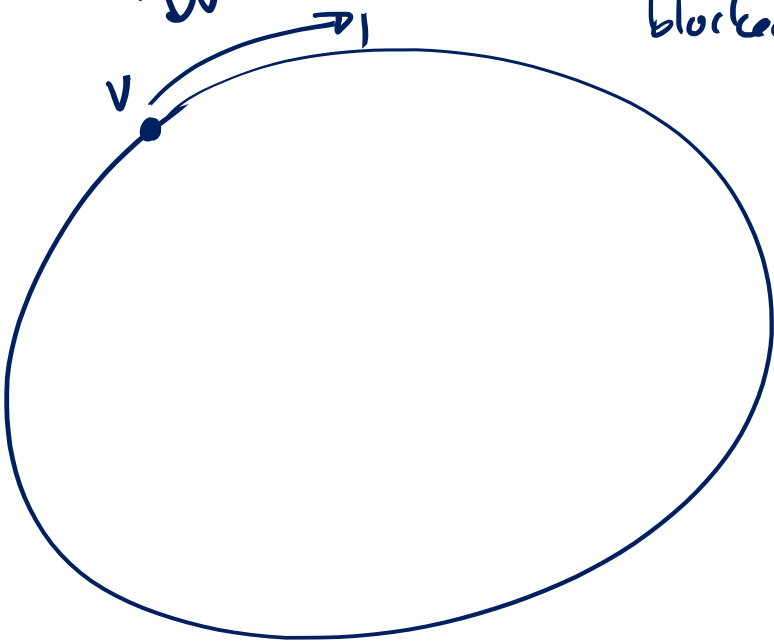
How can we improve the message complexity?

- choose the IDs at random
- choose some random candidates ...

Randomized Clockwise Leader Election

Theorem: With random IDs, the clockwise leader election algorithm has an expected message complexity of $O(n \log n)$.

D_v : # steps of ID(v) before it is blocked



msg. compl. M

$$M = \sum_{v \in V} D_v$$

$$\mathbb{E}[M] = \mathbb{E}\Big[\sum_{v \in V} D_v\Big]$$

linearity of exp.

$$= \sum_{v \in V} \mathbb{E}[D_v]$$

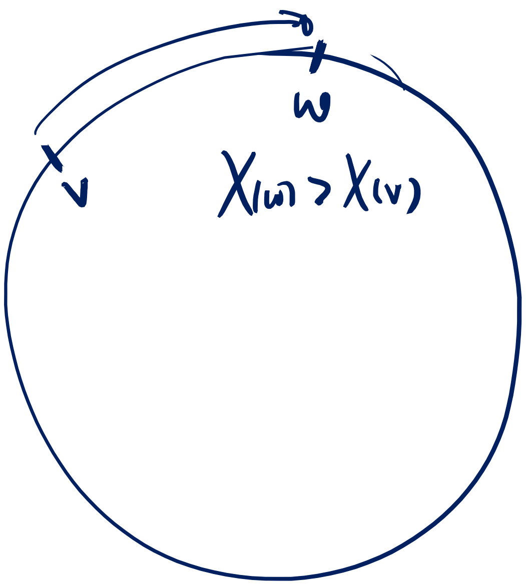
$$= n \cdot \mathbb{E}[D_v]$$

Randomized Clockwise Leader Election

Theorem: With random IDs, the clockwise leader election algorithm has an expected message complexity of $O(n \log n)$.

What is $E[D_v]$

$X(v)$: random number (random ID)



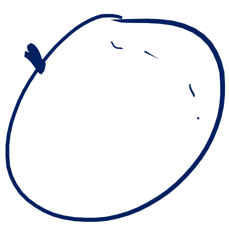
R_v: rank of the rand. # of v

($R_v = 1$: v has largest #)

($R_v = n$: v has smallest #)

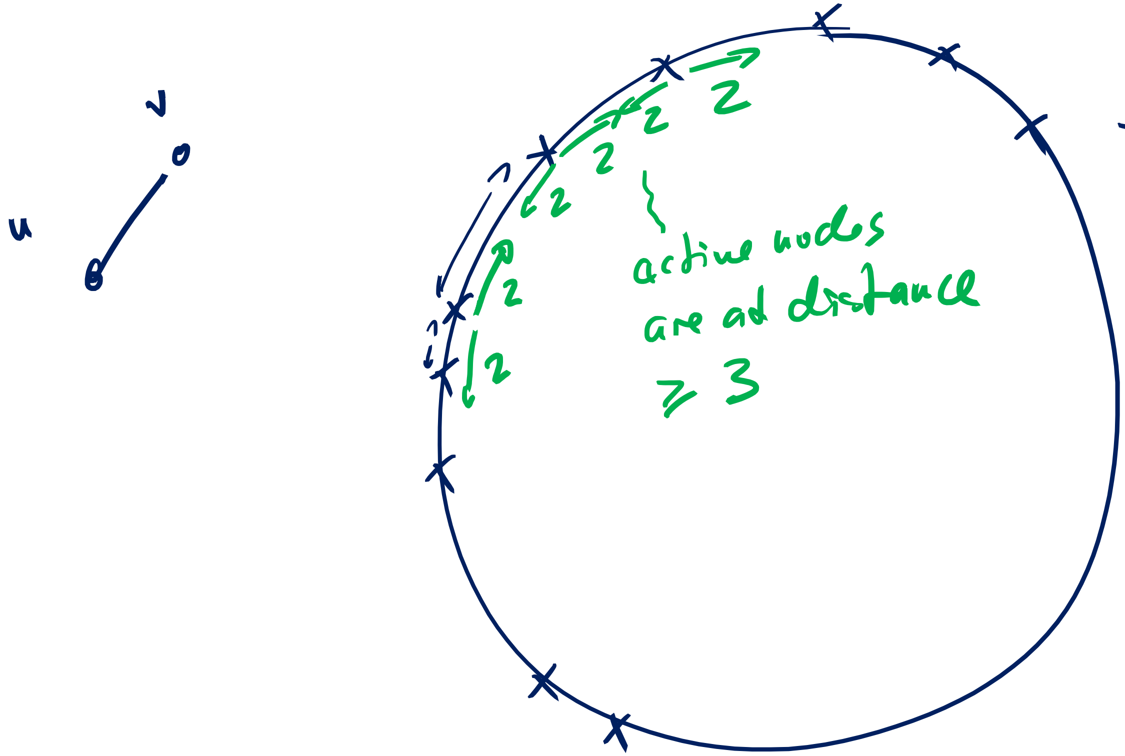
$$E[D_v] = \sum_{r=1}^{n} \underbrace{E[D_v \mid R_v = r]}_{\le \frac{n}{r}} \cdot \underbrace{P(R_v = r)}_{= \frac{1}{n}}$$

$$\le \sum_{r=1}^{n} \frac{1}{r} = H_n \le \ln n + 1$$



A Deterministic Message-Efficient Algorithm?

- Try to make sure that most IDs are not sent very far



- at the start all nodes are active
- exchange IDs with their neighbors
- become inactive if some neighbor has a larger ID

→ no two neighboring active nodes

→ at most $n/2$ active nodes

Radius Growth Algorithm

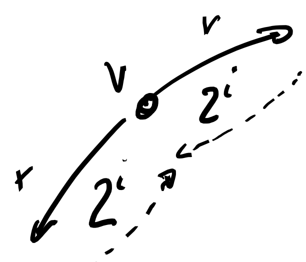
Basic idea:

- The algorithm consists of phases, initially all nodes (IDs) are active
- After phase $i \geq 1$, distance between any two active nodes is $> 2^{i+1}$

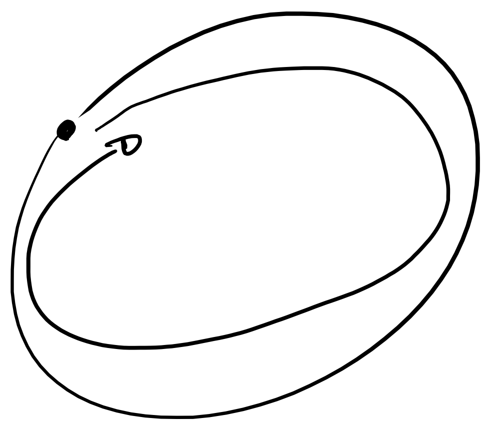
Algorithm:

phase i

active nodes send message to distance 2^i in each dir.
and send back echo



at some point $2^i \geq n$



Radius Growth Algorithm: Analysis

Theorem: The radius growth algorithm solves uniform, asynchronous leader election in time $O(n)$ with message complexity $\underline{O(n \log n)}$.

time compl.: phase i : $O(2^i)$

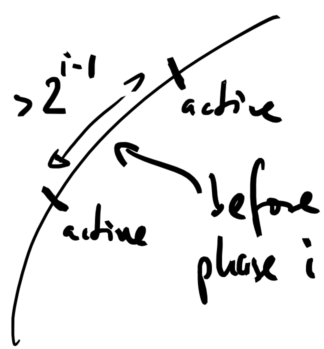
largest phase $\hat{i} = \lceil \log_2 n \rceil$

total time : $2 \cdot \sum_{i=0}^{\lceil \log_2 n \rceil} 2^i = 2 \cdot \left(2^{\lceil \log_2 n \rceil} - 1\right)$

$\leq 2 \cdot 2^{\log_2 n + 1} = \underline{4 \cdot n}$

msg. compl.:

phase i



$\Rightarrow$ # ~~at~~ active nodes $\leq \frac{n}{2^{i-1}} = \frac{2n}{2^i}$

each sends $4 \cdot 2^i$ msg.

total #msg.: $\leq \frac{2n}{2^i} \cdot 4 \cdot 2^i = \underline{\underline{8n}}$

Message Complexity Lower Bound

Recall: The asynchronous execution / schedule of a message passing algorithm is defined by the sequence of send and receive events

Remarks:

- We will assume that no two events happen at the same time
  - Such events can be ordered arbitrarily
- An execution of an asynchronous algorithm is determined by the algorithm and by an "adversarial" scheduler that decides about message delays, etc.
  - When proving a lower bound, we take the role of the scheduler
- We assume FIFO order for messages on the same edge
  - Only makes a lower bound stronger (and can always be enforced)

Message Complexity Lower Bound

Assumptions: For simplicity, we make the following assumptions:

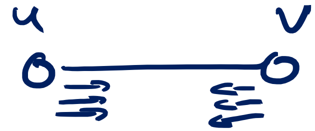
1. Asynchronous ring, where nodes may wake up at arbitrary times (but at the latest when receiving the first message)
   - For convenience, we will assume that $n = 2^k$

2. Uniform algorithms where the maximum ID node is elected as the leader
   - Assumption can be dropped with a more careful analysis

3. Explicit leader election (every node needs to learn the max. ID)
   - Can be enforced with additional $O(n)$ messages (at the end, the leader can send its ID around the ring)

4. For the proof, we have to play the adversary and specify in which order the messages are delivered…

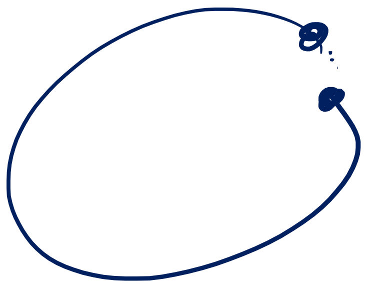
Open Schedule



Open Edge: Given a (partial) schedule, an edge $\{u, v\}$ is called open if no message has been received over this edge.

- Some messages might have been sent but not received over the edge

Open Schedule: A schedule for a ring is open if there is an open edge.

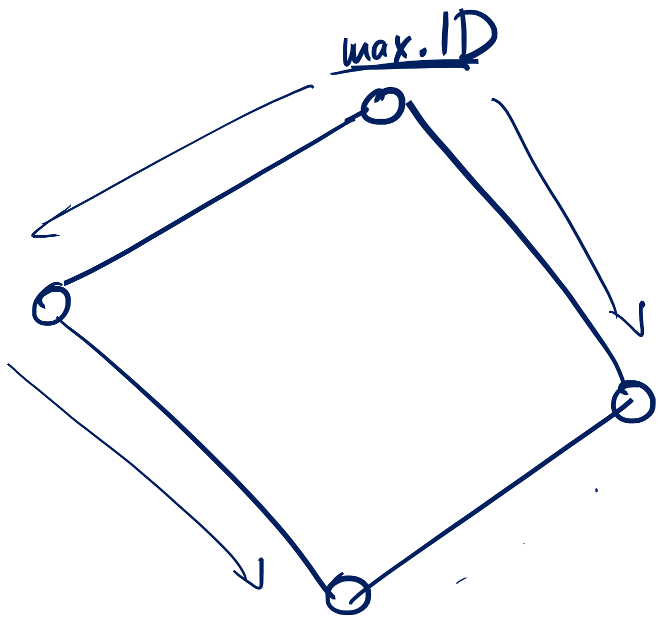


Open schedule message complexity:

- $\boldsymbol{M(n)}$: Given a ring of size n, for every asynchronous uniform leader election algorithm (and every possible assignment of IDs), there is an execution that produces an open schedule in which at least $M(n)$ messages have been received.
  - We will show that $M(n) = \Omega(n \cdot \log n)$ (by induction on n).

Open Schedule: Base Case

Lemma: Consider a cycle with $n = 4$ nodes. We can create an open schedule in which at least 3 messages are received.

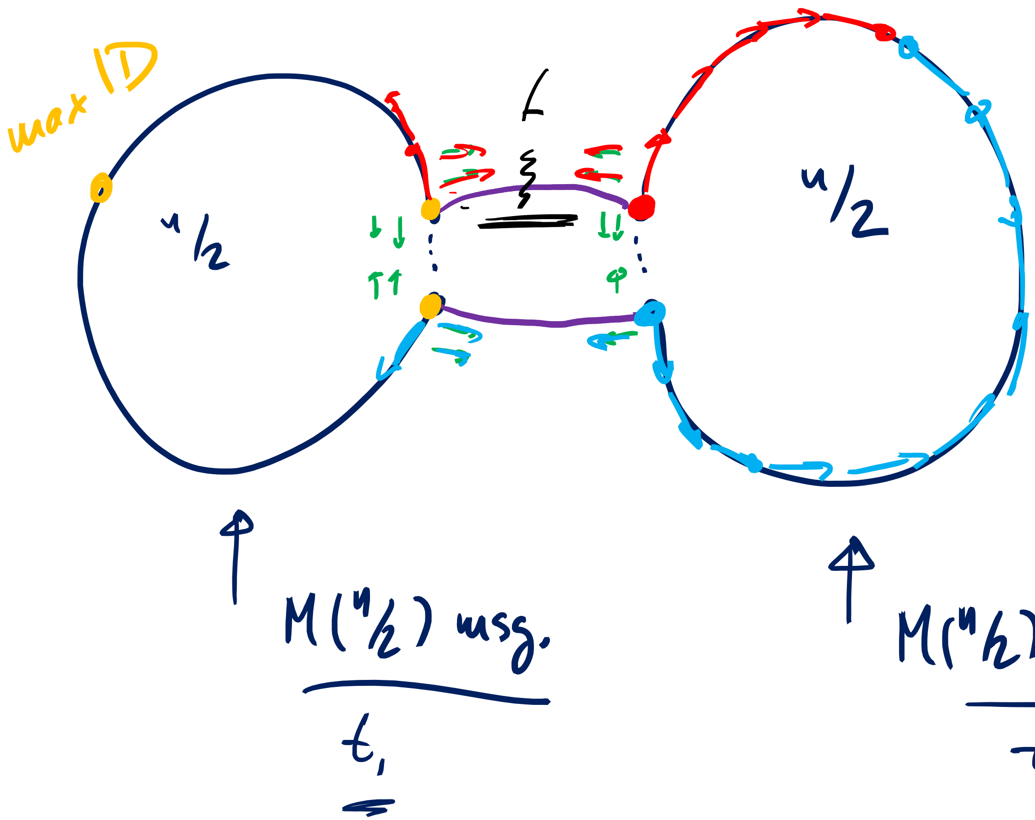


$M(4) \geq 3$

Open Schedule: Induction Step

Lemma: For $n = 2^k$ and integer $k \geq 3$, we have

$$M(n) \geq 2 \cdot M(n/2) + n/4.$$



red + light blue msg. ≥ n/2

Open Schedule: Induction Step

Lemma: For $n = 2^k$ and integer $k \geq 3$, we have

$$M(n) \geq 2 \cdot M\left({}^{n}/_{2}\right) + {}^{n}/_{4}.$$

Message Complexity Lower Bound

Theorem: Any uniform leader election algorithm in uniform rings of size n ($n = 2^k$ for $k \geq 2$) has message complexity at least

$$M(n) \geq \underline{{}^{n}/_{4} \cdot (\log n + 1)} = \Omega(n \log n).$$

by induction on n

$M(4) \geq 3$ ✓

ind. step: $M(n) \geq 2\underline{M(n/2)} + n/4$

$$\stackrel{(I.H.)}{\geq} 2\left(\frac{n}{8}\left(\underbrace{\log\frac{n}{2}+1}_{\log n}\right)\right) + \frac{n}{4} = \underline{\frac{n}{4}\log n + \frac{n}{4}} \quad ✓$$

Leader Election in Synchronous Rings

- Can we improve the message complexity for synchronous rings?
  - Assume that the algorithm is non-uniform (n is known)
  - Assume IDs are positive integers from $\{1, \dots, N\}$

Synchronous Leader Election Algorithm

- Algorithm consists of phases $i = 1,2, \dots$ of length n
- Every node v does the following

if phase $i = \mathrm{ID}(v)$ and v has not yet received a message **then**
    v becomes the leader
    v sends message "v is leader" arounds the ring

Leader Election in Synchronous Rings

Synchronous Leader Election Algorithm

- Algorithm consists of phases $i = 1,2, \ldots$ of length n
- Every node v does the following

if phase $i = \mathrm{ID}(v)$ and v has not yet received a message **then**
  v becomes the leader
  v sends message "v is leader" arounds the ring