

Minimum Spanning Tree,
Leader Election,
Synchronizers

Minimum Spanning Tree

Given a weighted graph $G = (V, E)$, generate a **spanning tree** $T = (V, E')$ such that the *sum of the weights of all the edges is minimum*.

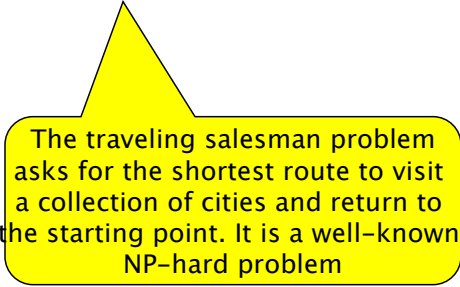
A few applications

Minimum cost vehicle routing.

A cable TV company will use this to lay cables in a new neighborhood.

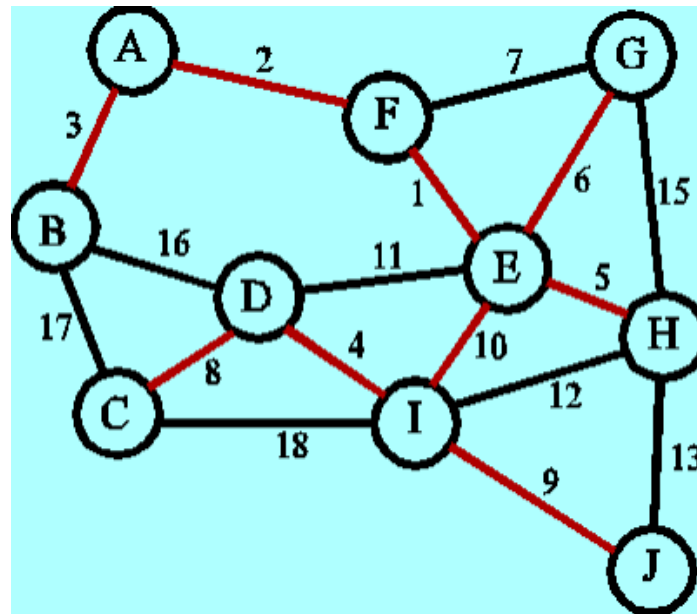
On Euclidean plane, *approximate* solutions to the **traveling salesman problem**,

We are interested in distributed algorithms only



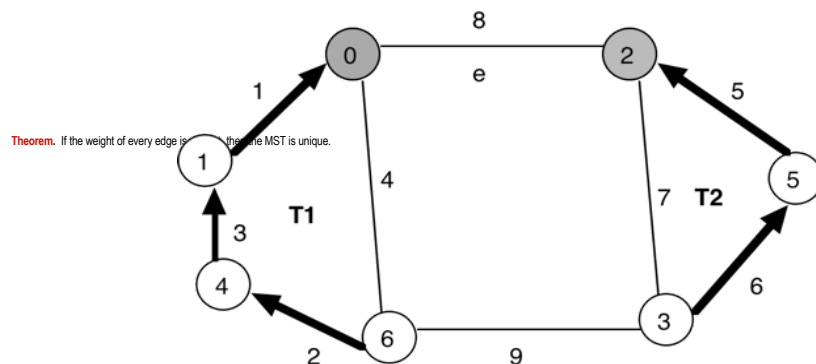
The traveling salesman problem asks for the shortest route to visit a collection of cities and return to the starting point. It is a well-known NP-hard problem

Example



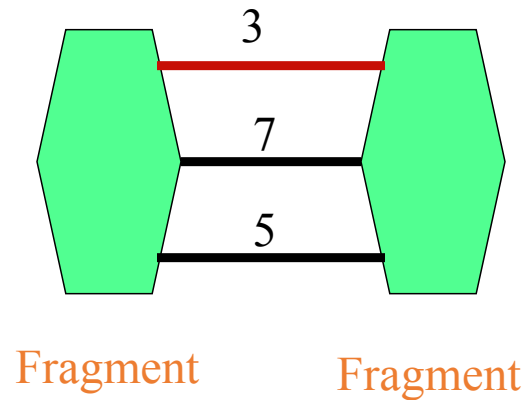
Sequential algorithms for MST

Review (1) Prim's algorithm and (2) Kruskal's algorithm (greedy algorithms)

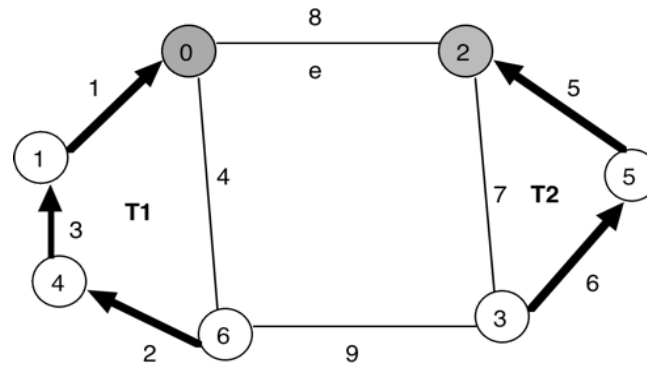


Gallagher-Humblet-Spira (GHS) Algorithm

- GHS is a distributed version of Prim's algorithm.
- Bottom-up approach. MST is recursively constructed by fragments joined by an edge of least cost.



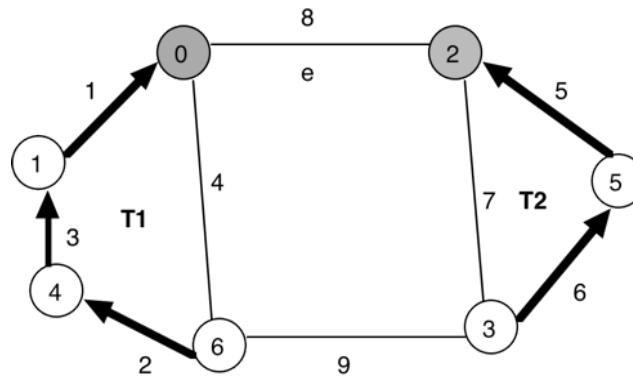
Challenges



Challenge 1. How will the nodes in a given fragment identify the edge to be used to connect with a different fragment?

A root node in each fragment is the coordinator

Challenges



Challenge 2. How will a node in **T1** determine if a given edge connects to a node of a different tree **T2** or the same tree **T1**? Why will node 0 choose the edge **e** with weight **8**, and not the edge with weight **4**?

*Nodes in a fragment acquire the **same name** before augmentation.*

Two main steps

- Each fragment has a **level**. Initially each node is a fragment at level 0.
- **(MERGE)** Two fragments at the same level L combine to form a fragment of level $L+1$
- **(ABSORB)** A fragment at level L is absorbed by another fragment at level L' ($L < L'$). The new fragment has a level L' .

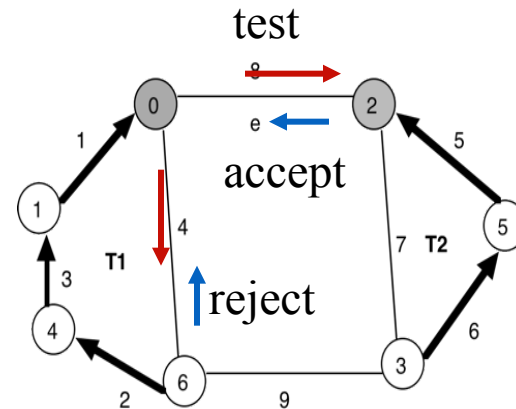
(Each fragment in level L has at least 2^L nodes)

Least weight outgoing edge

To test if an edge is **outgoing**, each node sends a **test** message through a candidate edge. The receiving node may send **accept** or **reject**.

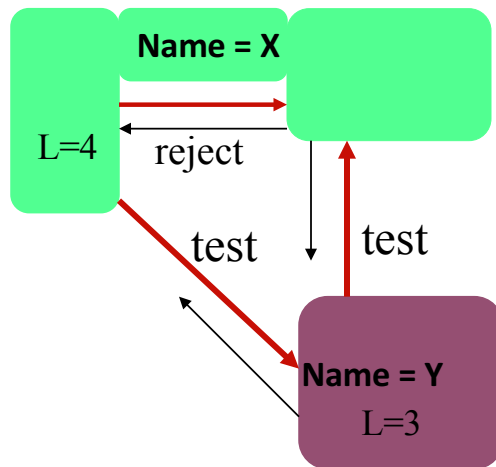
Root broadcasts **initiate** in its own fragment, **collects the report from other nodes** about eligible edges using a **convergecast**, and determines the **least weight outgoing edge**.

(Broadcast and Convergecast are two handy tools)



Accept or reject?

Let i send **test** to j



Case 1. If $\text{name}(i) = \text{name}(j)$ then send **reject**

Case 2. If $\text{name}(i) \neq \text{name}(j)$ AND $\text{level}(i) \leq \text{level}(j)$ then node j sends **accept**

Case 3. If $\text{name}(i) \neq \text{name}(j)$ AND $\text{level}(i) > \text{level}(j)$ then **wait until** $\text{level}(j) = \text{level}(i)$ and then send **accept/reject**. **WHY?** (See note below)

(Also note that levels can only increase).

Q: Can fragments wait for ever and lead to a deadlock?

Note. It may be the case that the responding node belongs a different fragment when it received the test message, but it is also trying to merge with the sending fragment.

The major steps

repeat

- 1 Test edges as outgoing or not
- 2 Determine **least weight outgoing edge** - it becomes a tree edge
- 3 Send **join** (or respond to **join**)
- 4 Update level & name & identify new coordinator/root

until there are no outgoing edges

Types of messages

(Initiate) Root initiates the “lwoe” search

(report) Nodes respond to the root with info about outgoing edges

(test) Nodes test if an edge is outgoing

(accept) The recipient of the test message certifies the edge as “*outgoing*”

(reject) The recipient of the test message certifies the edge as “*not outgoing*”

(join) Nodes bordering the edge send **join** to the fragment at the other end

(change_root) *Nodes broadcast the id of the new root in the fragment*

Classification of edges

- **Basic** (initially all branches are basic)
- **Branch** (all tree edges)
- **Rejected** (not a tree edge)

Branch and **rejected** are **stable attributes**

(once tagged as **rejected**, it remains so for ever.

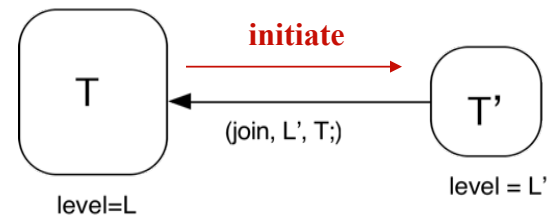
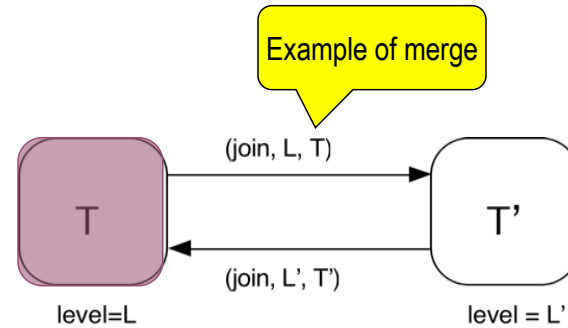
The same thing holds for **tree edges** too.)

Wrapping it up

Merge

The edge through which the **join** message is exchanged, changes its status to *branch*, and it becomes a tree edge.

The new root broadcasts an **(initiate, L+1, name)** message to the nodes in its own fragment.

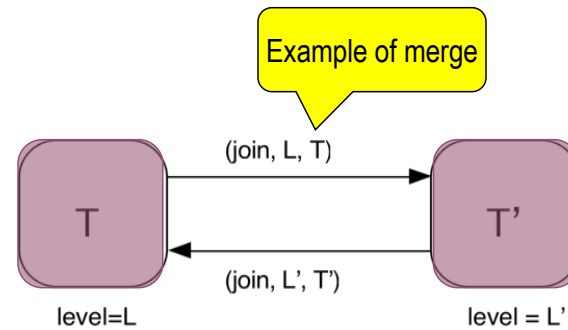


Wrapping it up

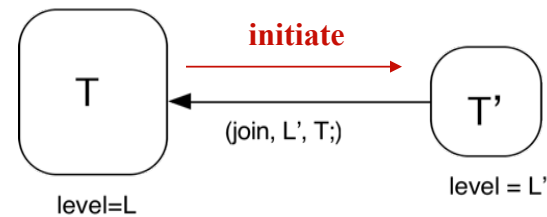
Merge

The edge through which the **join** message is exchanged, changes its status to *branch*, and it becomes a tree edge.

The new root broadcasts an **(initiate, L+1, name)** message to the nodes in its own fragment.



(a) $L = L'$

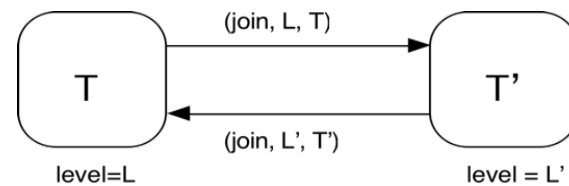


(b) $L > L'$

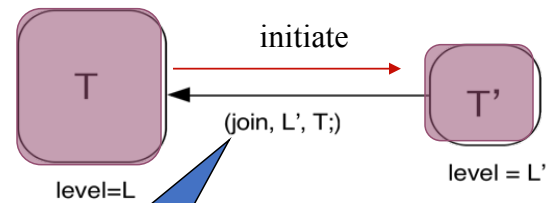
Wrapping it up

Absorb

T' sends a **join** message to T, and receives an **initiate** message. This indicates that the fragment at level L has been absorbed by the other fragment at level L'. They collectively search for the **lwoe**. The edge through which the **join** message was sent, changes its status to **branch**.



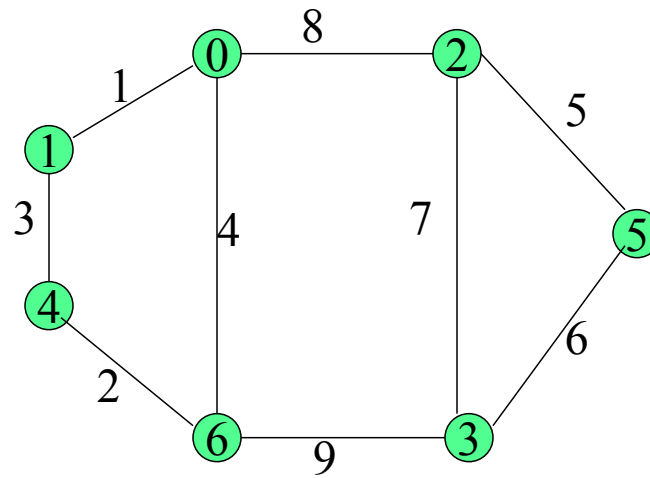
(a) $L = L'$



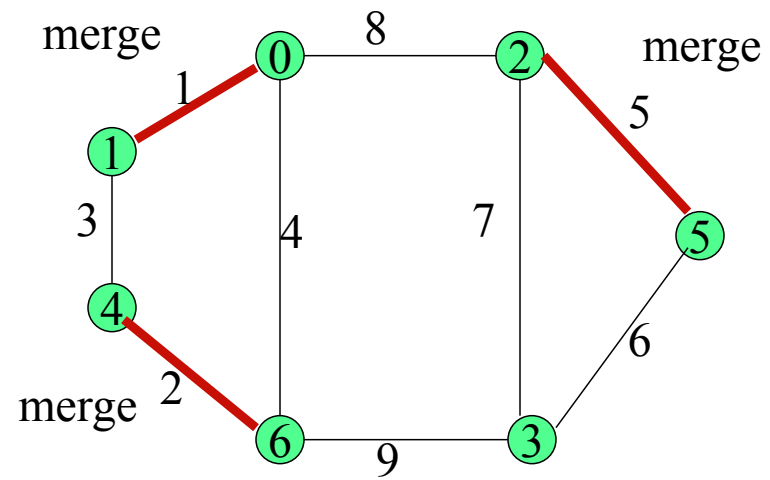
Example of absorb

(b) $L > L'$

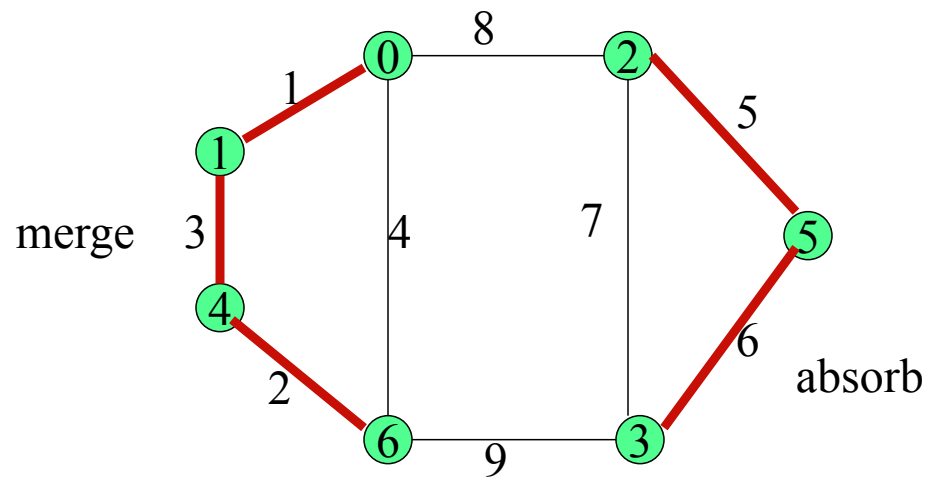
Example



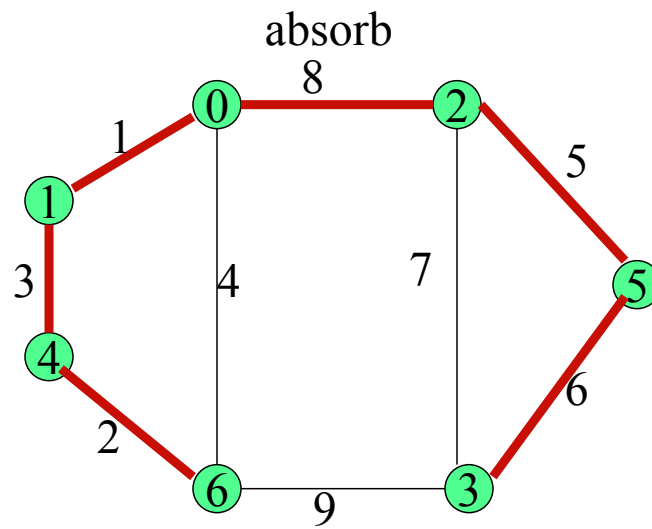
Example



Example



Example



Message complexity

Each edge may be rejected at most once. It requires two messages (*test + reject*). The upper bound is $2|E|$ messages.

At each of the (max) $\log N$ levels, a node RECEIVES at most (1) one *initiate* message and (2) one *accept* message and SENDS (3) one *report* message (4) one *test* message *not* leading to a rejection, and (5) one *changeroot* or *join* message.

So, the total number of messages has an upper bound of $2|E| + 5N \log N$

Coordination Algorithms:

Leader Election

Leader Election

Let $G = (V, E)$ define the network topology. Each process i has a variable $L(i)$ that defines the *leader*. The goal is to reach a configuration, where

$\forall i, j \in V : i, j$ are non-faulty ::

- (1) $L(i) \in V$ **and**
- (2) $L(i) = L(j)$ **and**
- (3) $L(i)$ is non-faulty

Often reduces to *maxima (or minima) finding problem*.
(if we ignore the failure detection part)

Leader Election

Difference between mutual exclusion & leader election

The similarity is in the phrase “at most one process.” But,

Failure is not an issue in mutual exclusion, a new leader is elected only after the current leader fails.

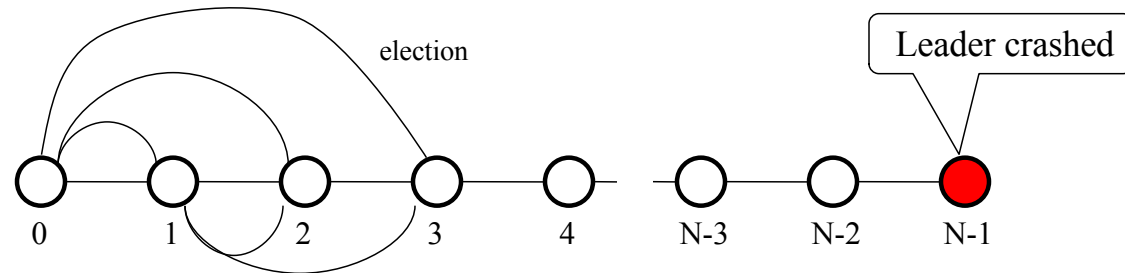
No fairness is necessary - it is not necessary that every aspiring process has to become a leader.

Bully algorithm

(Assumes that the topology is completely connected)

1. Send *election* message (*I want to be the leader*) to processes with *larger id*
2. Give up your bid if a process with *larger id* sends a *reply* message (*means no, you cannot be the leader*). In that case, wait for the *leader* message (*I am the leader*). Otherwise elect yourself the leader and send a *leader* message
3. If *no reply is received*, then elect yourself the leader, and broadcast a *leader* message.
4. If you receive a reply to the election message, but later don't receive a *leader* message from a process of larger id (i.e. the leader-elect has crashed), then re-initiate election by sending *election* message.

Bully algorithm



Node 0 sends N-1 **election** messages
Node 1 sends N-2 **election** messages
Node N-2 sends 1 **election** messages etc

Finally, node N-2 will be elected leader, but before it sent the **leader** message, it crashed. So, 0 starts all over again

The worst-case message complexity = **$O(n^3)$** (This is bad)

Maxima finding on a unidirectional ring

Chang-Roberts algorithm (asynchronous)

Initially all initiator processes are **red**.

Each initiator process i sends out **token $\langle i \rangle$**

{For each initiator i }

do token $\langle j \rangle$ received $\wedge j < i \rightarrow$ skip (do nothing)

token $\langle j \rangle \wedge j > i \rightarrow$ send token $\langle j \rangle$; color := **black**

token $\langle j \rangle \wedge j = i \rightarrow L(i) := i$ **{ i becomes the leader}**

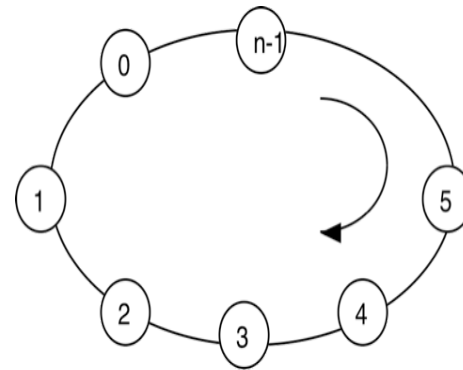
od

{Non-initiators remain **black**, and act as routers}

do token $\langle j \rangle$ received \rightarrow send $\langle j \rangle$ **od**

Message complexity = $O(n^2)$. Why?

What are the best and the worst cases?



The ids may not be nicely ordered like this

Bidirectional ring

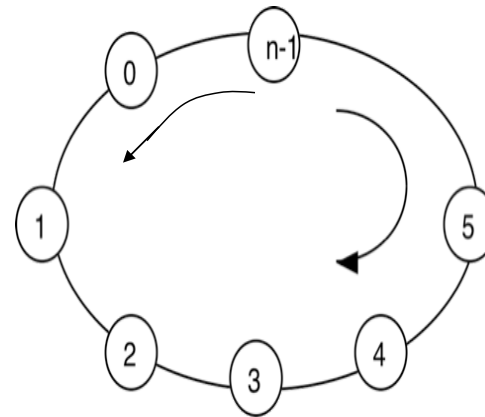
Franklin's algorithm (round based)

In each round, every process sends out *probes (same as tokens)* in **both** directions to its neighbors.

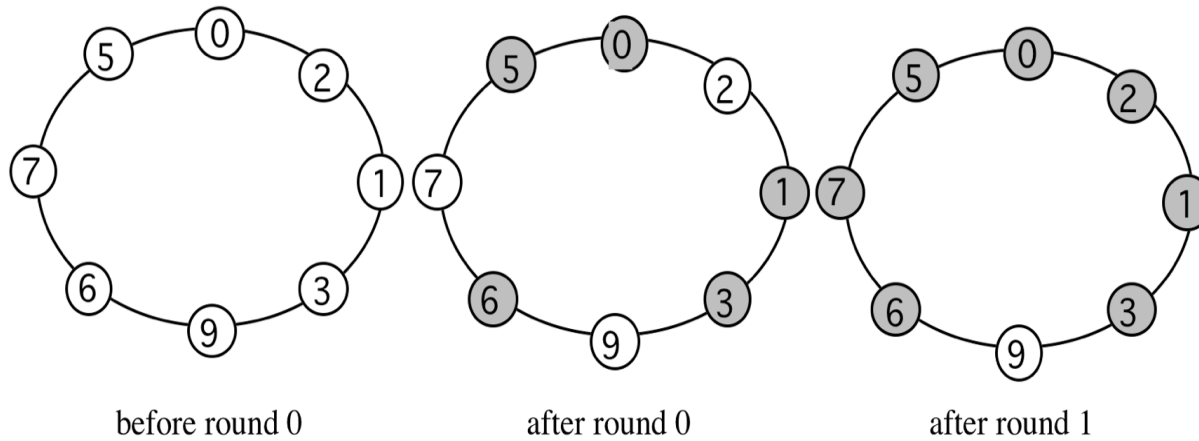
Probes from higher numbered processes will knock the lower numbered processes out of competition.

In each round, out of two neighbors, **at least one must quit**. So at least $1/2$ of the current contenders will quit.

Message complexity = $O(n \log n)$. Why?



Sample execution



Peterson's algorithm

initially $\forall i : \text{color}(i) = \mathbf{red}, \text{alias}(i) = \mathbf{i}$

{program for each round and for each red process}

send **alias**; receive **alias (N)**;

if $\text{alias} = \text{alias (N)} \rightarrow \mathbf{I am the leader}$

$\text{alias} \neq \text{alias (N)} \rightarrow$ send **alias(N)**; receive **alias(NN)**;

$\text{if } \text{alias(N)} > \max(\text{alias}, \text{alias (NN)}) \rightarrow \text{alias} := \text{alias (N)}$

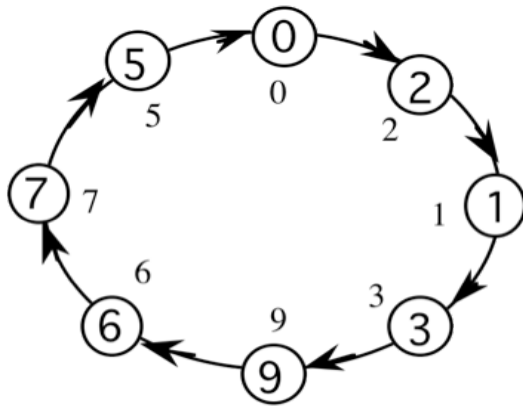
$\text{alias(N)} < \max(\text{alias}, \text{alias (NN)}) \rightarrow \text{color} := \mathbf{black}$

fi

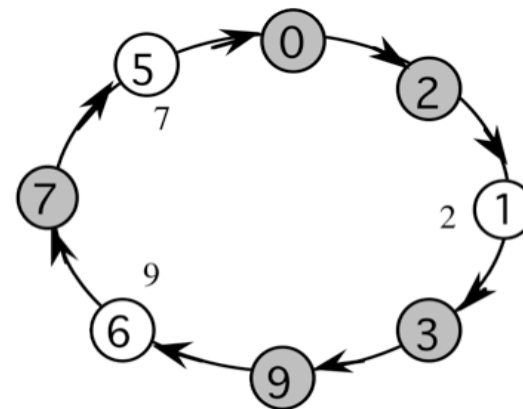
fi

{N(i) and NN(i) denote neighbor and neighbor's neighbor of i}

Peterson's algorithm



before round 0



after round 0

Round-based. Finds maxima on a **unidirectional ring** using $O(n \log n)$ messages. Uses an **id** and an **alias** for each process.

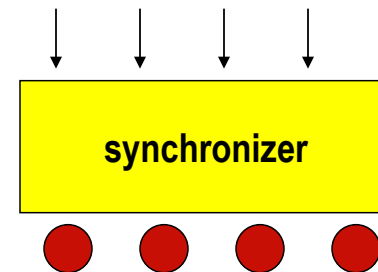
Synchronizers

Synchronous algorithms (round-based, where processes execute actions in lock-step synchrony) are easier to deal with than **asynchronous algorithms**. In each **round** (or **clock tick**), a process

- (1) receives messages from neighbors,
- (2) performs local computation
- (3) sends messages to ≥ 0 neighbors

A **synchronizer** is a protocol that enables synchronous algorithms to run on an asynchronous system.

Synchronous algorithm



Asynchronous system

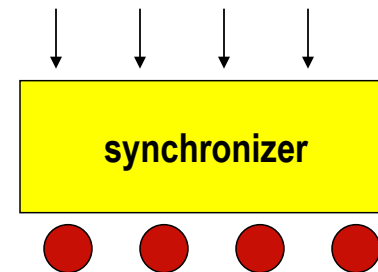
Global vs. Local synchronizers

In a **global synchronizer**, **no process** will receive or send any message for round r , **until all processes have sent and received their message** for round $(r-1)$

In a **local synchronizer**, **no process** will receive or send any message for round r , **until all its neighbors have sent and received their message** for round $(r-1)$. Thus, if a process receives a message for round r from one neighbor, and then receives a message for round $(r-1)$ from another neighbor, then the requirements of local synchronization are violated.

A global synchronizer is also a local synchronizer, but the converse is not true

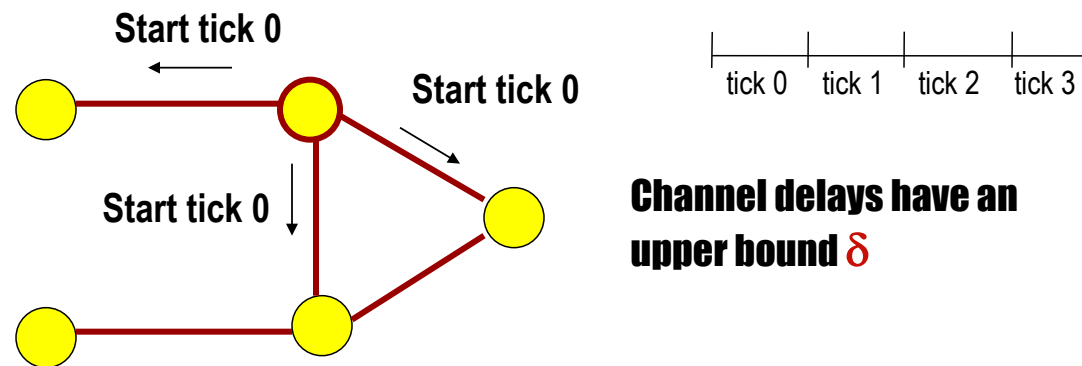
Synchronous algorithm



Asynchronous system

(Local) Synchronizers

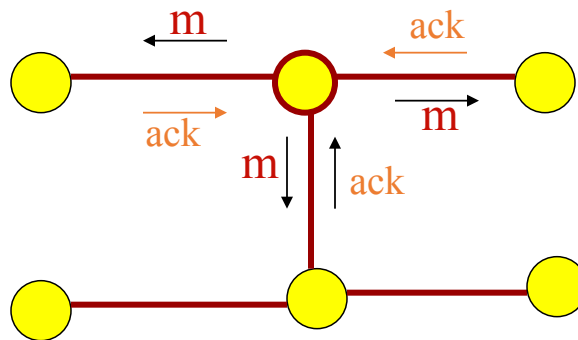
“Every message sent in *clock tick k* must be received by the neighbors in the *clock tick k*.” This is not automatic - some extra effort is needed.
Consider a basic *Asynchronous Bounded Delay (ABD)* synchronizer



Each process will *start the simulation of a new clock tick after 2δ time units*, where δ is the maximum propagation delay of each channel

α -synchronizers

It is a local synchronizer that does not use the delay bounds. Even when the propagation delay is arbitrarily large but finite, the α -synchronizer can handle this.



Simulation of each clock tick

1. Send and receive **messages** for the current tick.
2. Send **ack** for each incoming message, and receive **ack** for each outgoing message
3. Send a **safe message** to each neighbor after sending and receiving all **ack** messages (then follow steps 1-2-3-1-2-3- ...)

Complexity of α -synchronizer

Message complexity $M(\alpha)$

Defined as the number of messages passed around the *entire network* for the simulation of each clock tick.

$$M(\alpha) = O(|E|)$$

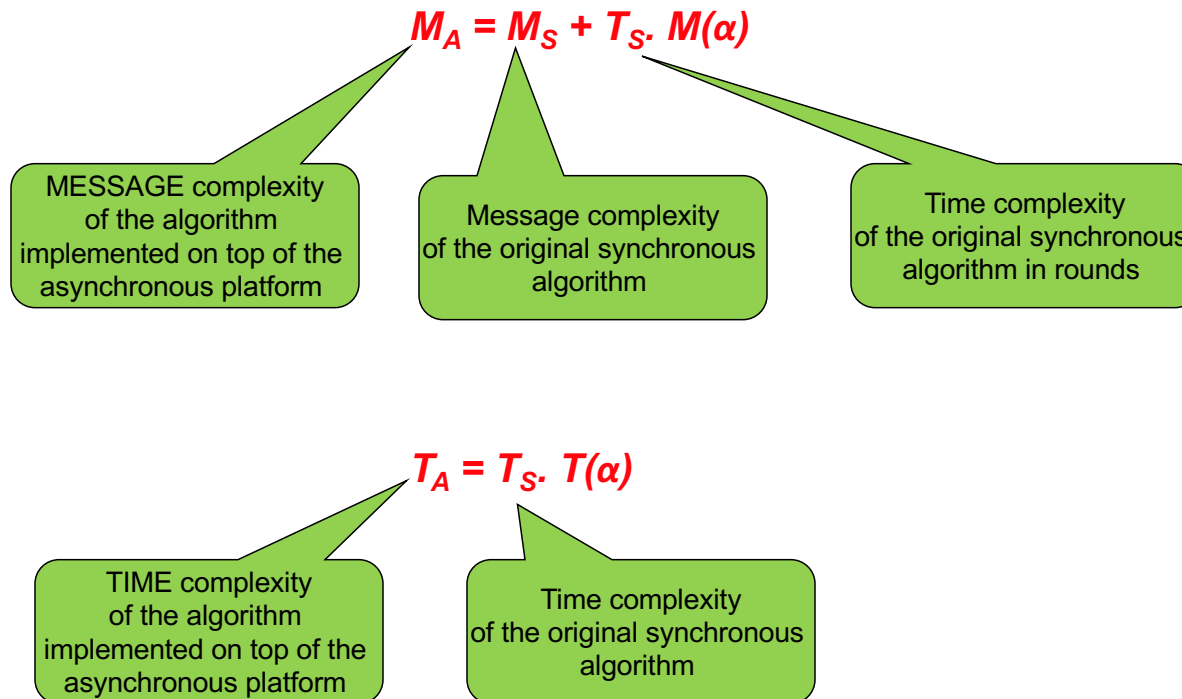
Time complexity $T(\alpha)$

Defined as the number of *asynchronous rounds* needed for the simulation of each clock tick.

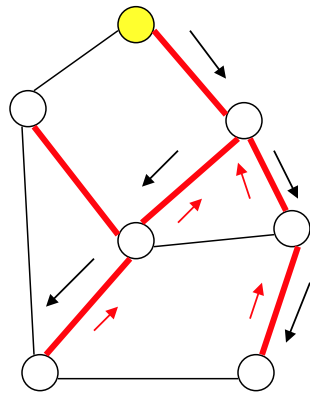
$$T(\alpha) = 3$$

(since each process exchanges *m*, *ack*, *safe*)

Complexity of α -synchronizer



The β -synchronizer



Form a *spanning tree* with any node as the root. The **root** initiates the simulation of each tick by sending message $m(j)$ for each clock **tick** j . Each process responds with $ack(j)$ and then with a **safe(j)** message **along the tree edges** (that represents the fact that the entire subtree under it is safe). When the root receives **safe(j)** from every child, it initiates the simulation of clock tick $(j+1)$ using a **next** message.

To compute the message complexity $M(\beta)$, note that in each simulated tick, there are m messages of the original algorithm, $(N-1)$ acks, and $(N-1)$ safe messages and $(N-1)$ next messages along the tree edges. So the message overhead is $O(N)$

*Time complexity $T(\beta)$ = depth of the tree.
For a balanced tree, this is $O(\log N)$*

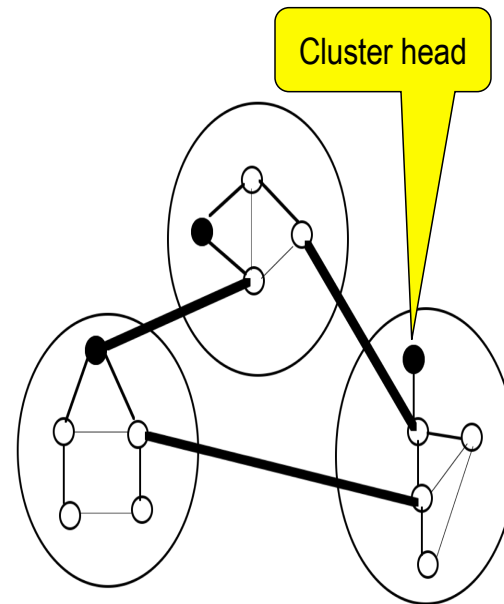
γ -synchronizer

Uses the best features of both α and β synchronizers. (*What are these?**)

The network is viewed as a tree of clusters. Each cluster has a cluster-head. Within each cluster, β -synchronizers are used, but for inter-cluster synchronization, α -synchronizer is used. For best complexity results, the cluster sizes must be carefully chosen.

Preprocessing overhead for cluster formation.

The number and the size of the clusters is a crucial issue in reducing the message and time complexities



* α -synch has lower time complexity, β -synchronizers have lower message complexity

Example of application: Shortest path

Consider Synchronous Bellman-Ford:

- $O(n |E|)$ messages, $O(n)$ rounds

– Asynchronous Bellman-Ford

- Many corrections possible (exponential), due to message delays.
- Message complexity can be exponential in n in the worst case
- Time complexity exponential in n , counting message pileups.

Using (e.g.) Synchronizer α :

- $M_A = M_S + T_S \cdot M(\alpha) = O(n \cdot |E|) + O(\text{diam}) \cdot O(|E|) = O(n \cdot |E|)$
- $T_A = T_S \cdot T(\alpha) = 3 \cdot \text{diam} \text{ rounds} = O(n) \text{ rounds}$