6.852: Distributed Algorithms
Fall, 2015

Lecture 14, Part 2

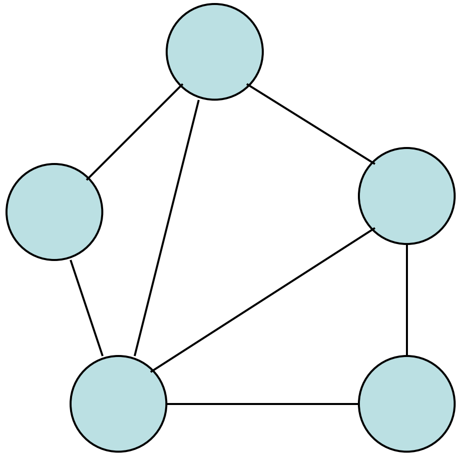
Today's plan

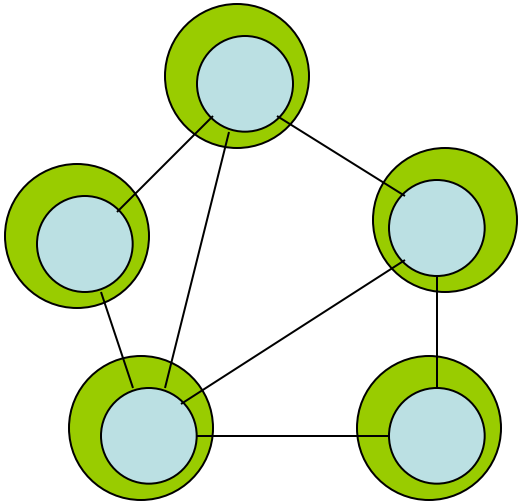
- Consistent global snapshots and stable property detection.
- Applications:
  - Distributed termination.
  - Deadlock detection.
  - Debugging
- Asynchronous shared memory model
- Readings: Chapter 19, Chapter 9
- Next:
  - Mutual exclusion
  - Reading: Sections 10.1-10.7

Consistent Global Snapshots and Stable Property Detection

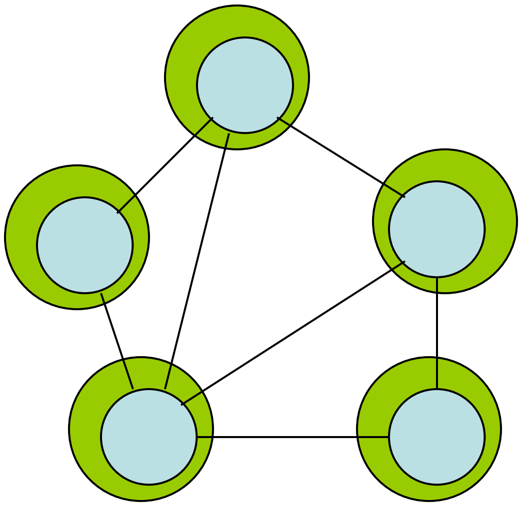
Consistent global snapshots and Stable property detection

- We have seen how logical time can be used to take a "global snapshot" of a running distributed system.
- Now examine global snapshots more closely.
- General idea:
  - Start with a distributed algorithm A, on an undirected graph G = (V,E).
  - Monitor A as it runs, and determine some property of its execution, e.g.:
    - Check whether certain invariants are true.
    - Check for termination, deadlock.
    - Compute some function of the global state, e.g., the total amount of money in a banking system.
    - Produce a complete snapshot for a backup.
- Monitored version: Mon(A)

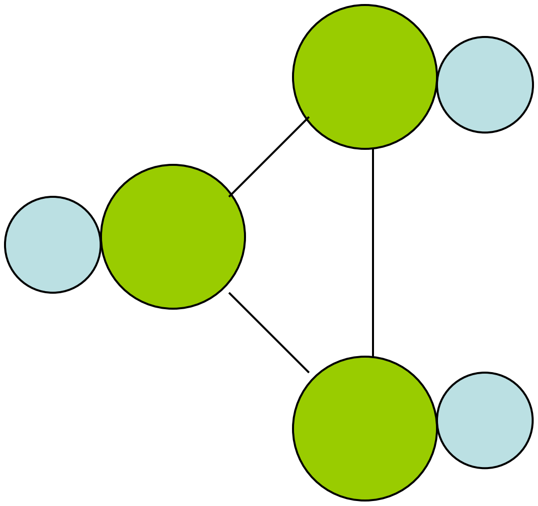




Mon(A)



- "Transformed version" of A.
- Mon(A) generally not obtained simply by composing each process A_i with a new monitor process.
- More tightly coupled.
- Monitoring process, $Mon(A)_i$, may "look inside" the corresponding A process, A_i, see the state.
- Superposition [Chandy, Misra]
  - Formalizes the permissible kinds of modifications.
  - Add new state components, new actions.
  - Modify old transitions, but only in certain permissible (nonintrusive) ways.



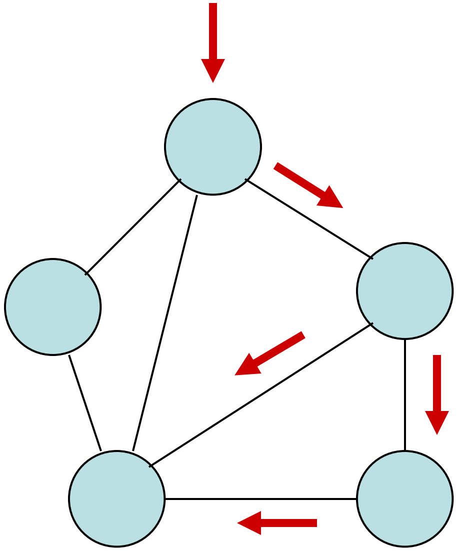
Key concepts

- Instantaneous snapshot:
  - Global state of entire distributed algorithm A, processes and channels, at some actual point in an execution.
  - Can use for checking invariants, checking for termination or deadlock, computing a function of the global state,…
- Consistent global snapshot:
  - Looks like an instantaneous snapshot, to every process and channel.
  - Good enough for checking invariants, checking for termination, …
- Stable property:
  - A property P of a global state such that, if P ever becomes true in an execution, P remains true forever thereafter.
  - E.g., termination, deadlock.
- Connection among these notions:
  - An instantaneous snapshot, or a consistent global snapshot, can be used to detect stable properties for a running distributed system.

Termination detection

[Dijkstra, Scholten]

Termination detection



- A simple stable property detection problem.
- Connected, undirected network graph G = (V,E).
- Assume:
  - Algorithm A begins with all nodes quiescent (only inputs enabled).
  - An input arrives at exactly one node.
  - Starting node need not be predetermined.
- From there, computation can “diffuse” throughout the network, or a portion of the network.
- At some point, the entire system may become quiescent again:
  - No non-input actions enabled at any node.
  - No messages in channels.
- Termination Detection problem:
  - If A ever reaches a quiescent state then the starting node eventually outputs “done”.
  - Otherwise, no one ever outputs “done”.
- To be solved by a monitoring algorithm Mon(A).

Dijkstra, Scholten Algorithm

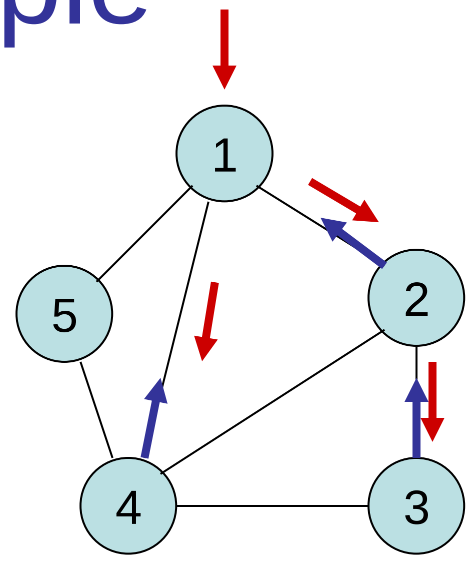
- Augment A with extra monitoring pieces that construct and maintain a tree, rooted at the starting node, and including all the nodes currently active in A.
- Grows, shrinks, grows,…as nodes become active, quiescent, active,…
- Algorithm:
  - Execute A as usual, but adding acks for all messages.
  - Messages of A treated like search messages in AsynchSpanningTree.
  - When a process receives an external input, it becomes the root, and begins executing A.
  - When any non-root process receives its first A message, it designates the sender as its parent in the tree, and begins participating in A.
  - Root process acks every message immediately.
  - Each other process acks all but the first message it receives immediately.
  - Convergecast for termination:

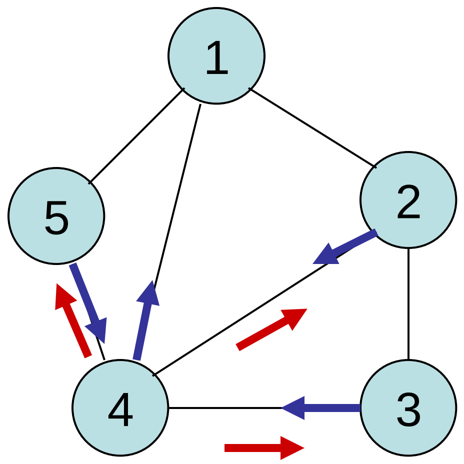
Dijkstra, Scholten Algorithm

- Algorithm:
  - Execute A as usual, adding acks for all messages.
  - When a process receives an external input, it becomes the root, and begins executing A.
  - When any non-root process receives its first A message, it designates the sender as its parent in the tree, and begins participating in A.
  - Root process acks every message immediately.
  - Other processes ack all but the first message immediately.
  - Convergecast for termination:
    - If a non-root process finds its A-state quiescent and all its A-messages acked, then it cleans up: acks the first A-message, deletes all info about the termination protocol, becomes idle.
    - If it later receives another A message, it treats it like the first A message wrt the termination protocol (defines a new parent, etc.), and resumes participating in A.
    - If root process finds its A-state quiescent and all its A-messages acked, it reports done.

DS Algorithm, example

- First, 1 gets awakened by an external A input, becomes the root, sends A messages to 2 and 4, 2 sends an A-message to 3, all set up parent pointers and start executing A.
- Next, 4 sends A message to 3, acked immediately.
- 4 sends A message to 1, acked immediately.
- 1, 2, 3, and 4 send A messages to each other for a while, everything gets acked immediately.
- Tree remains unchanged.
- Next, 2 and 3 quiesce locally; 3 cleans up, sends ack to 2, 2 receives ack, 2 cleans up, sends ack to 1.
- Next, 4 sends A messages to 2, 3, and 5, yielding a new tree:
- Etc.





Correctness

- Claim this correctly detects termination of A: that all A-processes are in quiescent states and no A-messages are in the channels.

- Theorem 1: If Mon(A) outputs “done” then A has really terminated.
- Proof sketch:
  - Depends on key invariants:
    - If root is idle (not actively engaged in the termination protocol), then all nodes are idle, and the channels are empty.
    - If a node is idle then the part of A running at that node is quiescent.

Correctness

- Theorem 2: If A ever becomes quiescent, then eventually Mon(A) outputs “done”.
- Proof sketch: [See book]
  - Depends on key invariants:
    - If the root is not idle, then the parent pointers form a directed tree directed toward the root, spanning exactly the non-idle nodes.
    - Conservation of acks.
  - Suppose for contradiction that A quiesces, but the termination protocol does not output “done”.
  - Then the spanning tree must eventually stabilize to some final tree.
    - Because no new A-messages are sent or received, and acks are eventually finished.
  - But then any leaf node of the final tree is able to clean up and become idle.
  - Shrinks the final tree further, contradicting stability.
  - Implies that the root must output “done”.

Complexity

- Messages:
  - 2m, where m is the number of messages sent in A.
- Time from quiescence of A until output “done”:
  - O(m d), where d = upper bound on message delay, ignore local processing time
  - Time to clean up the spanning tree.
- Bounds are most interesting if m << n.
  - E.g., for algorithms that involve only a limited computation in a small portion of a large network.

Application: Asynchronous BFS

- Recall Asynchronous Breadth-First Search algorithm (AsynchBFS).
- Allows corrections.
- Doesn’t terminate on its own; described ad hoc termination strategy earlier.
- It’s a diffusing algorithm:
  - Wakeup input at the root node.
- So we can apply [Dijkstra, Scholten] to get a simple terminating version.
- Similarly for AsynchBellmanFord shortest paths.

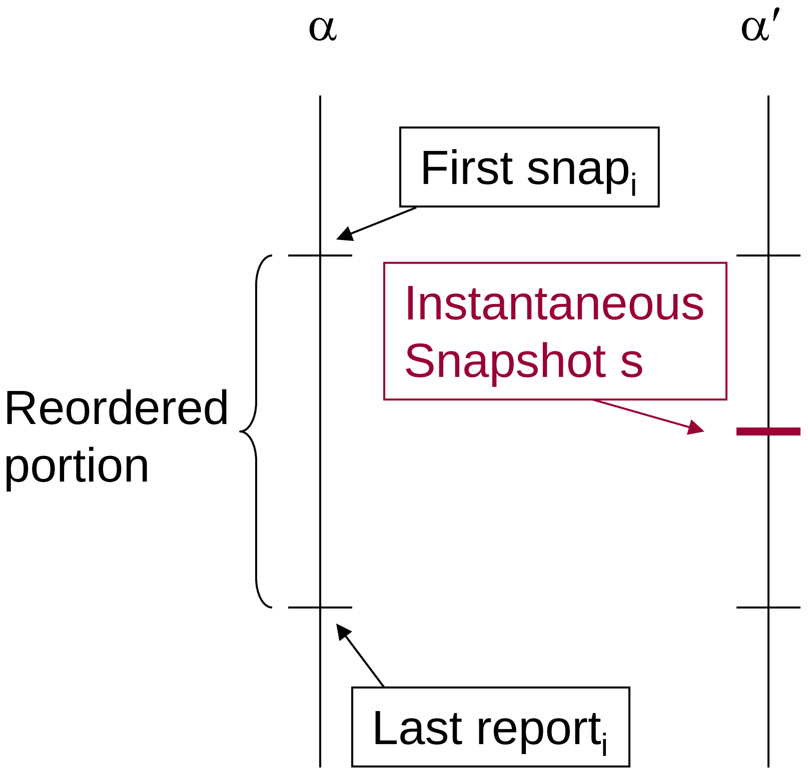
Consistent Global Snapshots
[Chandy, Lamport]

Consistent Global Snapshots

- Connected, undirected network graph G = (V,E).
- A is an arbitrary asynchronous distributed network algorithm.
- Mon(A) is supposed to take a "snapshot".
- Any number ($\geq$ 1) of nodes may receive $snap_i$ inputs, triggering the snapshot.
- Every node i should output $report_i$ containing:
  - A state for A_i.
  - States for all of A_i's incoming channels.
- Combination is a global state s.
- Must satisfy: If α is the actual underlying execution of A, then there is another execution, α', of A such that:
  - α and α' are indistinguishable to each individual A_i.
  - α and α' are identical up to the first snap and after the last report.
  - s is the global state at some point in α' in the snapshot interval.
- Implies the algorithm returns a Consistent Global Snapshot of A,
  - One obtained by reordering only the events occurring during the snapshot interval, and taking an instantaneous snapshot of the reordered execution, at some time during the snapshot interval.

Consistent Global Snapshot problem

- If α is the actual underlying execution of A, then there is another execution, α', of A such that:
  - α and α' are indistinguishable to each individual A_i.
  - α and α' are identical up to the first snap and after the last report.
  - s is the actual global state at some point in α' in the snapshot interval.



Chandy-Lamport algorithm

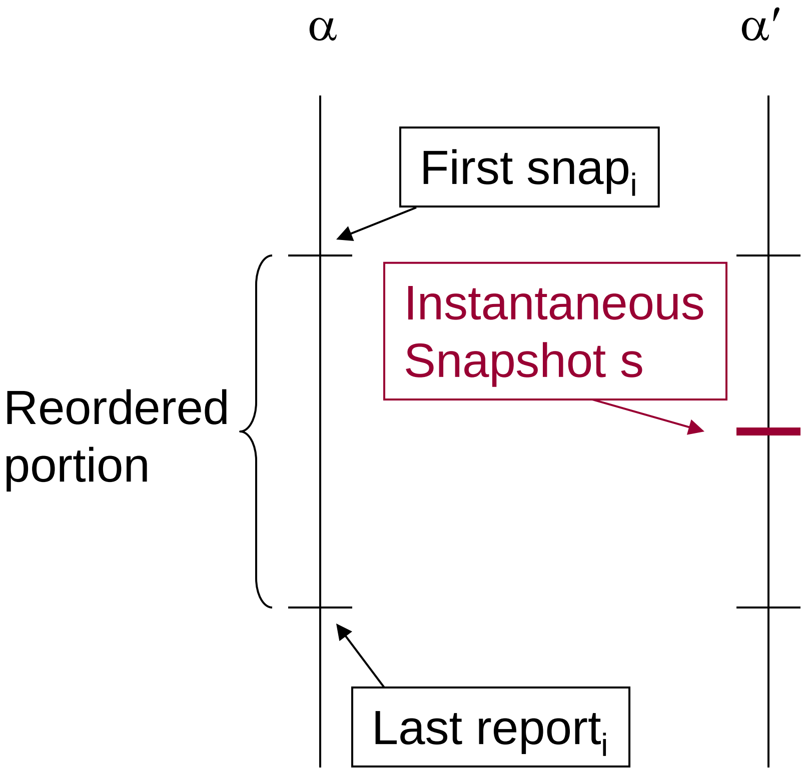
- Recall logical-time-based snapshot algorithm
  - Gets snapshot at a particular logical time t.
  - Depends on finding a good value of t.
- Chandy-Lamport algorithm can be viewed as running the same algorithm, but without explicitly using any particular logical time t.
- Instead, use marker messages to indicate where the logical time of interest occurs:
  - Put marker messages between messages sent at logical time $\leq$ t and those sent at logical times > t.
  - Relies on FIFO property of channels.

Chandy-Lamport algorithm

- Algorithm:
  - When not-yet-involved process i receives $snap_i$ input:
    - Snaps A_i's state.
    - Sends marker on each outgoing channel, thus marking the boundary between messages sent before and after the $snap_i$.
    - Thereafter, records all messages arriving on each incoming channel, up to the marker.
  - When process i receives a first marker message without having previously received $snap_i$:
    - Snaps A_i's state, sends out markers, and begins recording messages as before.
    - Channel on which it got the marker is recorded as empty.

Correctness

- Termination: Easy to see
  - All snap eventually, because of either snap input or marker message.
  - Markers eventually sent and received on all channels.
- Returns a correct global state:
  - Let α be the underlying execution of A.
  - We must produce α', show that the returned state is an instantaneous snapshot of α'.



Returns a correct global state

- Let α be the underlying execution of A.
- Divide events of α into:
  - S_1: Those before the snap at their processes
  - S_2: Those after the snap at their processes
- Every event of α belongs to some process, so is in S_1 or S_2.
- Obtain α' by reordering events of α between first snap and last report, putting all S_1 events before all S_2 events, preserving causality order.
  - Causality: Orders events at each process and sends vs. receives.
- Q: How do we know we can do this?
- Claim that no send appears in S_2 whose corresponding receive is in S_1.
- In other words, for every send in S_2, the corresponding receive is also in S_2.
- The points between S_1 and S_2 at all processes form a consistent cut.

Returns a correct global state

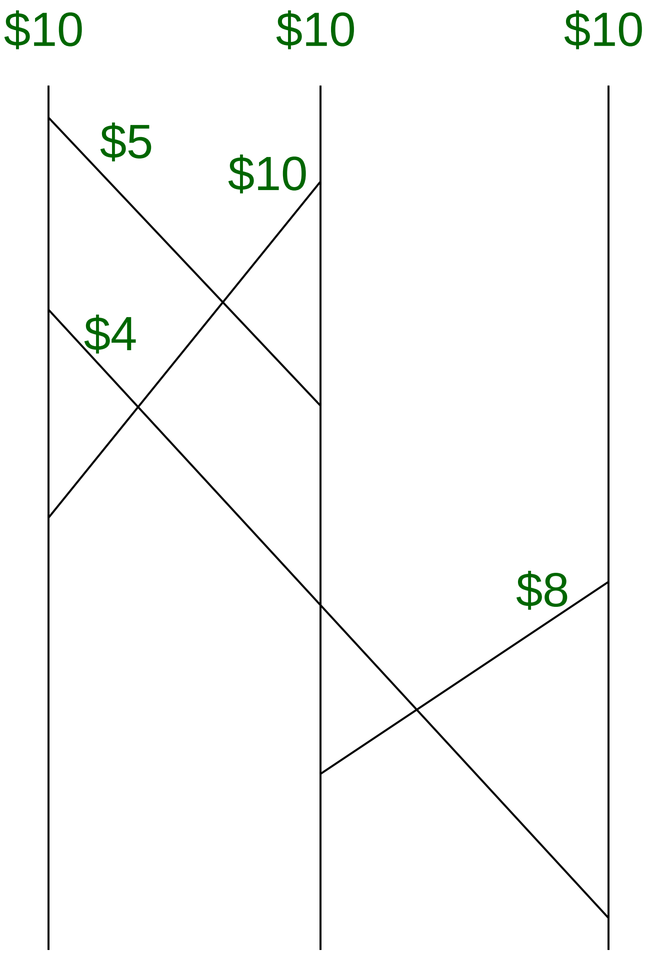
- Divide events of α into: S_1 (before snap) and S_2 (after snap).
- Obtain α' by reordering events of α between first snap and last report, putting all S_1 events before all S_2 events, preserving causality order.
- Can do this because no send appears in S_2 whose corresponding receive appears in S_1:
  - Follows from the marker discipline.
  - A send in S_2 occurs after the local snap, so after the marker is sent.
  - So the send produces a message that follows the marker on its channel.
  - Recipient snaps when it receives the marker (or sooner), so before receiving the message.
  - So the receive event is also in S_2.
- Returned state is exactly the global state of α' between the S_1 and S_2 events, that is, after all the pre-snap events and before all the post-snap events.
- Thus, returned state is an instantaneous snapshot of α'.

Remark

- Algorithm works in strongly-connected digraphs, as well as undirected graphs.

Example: Bank audit

- Distributed bank, money sent in reliable messages.
- Audit problem:
  - Count the total money in the bank.
  - While money continues to flow around.
  - Assume total amount of money is conserved (no deposits or withdrawals).

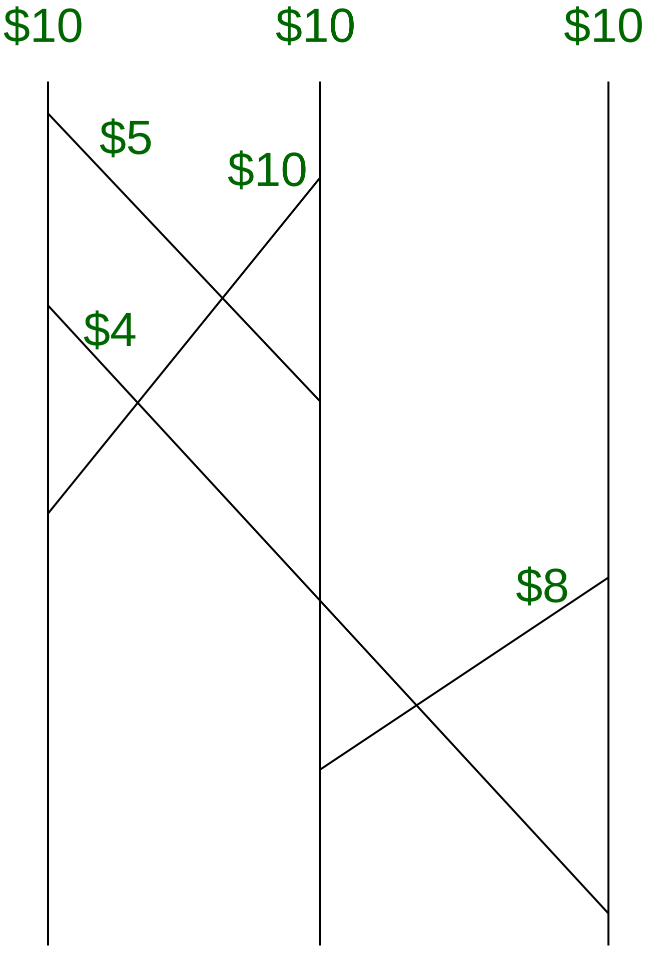


Trace

- Nodes 1,2,3 start with $10 apiece.

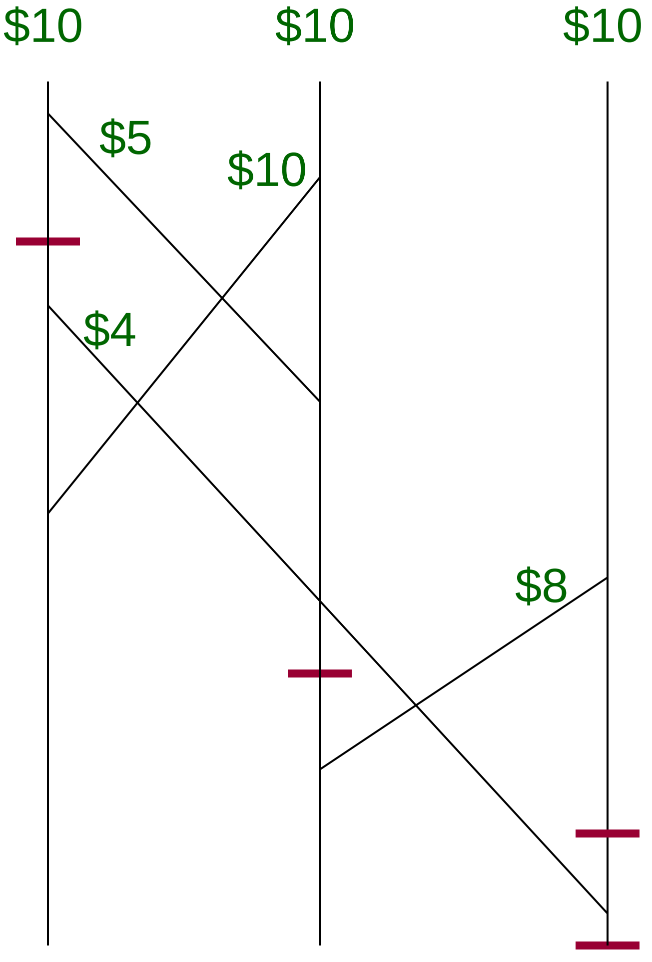
- Node 1 sends $5 to node 2.
- Node 2 sends $10 to node 1.
- Node 1 sends $4 to node 3.
- Node 2 receives $5 from node 1.
- Node 1 receives $10 from node 2.
- Node 3 sends $8 to node 2.
- Node 2 receives $8 from node 3.
- Node 3 receives $4 from node 1.

- Count the money?



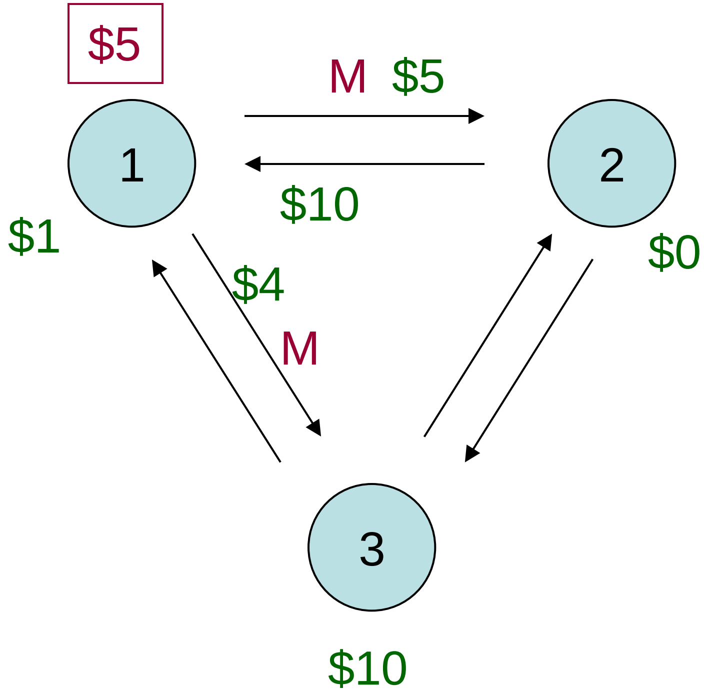
Chandy-Lamport audit

- Add snap input events:
- Q: Will local snapshots actually occur at these points?
- No, node 3 will snap before processing the $4 message, since it will receive the marker first.
- So actual local snapshot points are:



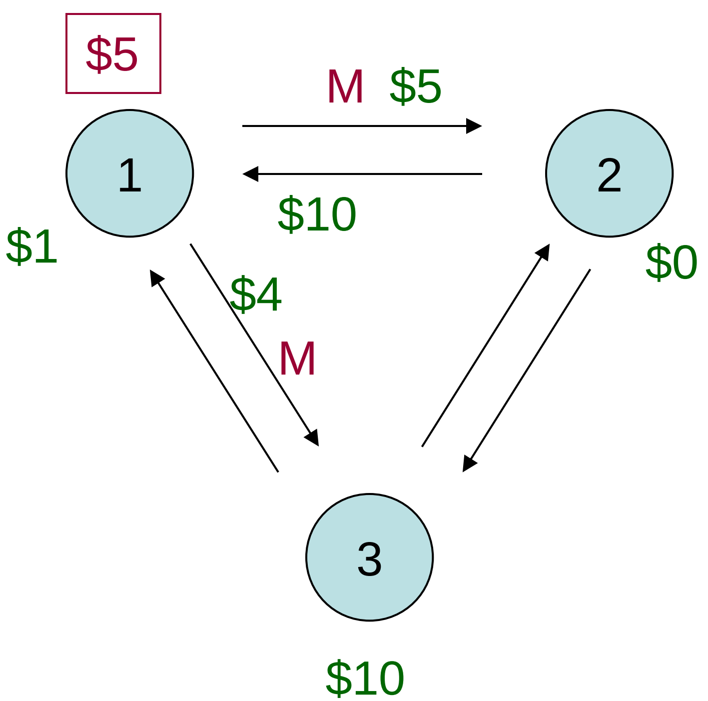
Trace, with snapshots

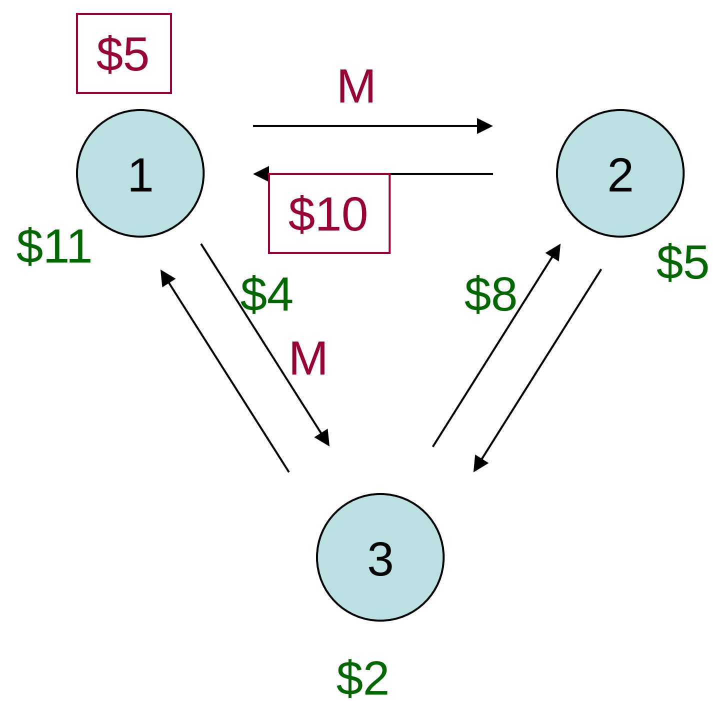
- Node 1 sends $5 to node 2.
- Node 2 sends $10 to node 1.
- Node 1 receives snap input, takes a snapshot, records state of A_1 as $5, sends markers.
- Node 1 sends $4 to node 3.



Trace, cont'd

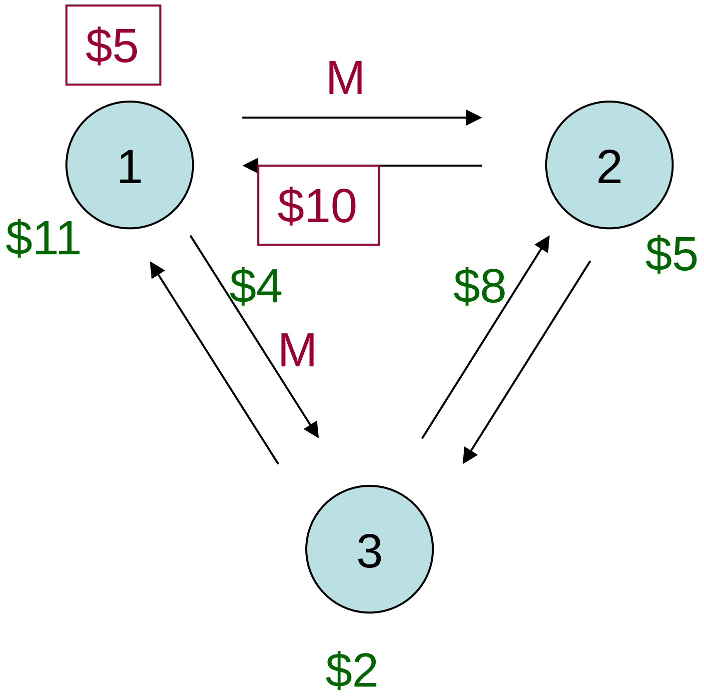
- Node 2 receives \$5 from node 1.
- Node 1 receives \$10 from node 2, accumulates it in its count for $C_{2,1}$.
- Node 3 sends \$8 to node 2.

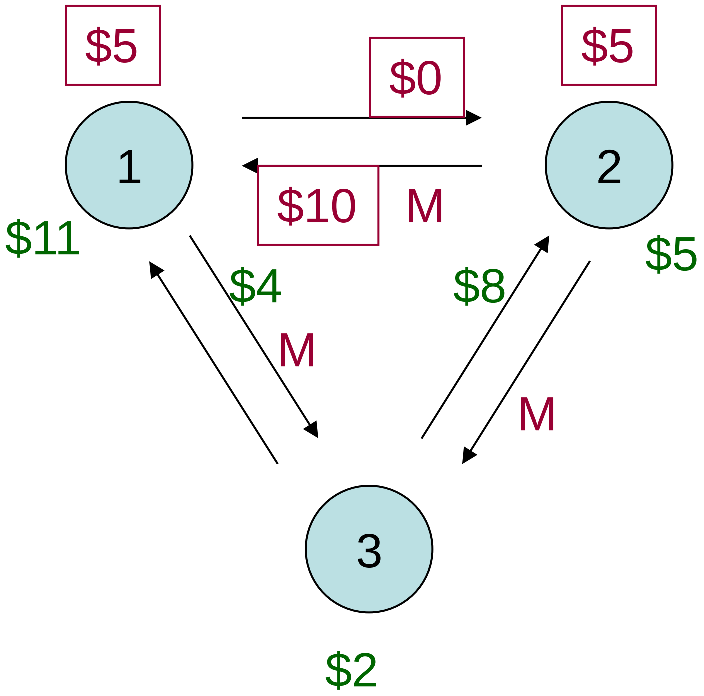




Trace, cont'd

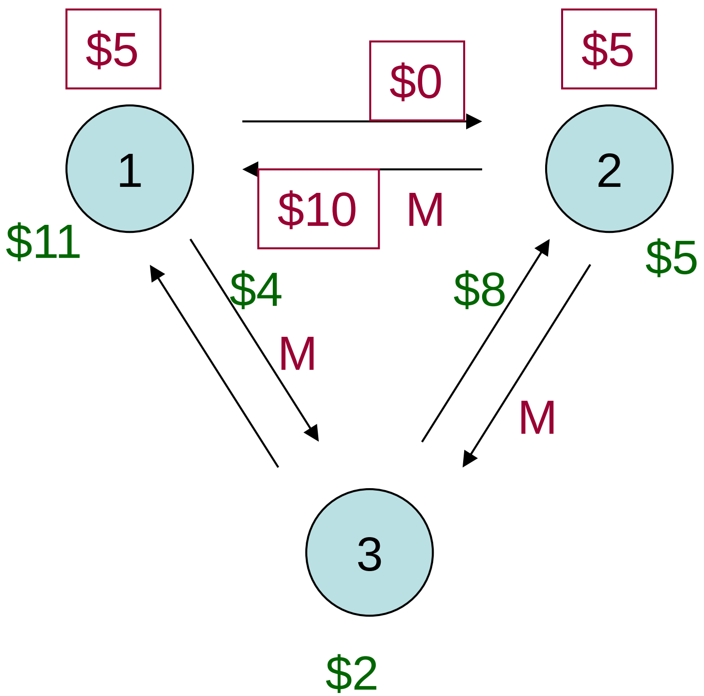
- Node 2 receives M from node 1, takes a snapshot, records state of A_2 as \$5, records state of $C_{1,2}$ as \$0, sends markers.

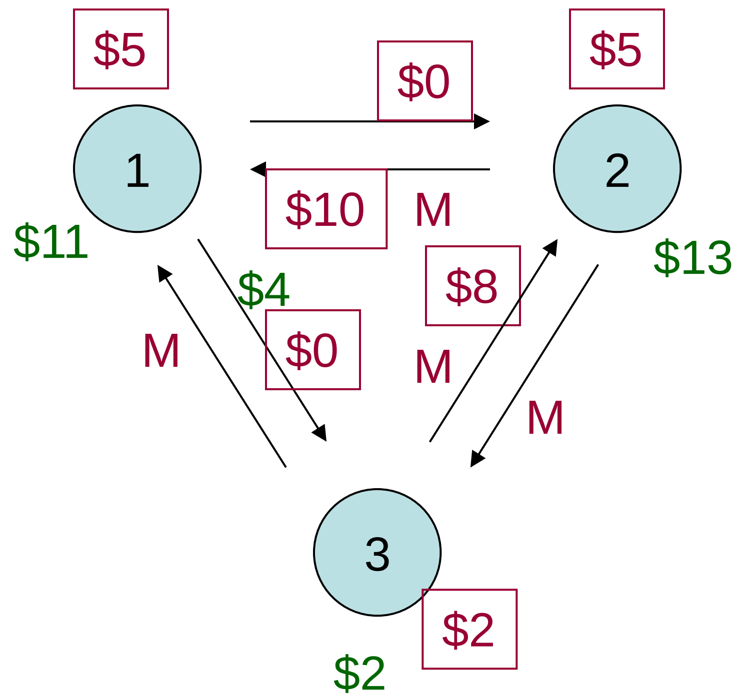




Trace, cont'd

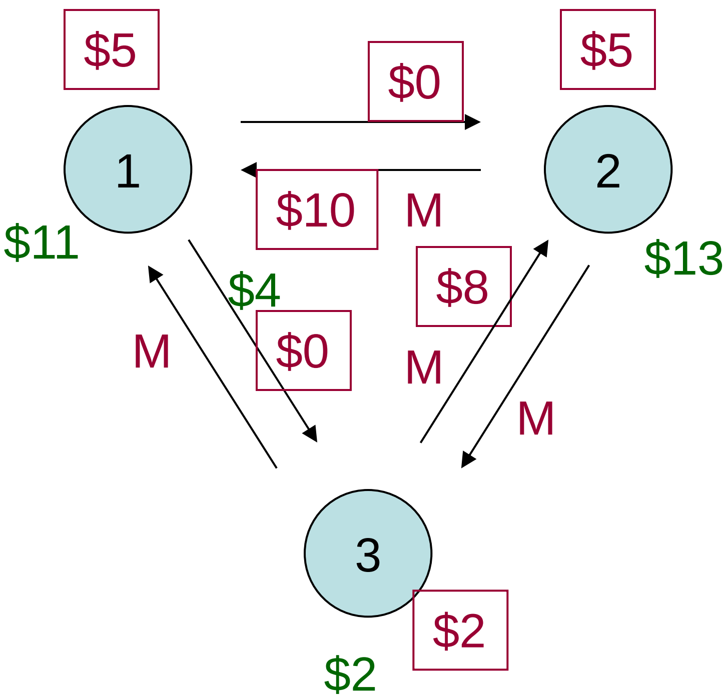
- Node 2 receives \$8 from node 3, accumulates it in its count for $C_{3,2}$.
- Node 3 receives M from node 1, takes a snapshot, records state of A_3 as \$2, records state of $C_{1,3}$ as \$0, sends markers.

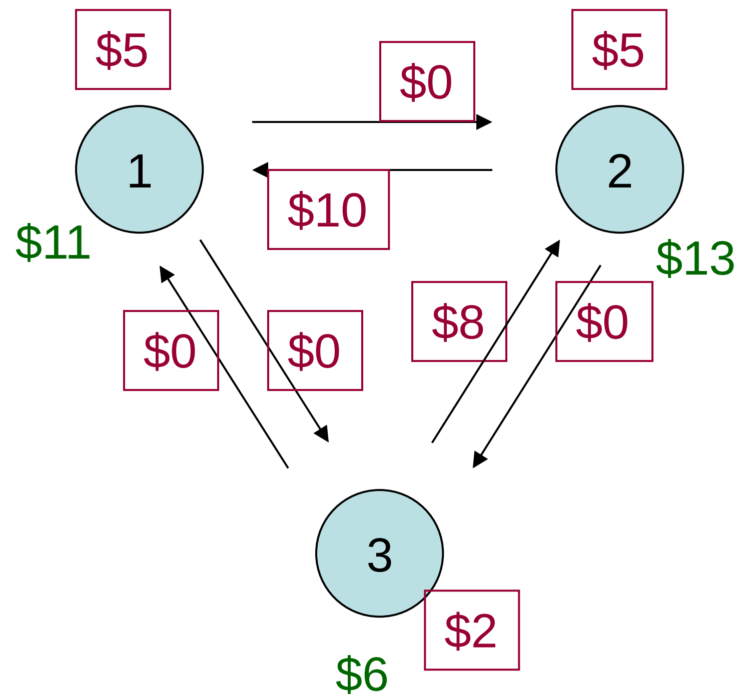




Trace, cont'd

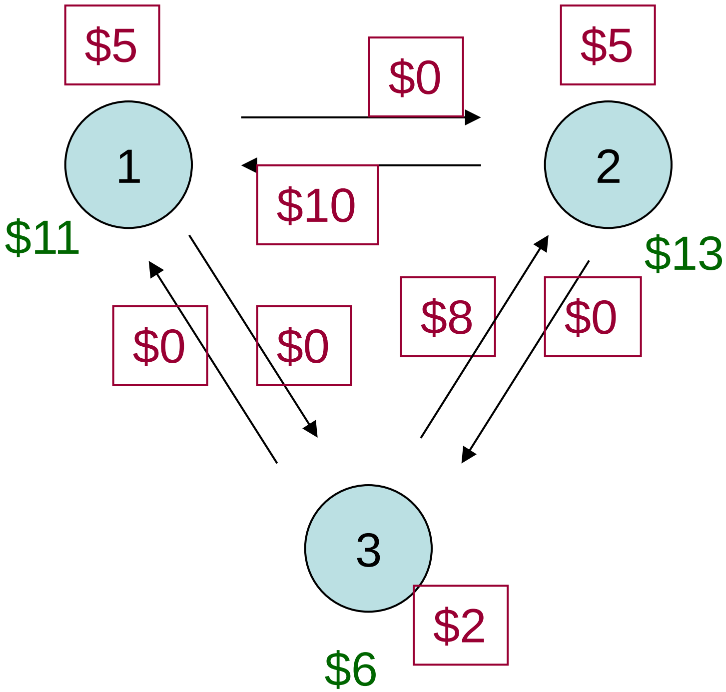
- Node 3 receives $4 from node 1, ignored by snapshot algorithm.
- Remaining markers arrive, finalizing the counts for the remaining channels.





Total amount of money

- At beginning: \$10 at each node = \$30
- At end: \$11 + \$13 + \$6 = \$30
- In the snapshot:
  - Nodes: \$5 + \$5 + \$2 = \$12
  - Channels: \$0 + \$10 + \$0
    + \$8 + \$0 + \$0 = \$18
  - Total: \$30

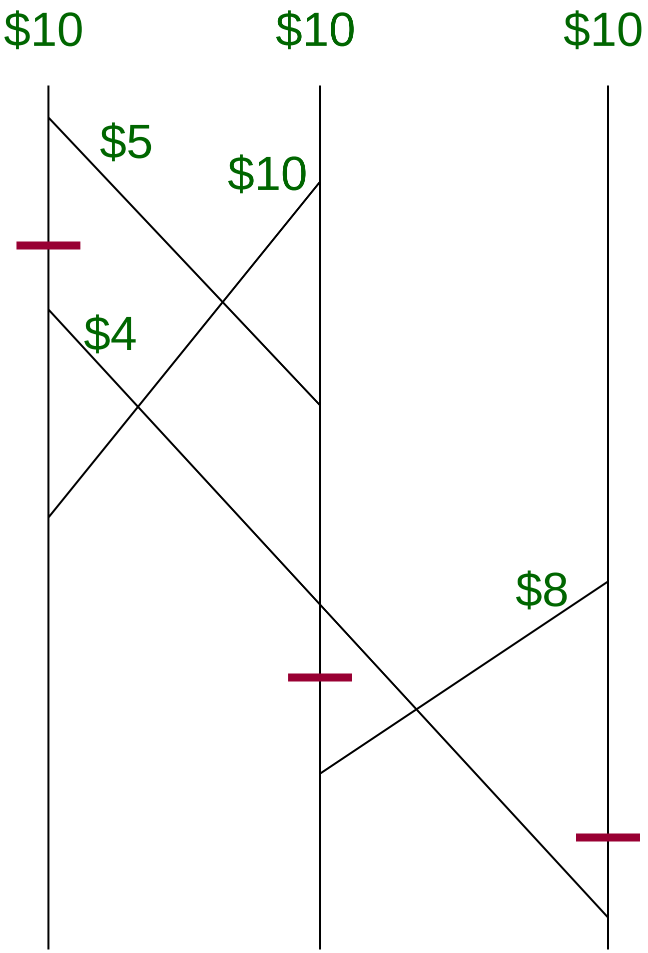


- Note:
  - The snapshot state never actually appears in the underlying execution α of the bank.
  - But it does appear in an alternative execution α' obtained by reordering events, aligning the local snapshots.

Original execution α

- Nodes 1,2,3 start with $10 apiece.

- Node 1 sends $5 to node 2.
- Node 2 sends $10 to node 1.
- Node 1 snaps.
- Node 1 sends $4 to node 3.
- Node 2 receives $5 from node 1.
- Node 1 receives $10 from node 2.
- Node 3 sends $8 to node 2.
- Node 2 snaps.
- Node 2 receives $8 from node 3.
- Node 3 snaps.
- Node 3 receives $4 from node 1.



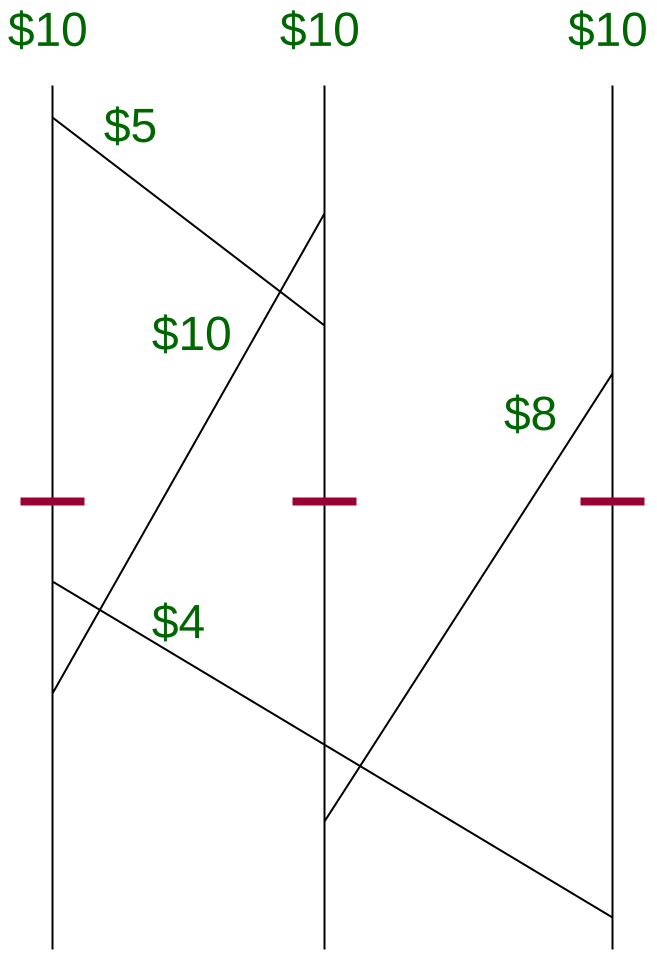
Reordered execution α'

- Nodes 1,2,3 start with $10 apiece.

- Node 1 sends $5 to node 2.
- Node 2 sends $10 to node 1.
- Node 2 receives $5 from node 1.
- Node 3 sends $8 to node 2.

- Everyone snaps.

- Node 1 sends $4 to node 3.
- Node 1 receives $10 from node 2.
- Node 2 receives $8 from node 3.
- Node 3 receives $4 from node 1.



Complexity

- Messages: O(|E|)
  - Traverse all edges, unlike [Dijkstra, Scholten]
- Time:
  - O(diam d), ignoring local processing time and pileups.

Applications of global snapshot

- Bank audit: As above.
- Checking invariants:
  - Global states returned are reachable global states, so any invariant of the algorithm should be true in these states.
  - Can take snapshot, check invariant (before trying to prove it).
- Checking requires some work:
  - Collect entire snapshot in one place and test the invariant there.
  - Or, keep the snapshot results distributed and use some distributed algorithm to check the property.
  - For "local" properties, this is easy:
    - E.g., consistency of values at neighbors: send-count_i = receive-count_j + number of messages in transit on channel from i to j.
  - For global properties, harder:
    - E.g., no global cycles in a "waits-for" graph, expressing which nodes are waiting for which other nodes.
    - Requires another distributed algorithm, for a static system.

Stable Property Detection

- Stable property:
  - A property P of a global state such that, if P ever becomes true in an execution, P remains true forever thereafter.
  - Similar to an invariant, but needn't hold in all reachable states; rather, once it's true, it remains true.
- Example: Termination
  - Assume distributed algorithm A has no external inputs, but need not start in a quiescent state.
  - Essentially, inputs appear in the initial states.
  - Terminates when:
    - All processes are in quiescent states, and
    - All channels are empty.
- Example: Deadlock
  - A set of processes are waiting for each other to do something, e.g., release a needed resource.
  - Cycle in a waits-for graph.

Snapshots and Stable Properties

- Can use [Chandy, Lamport] consistent global snapshots to detect stable properties.
- Run [CL], check whether stable property P is true in the returned snapshot state.
- Q: What does this show?
  - If P is true in the snapped state, then it is true in the real state after the final report output, and thereafter.
  - If P is false in the snapped state, then false in the real state just before the first snap input, and every state before that.
- Proof: Reachability arguments.

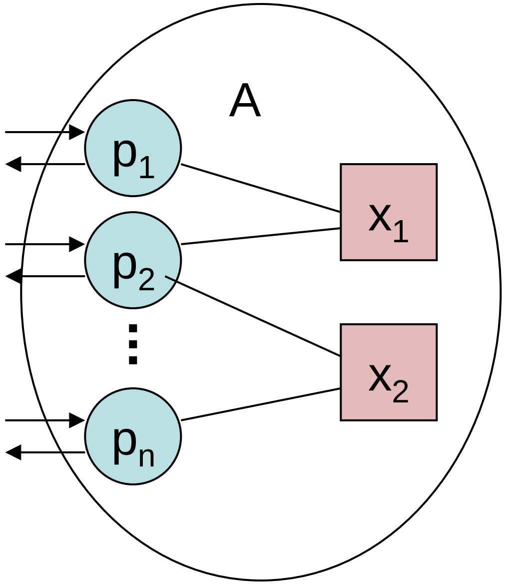
- Q: How can we be sure of detecting a stable property P, if it ever occurs?
- Keep taking snapshots.

Application: Asynchronous BFS

- Again recall AsynchBFS.
  - Allows corrections.
  - Doesn't terminate on its own; described ad hoc termination strategy earlier.
  - Diffusing algorithms, so we can apply [Dijkstra, Scholten] to get a simple terminating version.
- Alternatively, can use [Chandy, Lamport] algorithm to detect termination, using repeated snapshots.
- Eventually AsynchBFS actually terminates, and any snapshot thereafter will detect this.

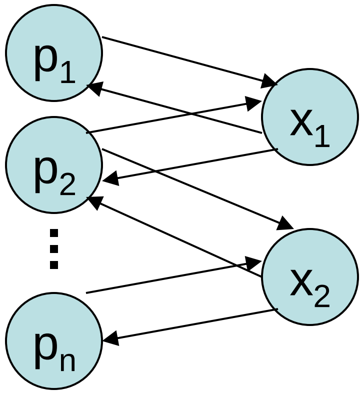
- Similarly for AsynchBellmanFord shortest paths.

Asynchronous Shared-Memory Systems



Asynchronous Shared-Memory Systems

- We’ve covered basics of non-fault-tolerant asynchronous network algorithms:
  - How to model them.
  - Basic asynchronous network protocols---broadcast, spanning trees, leader election,…
  - General methods for designing asynchronous network algorithms:
    - Synchronizers
    - Logical time
    - Global snapshots
- Now consider asynchronous shared-memory systems:
- Processes, interacting via shared objects, possibly subject to some access constraints.
- Shared objects have types, e.g.:
  - Read/write (weak)
  - Read-modify-write, compare-and-swap (strong)
  - Queues, stacks, others (in between)



Asynch Shared-Memory systems

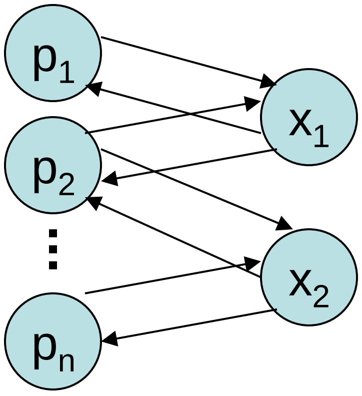
- Theory of ASM systems has much in common with theory of asynchronous networks:
  - Similar algorithms and impossibility results.
  - Even with failures.
  - Transformations from ASM model to asynch network model allow ASM algorithms to run in asynchronous networks.
    - "Distributed shared memory".
- Historically, theory for ASM started first.
- Arose in study of early operating systems, in which several processes can run on a single processor, sharing memory, with possibly-arbitrary interleavings of steps.
- Currently, ASM models are used to describe multiprocessor shared-memory systems, in which several processes can run on separate processors and share memory.

Topics

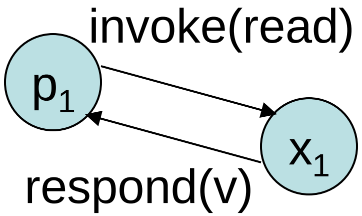
- Define the basic system model, without failures.
- Study basic problems:
  - Mutual exclusion.
  - Other resource-allocation problems.
- Then, introduce process failures into the model.
- Study more basic problems:
  - Distributed consensus
  - Implementing atomic objects:
    - Atomic snapshot objects
    - Atomic read/write registers
- Wait-free and fault-tolerant computability theory

Basic ASM Model, Version 1

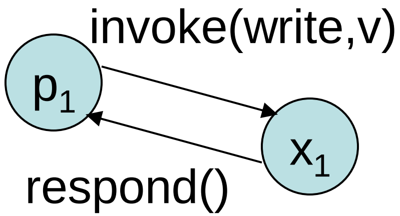
- Processes + objects, modeled as automata.
- Arrows:
  - Represent invocations and responses for operations on the objects.
  - Modeled as input and output actions.



- Fine-granularity model, can describe:
  - Delay between invocation and response.
  - Concurrent (overlapping) operations:
    - Object could reorder operations.
    - Could allow them to run concurrently, interfering with each other.

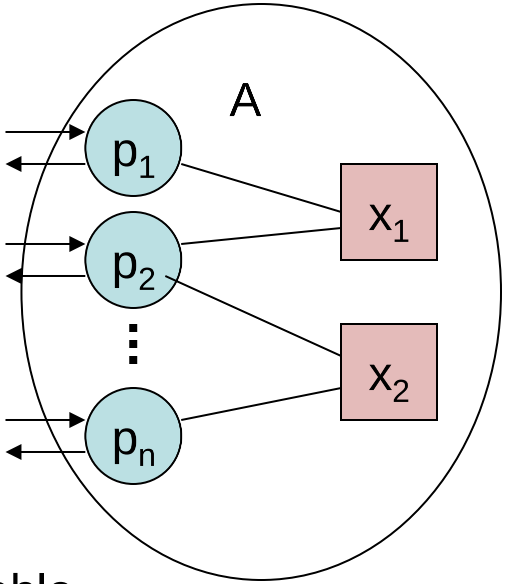


- We'll begin with a simpler, coarser model:
  - Object runs ops in invocation order, one at a time.
  - In fact, collapse each operation into a single step.
- Return to the finer model later.



Basic ASM Model, Version 2

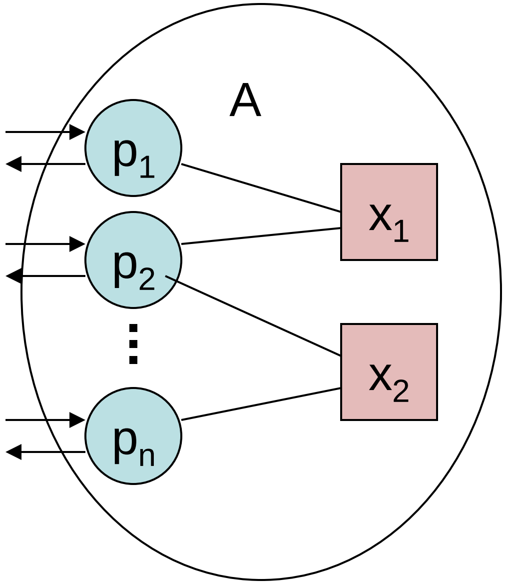
- One big shared memory system automaton A.
- External actions at process "ports".
- Each process i has:
  - A set $states_i$ of states.
  - A subset $start_{i,}$ of start states.
- Each variable x has:
  - A set $values_x$ of values it can take on.
  - A subset $initial_x$ of initial values.



- Automaton A:
  - States: State for each process, a value for each variable.
  - Start: Start states, initial values.
  - Actions: Each action associated with one process, and some also with a single shared variable.
  - Input/output actions: At the external boundary.
  - Transitions: Correspond to local process steps and variable accesses.
    - Action enabling, which variable is accessed, depend only on process state.
    - Changes to variable and process state depend also on variable value.
    - Must respect the type of the variable.
  - Tasks: One or more per process (threads).

Basic ASM Model

- Execution of A:
  - As specified by general definitions of executions, fair executions for I/O automata.
  - By fairness definition, each task gets infinitely many chances to take steps.
  - Model environment as a separate automaton, to express restrictions on environment behavior.



- Commonly-used variable types:
  - Read/write registers: Most basic primitive.
    - Allows access using separate read and write operations.
  - Read-modify-write: Most powerful primitive:
    - Atomically, read variable, do local computation, write to variable.
  - Compare-and-swap, fetch-and-add, queues, stacks, etc.
- Different computability and complexity results hold for different variable types.

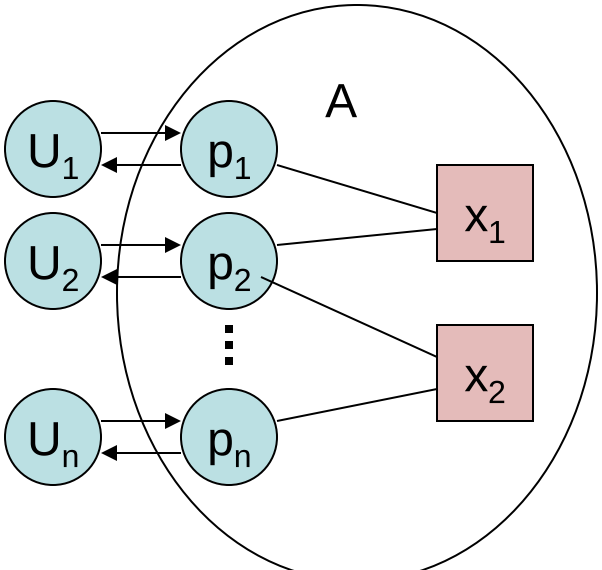
The Mutual Exclusion Problem

The Mutual Exclusion Problem

- Share one resource among n user processes, U_1, U_2,…,U_n.
  - E.g., printer, portion of a database.
- U_i has four "regions".
  - Subsets of its states, described by portions of its code.
  - C critical; R remainder; T trying; E exit

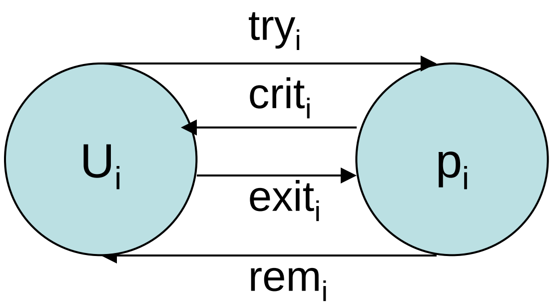
Protocols for obtaining and relinquishing the resource

- Cycle: $R \rightarrow T \rightarrow C \rightarrow E$
- Architecture:
  - U_is and A are IOAs, compose.

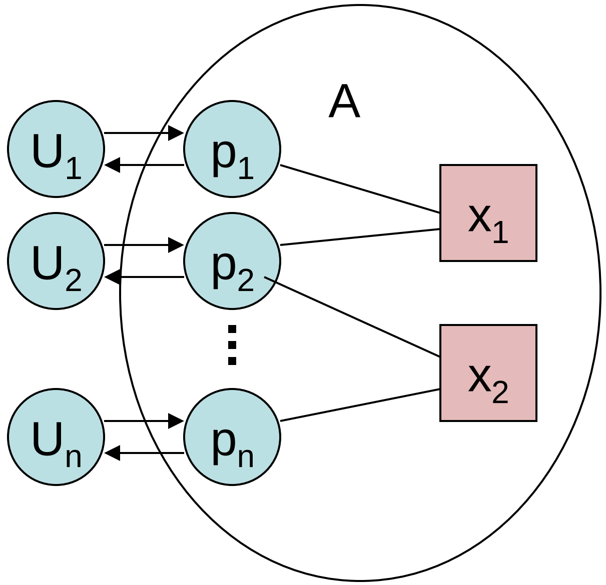


The Mutual Exclusion Problem

- Actions at user interface:
  - Connect U_i to P_i
  - p_i is U_i's "agent"

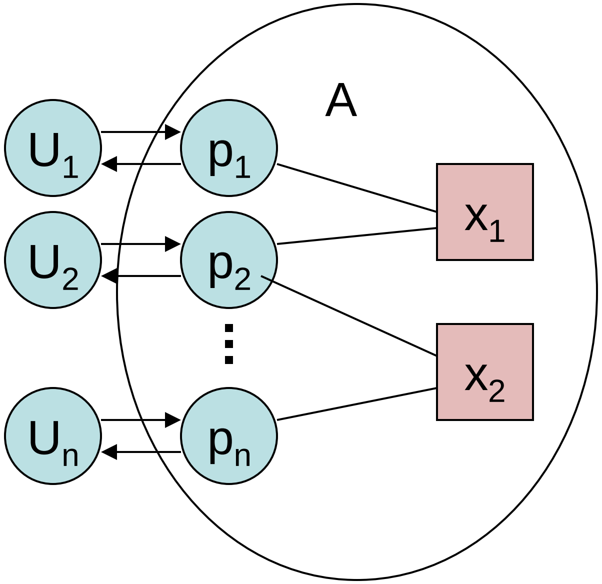


- Correctness conditions:
  - Well-formedness (Safety):
    - System also obeys cyclic discipline.
    - E.g., doesn't grant resource when it wasn't requested.
  - Mutual exclusion (Safety):
    - System never grants to > 1 user simultaneously.
    - Trace safety property.
    - Or, there's no reachable system state in which >1 user is in C at once.
  - Progress (Liveness):
    - From any point in a fair execution:
      - If some user is in T and no user is in C then at some later point, some user enters C.
      - If some user is in E then at some later point, some user enters R.



The Mutual Exclusion Problem

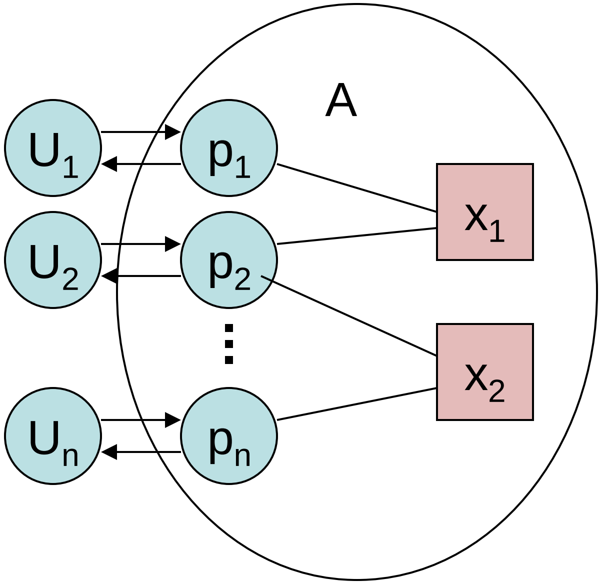
- Well-formedness (Safety):
  - System obeys cyclic discipline.
- Mutual exclusion (Safety):
  - System never grants to > 1 user.
- Progress (Liveness):
  - From any point in a fair execution:
    - If some user is in T and no user is in C then at some later point, some user enters C.
    - If some user is in E then at some later point, some user enters R.



- Conditions constrain the system automaton A, not the users.
  - System determines if/when users enter C and R.
  - Users determine if/when users enter T and E.
  - We don’t state any requirements on the users, except that users respect well-formedness.

The Mutual Exclusion Problem

- Well-formedness (Safety):
- Mutual exclusion (Safety):
- Progress (Liveness):
  - From any point in a fair execution:
    - If some user is in T and no user is in C then at some later point, some user enters C.
    - If some user is in E then at some later point, some user enters R.



- Fairness assumption:
  - Progress condition requires fairness assumption (all process tasks continue to get turns to take steps).
  - Needed to guarantee that some process enters C or R.
  - In general, in the asynchronous model, liveness properties require fairness assumptions.
  - Contrast: Well-formedness and mutual exclusion are safety properties, don't depend on fairness.

One more assumption…

- No permanently active processes.
  - Locally-controlled actions enabled only when user is in T or E.
  - No always-awake, dedicated processes.
  - Motivation:
    - Multiprocessor settings, where users can run processes at any time, but are otherwise not involved in the protocol.
    - Avoid “wasting a processor”.

Next time…

- Mutual exclusion algorithms:
  - Dijkstra’s algorithm
  - Peterson’s algorithms
  - Lamport’s Bakery Algorithm
- Reading: Sections 10.1-10.7