

Lecture 13: March 9

Lecturer: Prashant Shenoy

Scribe: Oindrila Saha

13.1 Logical and Vector Clocks

13.1.1 Recap from last lecture

For logical clocks, the problem we address is how to reason about the ordering of events in a distributed environment without the specific times at which events occurred. We can use the concept of a logical time, which depends on message exchanges among processes, to figure out the event order. A message involves two events: the sending of a message on the first process, and the receipt of the message in the second process. Since the message has to be sent before it is received, those two events are always ordered. This allows us to order two events across machines. In addition, local events within a process are also ordered based on the execution order. According to the transitive property of events ordering, events across processes are ordered.

Lamport’s logical clock is a partial ordering of events. Specifically, the events that occur after the send cannot be ordered w.r.t. another process unless there are more messages exchanged. Similarly, events that occur before the receipt of a message also cannot be ordered w.r.t. other processes unless there are other prior messages that have been received.

Another disadvantage of logical clocks is that it does not have causality. If event A has occurred before event B, then the clock value of A is less than the clock value of B. However, if a clock value is numerically less than the other, we cannot say the first event has happened before the second one. In Lamport’s clock if events are concurrent we cannot assign any ordering to them as there is no happen before relationship. Therefore, we only get a partial order using logical clocks.

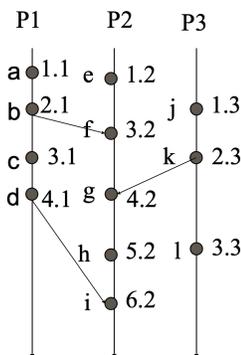


Figure 13.1: Creating total order by appending process id

13.1.2 Total Order

To convert a partial order into a total order, we can impose an arbitrary order by appending ‘.’ and a process’ id to a logical time value in each process. As a result, the process id can be used to break ties. What requires attention is that the total order is just a tie-breaking rule to assign an order for the events, so it does not actually tell us the real order of events. All the Lamport’s clock properties still hold. This is an approach that many system designers used to take Lamport’s clock which gives us a partial order and convert it into a set of events that gives us a total order.

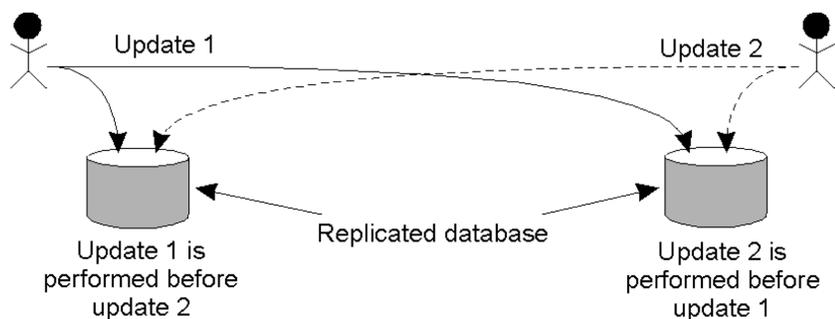


Figure 13.2: Example: Totally-Ordered Multicasting

13.1.3 Example: Totally-Ordered Multicasting

Here we use the example to show the idea of total orders. There are two replicated databases. Because they are replicated a query goes to both copies and both will update. With a partial ordering, the problem arises when there are queries that are sent in parallel. It may happen that when a user (User 1) sends a query to both replicas and is closer geographically to one of them (thus causing the query to be received sooner in the closer replica), and another user (User 2) sends another query at the same time and is closer to the other replica, the queries are executed out of order, thus producing inconsistent results.

To ensure the ordering is the same on each replica, we can use Lamport’s clock to order all the transactions by a totally-ordered multicasting approach. When a set of transactions comes in to any replica, we use a logical value to clock them similarly with the above idea of total order.

In this example, whenever a message is sent to one database, it is also sent to all the other databases. All queries and transactions are replicated. So messages are multicast to all the database replicas as shown in Figure 13.2. Like the total order example, the logical time in each machine can be appended with ‘.’ and its machine id, allowing us to break ties for the logical clock values across different machines. Thus, each database replica will respect to a consistent order.

The details of this algorithm are not expended too much in the lecture, but the totally-ordered multicasting approach can ensure each replica receives all the transactions and executes them in a consistent order.

Q: Why is there no 2.2 in Figure 13.1

A: When you send a message, you always piggyback the logical clock value. When b is sent it takes max of 1.1 and 2.1 and you get 2.1 and add 1 to get 3.1. You can skip Ids; it does not need to be sequential.

13.1.4 Causality

In Lamport’s algorithm we couldn’t say that if the clock value of A is less than B, then A has happened before B. Nothing can be said about events by comparing timestamps. In some cases, we may need to maintain causality, i.e., if A happened before B, A is causally related to B. *Causal delivery* means that if the send of a message happens before another message, the receive should also be in this order. We need a time stamping mechanism such that if $T(A) < T(B)$, A should have causally preceded (happened before) B.

13.1.5 Vector Clocks

Causality can be captured by means of **vector clocks**. Instead of just one integer, every process maintains a vector as long as the number of processes, so each process i has a vector V_i . $V_i[i]$ is the number of event that have occurred at i , which is effectively its Lamport’s clock. $V_i[j]$ is the number of events i knows have occurred at process j . Vector clocks are updated as follows in each of the below situations:

- Local event: increment $V_i[i]$
- Send a message: piggyback entire vector V
- Receipt of a message: $V_j[k] = \max(V_j[k], V_i[k])$
- Receiver is told about how many events the sender knows occurred at another process k
- Also $V_j[j] = V_j[j] + 1$

The idea of how vector clocks works is shown in Figure 13.3. Assume we have processes P_1, P_2 and P_3 . Each process keeps an integer logical clock and the logical clock will be ticked whenever a local event happens. However, rather than keeping a single clock value like before, each process maintains a vector of clock value, one for each process in the system. Essentially, all clocks get initialized to $(0,0,0)$. The three elements are for process P_1, P_2 , and P_3 respectively. Every time an event occurs in process i (i can be P_1, P_2 or P_3), the i -th element of process i ’s vector gets incremented. E.g., process P_1 get its logical clocks in the first element of its vector clock incremented in the first events. When a message is sent by process i , the vector clock is piggybacked onto that message. When a process j received the message, the i -th element will be the maximum of i -th element of vector clock i and local vector clock j , and the j -th element will get incremented by one.

Two examples are shown in Figure 13.3 when a message is passed from P_1 to P_2 and V_2 is updated to $(1,2,0)$ and then when a message is passed from P_2 to P_3 and V_3 is updated to $(1,3,4)$. Through sequence of messages each vector clock knows something about the other processes either directly or indirectly.

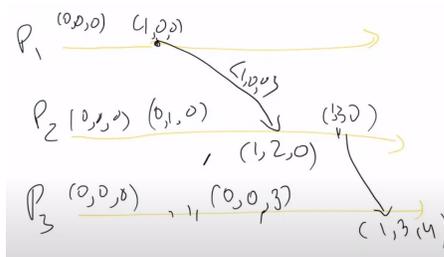


Figure 13.3: Vector Clocks

To check if there is a happen before relationship maintains lets take the send (S) from P_1 and the receive at P_3 (R). Clearly, S has happened before R. We need to have $(1,3,4) > (1,0,0)$. We define greater between vectors as $V1 > V2$ if every element of V1 is \geq every element of V2 and there is at least one element that is strictly greater. This is true in the case of S and M. We cannot compare $(1,4,0)$ and $(1,3,4)$. That means they are concurrent events. Vector clocks claims the causality property.

It is worth noting that, unlike the logical clocks which are all comparable integers, vector clocks give undefined order relationship to two concurrent or independent events, which respects the reality and causality.

13.1.6 Enforcing Causal Communication

Using vector clocks, it is now possible to ensure that a message is delivered only if all messages that causally precede it have also been received as well. To enable such a scheme, we will assume that messages are multicast within a group of processes. As an example, consider three processes P_0 , P_1 , and P_2 as shown in Figure 13.4. At local time $(1,0,0)$, P_0 sends message m to the other two processes. After its receipt by P_1 , the latter decides to send m^* , which arrives at P_2 sooner than m . At that point, the delivery of m^* is delayed by P_2 until m has been received and delivered to P_2 's application layer.

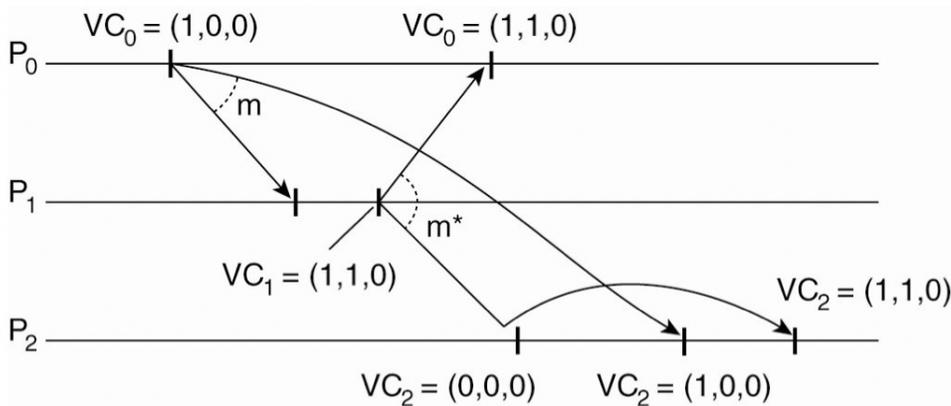


Figure 13.4: Enforcing causal communication

Q: Why is there VC1 not $(1,2,0)$ in Figure 13.4?

A: Using the algorithm defined in this lecture it should be $(1,2,0)$.

13.2 Global States and Distributed Snapshots

13.2.1 Global State

Problem definition: We want to run a distributed application where one of n processes crashes. Rather than killing all the processes and starting from the beginning, we can periodically take snapshots (or checkpoints) of the distributed application to keep a global state and start from the latest snapshot.

In a distributed system, when there are n processes communicating with each other, taking a checkpoint is harder. We need to take the snapshot in a consistent manner.

The global state includes the local state of each process and messages that are in transit (like the TCP buffers). The snapshot for a global state should be captured in a consistent fashion even without clock synchronization. A “consistent fashion” means that whenever restarting the computation from a checkpoint, you should get the same end result as if there was no cache at all. We will eliminate the notion of a clock and derive a technique independent of clock synchronization.

Specifically, when a message exchanging crosses the snapshot taking, the sending point of the message instead of the receipt should be captured in the state. As shown in Figure 13.5, a consistent cut (a) can achieve the consistent state while a inconsistent cut (b) can not. For (b), if you restart the computation from whenever the dotted line hits each of the processes, m2 received by P3 can be inconsistent – P3 already saw the m2 before the snapshot and it will see m2 again after re-computation from the snapshot and P2 resent it. the send should be to the left of the cut and the receive to the right.

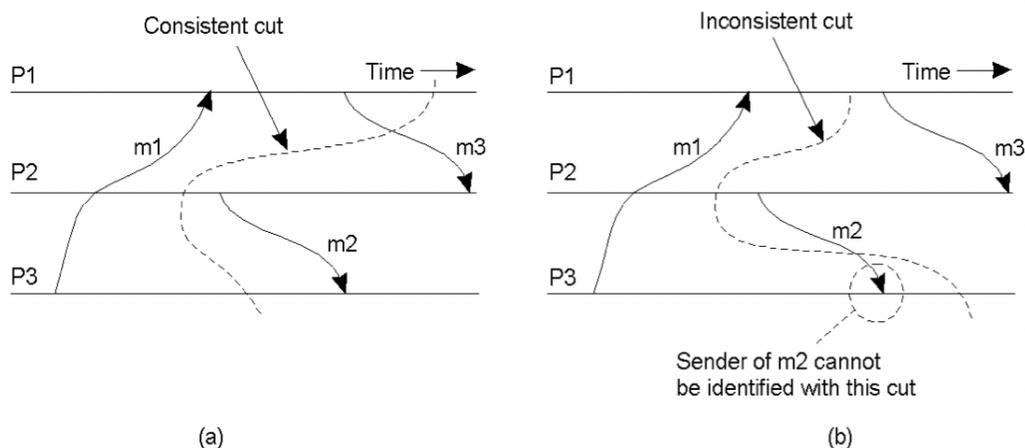


Figure 13.5: Consistent and Inconsistent cuts

13.2.2 Distributed Snapshot

A simple technique to capture is to assume a global synchronized clock and freeze all machines at the same time and capture whatever that are in all memories at the exactly same time and capture everything that is in transit. But there is no global clock synchronization. Distributed snapshotting is a mechanism that gives a consistent global state without a global clock synchronization.

The Photograph Example: Imagine a photographer taking a picture of the sky with birds flying around in it. They cannot capture entire sky in one picture, so instead they stitch pictures of the left, middle, and right parts of sky. If they take a picture of the left, and then take a picture of the middle, but a bird has flown from left to the middle, they have taken a picture of that bird twice because of inconsistency. The goal is to get a consistent photo—for it to be consistent, then all three pictures have captured all of the birds that were in the sky exactly once.

Photographs are like snapshots (in fact, this is where the name comes from). Birds are messages that are going from one process to another. You do not want a bird missing or duplicated in your picture just like you do not want a message lost or double counted.

13.2.3 Distributed Snapshot Algorithm

Assume each process communicates with another process using unidirectional point-to-point channels (e.g., TCP connections). Any process can initiate the snapshot algorithm. When a process initiates the algorithm, it will first checkpoint its local state, and then send a marker on every outgoing channel. When a process receives a marker, it will have different actions depending on whether it has already seen the marker for this snapshot. If the process sees a marker at the first time, it will checkpoint its state, send markers out and start saving messages on all the other channels. If the process sees a marker for the snapshot at the second time, it will stop saving messages for the channel from which the marker comes. A process will finish a snapshot until it receives a marker on each incoming channel and processes them all.

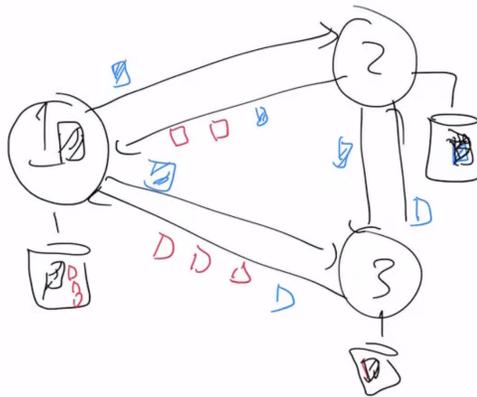


Figure 13.6: Distributed Snapshot Example

For example, in Figure 13.6, process 1, 2 and 3 communicates with each other through the duplex TCP connections. When process 1 initiate a snapshot, it will first save its memory contents (state) into the disk, and then send a marker (in blue) to 2 and 3. It will also start to save the incoming messages (in red) (TCP buffers) from 2 and 3. When 2 receives a marker from 1, it will start to checkpoint state, send a marker out to 1 and 3, and start saving messages from 3. Assume 3 sees the marker sent from 1 first, it will checkpoint its state, sends out a marker to 1 and 2, and start saving messages from 2. 1 will stop saving messages for 2 or 3 until a marker from 2 or 3 arrives. 2 will stop saving messages for 3 until a marker from 2 arrives. And P3 will stop saving messages for 2 until a marker from 3 arrives. Each process will finish this snapshot when it sees a marker from every incoming channel. Thus, a distributed snapshot captured.

13.2.4 Snapshot Algorithm Example

Q decides to take a snapshot. Consider Figure 13.7. b) It receives a marker for the first time and records its local state; c) Q records all incoming message; d) Q receives a marker for its incoming channel and finishes recording the state of the incoming channel. First incoming marker says start recording and all other incoming messages say stop recording.

Q: Is it guaranteed to be correct?

A: It is guaranteed to be correct under the assumption that all channels are TCP/IP sockets and we are delivering messages in order. It is guaranteed to terminate as well.

Q: In a distributed snapshot can you have a case where a send is before the cut and the receive is after?

A: In Fig this the case where you will record m_3 . It will keep recording incoming channels and will save m_3 as part of its disk. P2 will process m_3 and snapshot will save a copy on disk too.

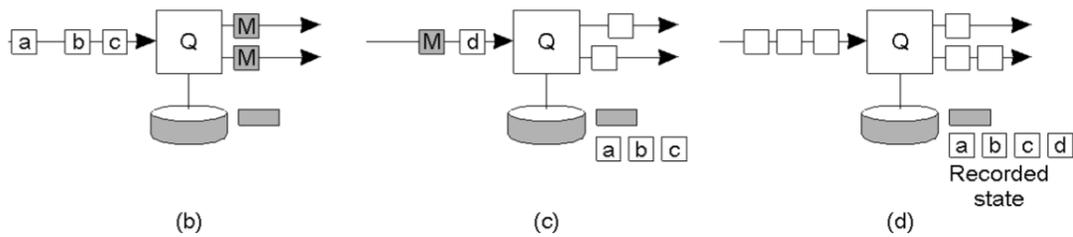


Figure 13.7: Snapshot algorithm example.

Q: Can the process not send new messages before you receive?

A: There is no restriction on what the process can do when a snapshot is happening.

13.3 Termination Detection

This involves detecting the end of a distributed computation. We need to detect this since a process cannot just exit after completing its events if another process wants to send it a message. Let sender be the predecessor, and receiver be the successor. There are two types of markers: Done and Continue. After finishing its part of the snapshot, process Q sends a Done or a Continue to its predecessor. Q can send a Done only when:

- All of Q's successors send a Done
- Q has not received any message since it check-pointed its local state and received a marker on all incoming channels
- Else send a Continue

Computation has terminated if the initiator receives Done messages from everyone.

Q: What happens when there is a failure?

A: Terminal detection algorithm does not handle failures. We need to go previous snapshot if there is a crash.