Synchronization & Coordination

Yao Liu
Clock synchronization

• When each machine has its own clock, an event that occurred after another event may nevertheless be assigned an earlier time.
Skew between computer clocks in a distributed system
Clock synchronization algorithms

- The relation between clock time and UTC when clocks tick at different rates.

\[
\frac{dC}{dt} > 1 \\
\frac{dC}{dt} = 1 \\
\frac{dC}{dt} < 1
\]

Clock time, C

UTC, t
Clock synchronization

- Physical clocks drift, therefore need for clock synchronization algorithms
  - Many algorithms depend upon clock synchronization
  - Clock synch. Algorithms – Cristian, NTP, Berkeley algorithm, etc.

- However, since we cannot perfectly synchronize clocks across computers, we cannot use physical time to order events
Clock synchronization using a time server

\[ p \leftrightarrow m_r \leftrightarrow m_t \rightarrow \text{Time server, } S \]
Cristian's algorithm

- Getting the current time from a time server.

Both $T_0$ and $T_1$ are measured with the same clock

I, Interrupt handling time
Clock synchronization algorithms

• Cristian’s algorithm
  • p should set its time to $t + \frac{T_{\text{round}}}{2}$
  • If the value of the minimum transmission time $\text{min}$ is known or can be conservatively estimated
    • Earliest time at which S could have placed its time in $m_t$ was $\text{min}$ after p dispatched $m_r$
    • Latest point at which it could do so was $\text{min}$ before $m_t$ arrived at p
    • Time by S’s clock when message arrives at p is in range $[t + \text{min}, t + T_{\text{round}} - \text{min}]$
    • Accuracy $\pm (T_{\text{round}}/2 - \text{min})$
The Berkeley algorithm

a) The time daemon asks all the other machines for their clock values
b) The machines answer
c) The time daemon tells everyone how to adjust their clock
An example synchronization subnet in an NTP implementation

Note: Arrows denote synchronization control, numbers denote strata.
Logical time & clocks

- Lamport proposed using logical clocks based upon the “happened before” relation
  - If two events occur at the same process, then they occurred in the order observed
  - Whenever a message is sent between processes, the event of sending occurred before the event of receiving
  - X happened before Y denoted by $X \rightarrow Y$
Events occurring at three processes
Lamport’s algorithm

• Each process has its own logical clock
• LC1: $C_p$ is incremented before each event at process $p$
• LC2:
  • When process $p$ sends a message it piggybacks on it the value $C_p$
  • On receiving a message $(m, t)$, a process $q$ computes $C_q = \max(C_q, t)$ and then applies LC1 before timestamping the receive event
Lamport timestamps for the events

Physical time

```
p1
  1 a
  2 b
  m1

p2
  3 c
  4 d
  m2

p3
  1 e
  5 f
```
Vector timestamps

- Shortcoming of Lamport’s clocks: if \( L(e) < L(f) \), we cannot conclude that \( e \rightarrow f \)

- Vector clocks
  - A process keeps an array of clocks, one for each process
  - Like Lamport timestamps, processes piggyback vector timestamps on messages they send each other

\[
\begin{align*}
V &= W \text{ iff } V[j] = W[j] \text{ for } j = 1, 2, \ldots, N \\
V &\leq W \text{ iff } V[j] \leq W[j] \text{ for } j = 1, 2, \ldots, N \\
V &< W \text{ if } V \leq W \text{ and } V \neq W
\end{align*}
\]
Vector timestamps for the events

- $p_1$: (1,0,0) (2,0,0)
- $p_2$: (2,1,0) (2,2,0)
- $p_3$: (0,0,1) (2,2,2)

Physical time
Totally ordered logical clocks

• Logical clocks only impose partial ordering
• For total order, use \((T_a, P_a)\) where \(P_a\) is process id
• \((T_a, P_a) < (T_b, P_b)\) if and only if either \(T_a < T_b\) or \(T_a = T_b\) and \(P_a < P_b\)
• This ordering has no physical significance, but it is sometime useful, e.g., to break a tie between two processes trying to enter a critical section
Example: totally-ordered multicasting

- Updating a replicated database and leaving it in an inconsistent state.
  - $P1$ adds $100$ to an account (initial value: $1000$)
  - $P2$ increments account by $1\%$
  - There are two replicas

**Diagram:**
- Update 1 is performed before update 2
- Replicated database
- Update 2 is performed before update 1
Example: totally-ordered multicasting

• Receiver multicasts an ack to all other processes
• According to Lamport’s algorithm
  • timestamp on message is less than timestamp on ack
• Applications receive queued messages
  • only if it is in the head of the queue
  • has been acknowledged by every other process
• Ultimately
  • All processes eventually will have the same copy of the local queue!
Distributed mutual exclusion

- Central server algorithm
- Ricart and Agrawala algorithm
  - A distributed algorithm that uses logical clocks
- Ring-based algorithms

NOTE: the above algorithms are not fault-tolerant and not very practical. However, they illustrate issues in the design of distributed algorithms

- Several other mutual exclusion algorithms have been proposed
  - Quorum consensus algorithms – Maekawa’s algorithm
  - We will discuss majority voting in the context of replicated data management
Server managing a mutual exclusion token for a set of processes

1. Request token
2. Release token
3. Grant token

Queue of requests

$p_1$

$p_2$

$p_3$

$p_4$
On initialization
\[ \text{state} := \text{RELEASED}; \]

To enter the section
\[ \text{state} := \text{WANTED}; \]
Multicast request to all processes;
\[ T := \text{request’s timestamp}; \]
Wait until (number of replies received = \((N - 1)\));
\[ \text{state} := \text{HELD}; \]

On receipt of a request \(<T_i, p_i>\) at \(p_j\) \((i \neq j)\)
\[ \text{if} \ ( \text{state} = \text{HELD} \ \text{or} \ ( \text{state} = \text{WANTED} \ \text{and} \ (T_j, p_j) < (T_i, p_i)) \)
\]
then
\[ \text{queue request from} \ p_i \ \text{without replying}; \]
else
\[ \text{reply immediately to} \ p_i; \]
end if

To exit the critical section
\[ \text{state} := \text{RELEASED}; \]
reply to any queued requests;
Multicast synchronization
A ring of processes transferring a mutual exclusion token
Comparison

- A comparison of three mutual exclusion algorithms.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Messages per entry/exit</th>
<th>Delay before entry (in message times)</th>
<th>Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centralized</td>
<td>3</td>
<td>2</td>
<td>Coordinator crash</td>
</tr>
<tr>
<td>Distributed</td>
<td>$2(n - 1)$</td>
<td>$2(n - 1)$</td>
<td>Crash of any process</td>
</tr>
<tr>
<td>Token ring</td>
<td>1 to $\infty$</td>
<td>0 to $n - 1$</td>
<td>Lost token, process crash</td>
</tr>
</tbody>
</table>
Maekawa’s algorithm

- Every node needs permission from other nodes in its quorum before it enters critical section
- Quorums are constructed in such a way that no two nodes can be in their critical section at the same time
- The size of each node’s quorum is $O(\sqrt{N})$, which can be shown to be optimal
Construction of quorum sets

Consider a system with 9 nodes

The quorum for any node includes the nodes in its row and column

Quorum for Node 1 = \{1,2,3,4,7\}
Quorum for Node 9 = \{3,6,7,8,9\}

There is a non-null intersection for the quorums of any two nodes
Maekawa’s algorithm

On initialization

\[ \text{state} := \text{RELEASED}; \]
\[ \text{voted} := \text{FALSE}; \]

For \( p_i \) to enter the critical section

\[ \text{state} := \text{WANTED}; \]
Multicast \textit{request} to all processes in \( V_i - \{p_i\} \);
Wait until (number of replies received = \( (K - 1) \));
\[ \text{state} := \text{HELD}; \]

On receipt of a request from \( p_i \) at \( p_j \) (\( i \neq j \))

\[ \text{if} \ (\text{state} = \text{HELD} \text{ or voted} = \text{TRUE}) \]
\[ \text{then} \]
\[ \quad \text{queue request from} \ p_i \ \text{without replying}; \]
\[ \text{else} \]
\[ \quad \text{send reply to} \ p_i; \]
\[ \quad \text{voted} := \text{TRUE}; \]
\[ \text{end if} \]
Maekawa’s algorithm – cont’d

For \( p_i \) to exit the critical section

\[
\text{state} := \text{RELEASED};
\]

Multicast \text{release} to all processes in \( V_i - \{p_i\} \);

On receipt of a release from \( p_i \) at \( p_j \) (\( i \neq j \))

\[
\text{if (queue of requests is non-empty)}
\]

\[
\text{then}
\]

\[
\text{remove head of queue – from } p_k, \text{ say;}
\]

\[
\text{send reply to } p_k;
\]

\[
\text{voted} := \text{TRUE};
\]

\[
\text{else}
\]

\[
\text{voted} := \text{FALSE};
\]

\[
\text{end if}
\]
Election algorithms

• An election is a procedure carried out to choose a process from a group, for example to take over the role of a process that has failed

• Main requirement: elected process should be unique even if several processes start an election simultaneously

• Algorithms:
  
  • Bully algorithm: assumes all processes know the identities and addresses of all the other processes
  
  • Ring-based election: processes need to know only addresses of their immediate neighbors
The bully algorithm

The election of coordinator $p_2$, after the failure of $p_4$ and then $p_3$
A ring-based election in progress

Note: The election was started by process 17.
The highest process identifier encountered so far is 24.
Participant processes are shown darkened.
Detecting global properties

a. Garbage collection

b. Deadlock

c. Termination
Global states and consistent cuts

- Capturing a global state would be straightforward if we had perfectly synchronized clocks.
- How to capture a meaningful global state from local states recorded at different real times?
- Each process records events that correspond to internal actions, e.g., updating a variable, and the sending or receipt of a message.
- A cut is a subset of the system’s global history that is a union of prefixes of process histories.
- A cut is **consistent** if for each event it contains, it also contains all events that happened before that event.
Consistent and inconsistent cuts

Physical time

Consistent cut

Inconsistent cut
Chandy and Lamport’s snapshot algorithm

- **Goal**: record a set of process and channel states for a set of processes such that even if the combination of recorded states may never have occurred at the same time, the recorded state is consistent
- **State recorded locally at processes**
- **Assumptions**
  - neither channels nor processes fail; communication is reliable
  - channels are unidirectional and provide FIFO message delivery
  - the graph of processes and channels is strongly connected
  - any process may initiate a global snapshot at any time
  - processes may continue with their execution and send and receive normal messages while the snapshot takes place
Chandy and Lamport algorithm

- Organization of a process and channels for a distributed snapshot
Chandy and Lamport’s “snapshot” algorithm

*Marker receiving rule for process $p_i$*

On $p_i$’s receipt of a *marker* message over channel $c$:

- *if* ($p_i$ has not yet recorded its state) it
  - records its process state now;
  - records the state of $c$ as the empty set;
  - turns on recording of messages arriving over other incoming channels;
- *else*
  - $p_i$ records the state of $c$ as the set of messages it has received over $c$ since it saved its state.

*Marker sending rule for process $p_i$*

After $p_i$ has recorded its state, for each outgoing channel $c$:

- $p_i$ sends one marker message over $c$
- (before it sends any other message over $c$).
Chandy and Lamport algorithm

b) Process Q initiates snapshot and records its local state

c) Q records all incoming messages

d) Q receives a marker for its incoming channel and finishes recording the state of the incoming channel
Two processes and their initial states

- Assume that $p_2$ has already received an order for five widgets, which it will shortly dispatch to $p_1$.
Example: Chandy & Lamport’s algorithm

1. Global state $S_0$
   - Global state $S_0$
   - $p_1$: <$1000, 0$> (empty) $c_2$
   - $p_2$: <$50, 2000$> (empty) $c_1$

2. Global state $S_1$
   - Global state $S_1$
   - $p_1$: <$900, 0$> (empty) $c_2$
   - $p_2$: <$50, 2000$> (empty) $c_1$
   - (Order 10, $100$), M

3. Global state $S_2$
   - Global state $S_2$
   - $p_1$: <$900, 0$> (five widgets) $c_2$
   - $p_2$: <$50, 1995$> (empty) $c_1$
   - (Order 10, $100$), M

4. Global state $S_3$
   - Global state $S_3$
   - $p_1$: <$900, 5$> (empty) $c_2$
   - $p_2$: <$50, 1995$> (empty) $c_1$
   - (Order 10, $100$)
   - (M = marker message)

Final recorded state $p_1$: <$1000, 0$>; $p_2$: <$50, 1995$>, $c_1$: <five widgets>; $c_2$: <>
Reading

- Chapter 6 of Tbook, Chapter 14 of Cbook
- Papers on course homepage