Atomic Commit

The objective
Preserve data consistency for distributed transactions in the presence of failures

Model

- For each distributed transaction $T$:
  - one coordinator
  - a set of participants
- Coordinator knows participants; participants don’t necessarily know each other
- Each process has access to a Distributed Transaction Log (DT Log) on stable storage

The setup

- Each process $p_i$ has an input value $vote_i$:
  $vote_i \in \{Yes, No\}$
- Each process $p_i$ has output value $decision_i$:
  $decision_i \in \{Commit, Abort\}$
AC Specification

AC-1: All processes that reach a decision reach the same one.
AC-2: A process cannot reverse its decision after it has reached one.
AC-3: The Commit decision can only be reached if all processes vote Yes.
AC-4: If there are no failures and all processes vote Yes, then the decision will be Commit.
AC-5: If all failures are repaired and there are no more failures, then all processes will eventually decide.

Comments

AC1: We do not require all processes to reach a decision.
AC-2: A process cannot reverse its decision after it has reached one.
AC-3: The Commit decision can only be reached if all processes vote Yes.
AC-4: If there are no failures and all processes vote Yes, then the decision will be Commit.
AC-5: If all failures are reported and there are no more failures, then all processes will eventually decide.

Liveness & Uncertainty

A process is uncertain if it has voted Yes but does not have sufficient information to commit.
While uncertain, a process cannot decide unilaterally.
Uncertainty + communication failures = blocking!

Liveness & Independent Recovery

Suppose process $p$ fails while running AC.
If, during recovery, $p$ can reach a decision without communicating with other processes, we say that $p$ can independently recover.
Total failure (i.e. all processes fail) - independent recovery = blocking.
A few character-building facts

Proposition 1
If communication failures or total failures are possible, then every AC protocol may cause processes to become blocked.

Proposition 2
No AC protocol can guarantee independent recovery of failed processes.

2-Phase Commit

Coordinator $c$

1. sends VOTE-REQ to all participants

Participant $p_i$

II. sends vote$_i$ to Coordinator

- if vote$_i = NO$ then
  - decide$_i := ABORT$
  - halt

III. if (all votes YES) then
  - decide$_c := COMMIT$
  - send COMMIT to all

else
  - decide$_c := ABORT$
  - send ABORT to all who voted YES
  - halt
2-Phase Commit

Coordinator C
I. sends VOTE-REQ to all participants

Participant \(p_i\)
II. sends vote, to Coordinator if vote = NO then
   decide \(_i\) := ABORT
   halt
else
   decide \(_i\) := COMMIT
   send COMMIT to all
   halt

III. if (all votes YES) then
   decide := COMMIT
   send COMMIT to all
else
   decide := ABORT
   send ABORT to all who voted YES
   halt

IV. if received COMMIT then
   := COMMIT
else
   := ABORT      !
halt

Notes on 2PC

Satisfies AC-1 to AC-4

But not AC-5 (at least “as is”)
   i. A process may be waiting for a message that may never arrive
   □ Use Timeout Actions
   □ Processes save protocol state in DT-Log

Timeout actions

Processes are waiting on steps 2, 3, and 4

Step 2 \(p_i\) is waiting for VOTE-REQ from coordinator
Step 3 Coordinator is waiting for vote from participants
Step 4 \(p_i\) (who voted YES) is waiting for COMMIT or ABORT
Timeout actions

Processes are waiting on steps 2, 3, and 4

Step 2. $p_i$ is waiting for VOTE-REQ from coordinator
Since it is has not cast its vote yet, $p_i$ can decide ABORT and halt.

Step 3. Coordinator is waiting for vote from participants
Coordinator can decide ABORT, send ABORT to all participants which voted YES, and halt.

Step 4. $p_i$ (who voted YES) is waiting for COMMIT or ABORT

Termination protocols

I. Wait for coordinator to recover
   □ It always works, since the coordinator is never uncertain
   □ may block recovering process unnecessarily

II. Ask other participants

Cooperative Termination

- $c$ appends list of participants to VOTE-REQ
- when an uncertain process $p$ times out, it sends a DECISION-REQ message to every other participant $q$
- if $q$ has decided, then it sends its decision value to $p$, which decides accordingly
- if $q$ has not yet voted, then it decides ABORT, and sends ABORT to $p$
- What if $q$ is uncertain?
Logging actions

1. When $c$ sends VOTE-REQ, it writes START-2PC to its DT Log
2. When $p_i$ is ready to vote YES,
   i. $p_i$ writes YES to DT Log
   ii. $p_i$ sends YES to $c$ ($p_i$ writes also list of participants)
3. When $p_i$ is ready to vote NO, it writes ABORT to DT Log
4. When $c$ is ready to decide COMMIT, it writes COMMIT to DT Log before sending COMMIT to participants
5. When $c$ is ready to decide ABORT, it writes ABORT to DT Log
6. After $p_i$ receives decision value, it writes it to DT Log

$p$ recovers

1. When coordinator sends VOTE-REQ, it writes START-2PC to its DT Log
2. When participant is ready to vote Yes, writes Yes to DT Log before sending yes to coordinator (writes also list of participants)
   When participant is ready to vote No, it writes ABORT to DT Log
3. When coordinator is ready to decide COMMIT, it writes COMMIT to DT Log before sending COMMIT to participants
   When coordinator is ready to decide ABORT, it writes ABORT to DT Log
4. After participant receives decision value, it writes it to DT Log

$p$ recovers

1. When coordinator sends VOTE-REQ, it writes START-2PC to its DT Log
2. When participant is ready to vote Yes, writes Yes to DT Log before sending yes to coordinator (writes also list of participants)
   When participant is ready to vote No, it writes ABORT to DT Log
3. When coordinator is ready to decide COMMIT, it writes COMMIT to DT Log before sending COMMIT to participants
   When coordinator is ready to decide ABORT, it writes ABORT to DT Log
4. After participant receives decision value, it writes it to DT Log

$p$ recovers

1. When coordinator sends VOTE-REQ, it writes START-2PC to its DT Log
2. If DT Log contains START-2PC, then $p = c$
3. If DT Log contains a decision value, then decide accordingly
4. Else decide ABORT
5. If DT Log contains a decision value, then decide accordingly
6. Else decide ABORT
7. Otherwise, $p$ is a participant:
   a. If DT Log contains a decision value, then decide accordingly
   b. Else if it does not contain a Yes vote, decide ABORT
   c. Else (Yes but no decision) run a termination protocol
2PC and blocking

- Blocking occurs whenever the progress of a process depends on the repairing of failures.
- No AC protocol is non-block in the presence of communication or total failures.
- But 2PC can block even with non-total failures and no communication failures among operating processes!

3-Phase Commit

- Two approaches:
  1. Focus only on site failures
     - Non-blocking, unless all sites fail
     - Timeout \( \equiv \) site at the other end failed
     - Communication failures can produce inconsistencies
  2. Tolerate both site and communication failures
     - Partial failures can still cause blocking, but less often than in 2PC

Blocking and uncertainty

Why does uncertainty lead to blocking?

- An uncertain process does not know whether it can safely decide COMMIT or ABORT because some of the processes it cannot reach could have decided either
Blocking and uncertainty

Why does uncertainty lead to blocking?

- An uncertain process does not know whether it can safely decide COMMIT or ABORT because some of the processes it cannot reach could have decided either.

Non-blocking Property
- If any operational process is uncertain, then no process has decided COMMIT.

2PC Revisited

In state PC, a process knows that it will commit unless it fails.
3PC: The Protocol
Dale Skeen (1982)

I. $c$ sends VOTE-REQ to all participants.

II. When $p_i$ receives a VOTE-REQ, it responds by sending a vote to $c$:
    - if vote = No, then decide$_i$ := ABORT and $p_i$ halts.
    - otherwise, $p_i$ collects votes from all $p_j$.

III. $c$ collects votes from all $p_j$:
    - if all votes are Yes, then $c$ sends PRECOMMIT to all $p_j$.
    - otherwise, $c$ halts.

IV. If $p_i$ receives PRECOMMIT, then it sends ACK to $c$.

V. $c$ collects ACKs from all $p_j$:
    - When all ACKs have been received, decide$_c$ := COMMIT.
    - $c$ sends COMMIT to all $p_j$.

VI. When $p_i$ receives COMMIT, $p_i$ sets decide$_i$ := COMMIT and halts.

2PC Revisited

1. $c$ sends VOTE-REQ to all participants.

2. When participant $p_1$ receives a VOTE-REQ, it responds by sending a vote to $c$:
   - if vote = No, then decide$_i$ := ABORT and $p_i$ halts.
   - otherwise, $c$ collects vote from all $p_j$.

3. If all votes are Yes, then $c$ sends PRECOMMIT to all $p_j$.
   - otherwise, $c$ halts.

4. If $p_i$ receives PRECOMMIT, then it sends ACK to $c$.

5. $c$ collects ACKs from all $p_j$:
   - When all ACKs have been received, decide$_c$ := COMMIT.
   - $c$ sends COMMIT to all $p_j$.

6. When $p_i$ receives COMMIT, $p_i$ sets decide$_i$ := COMMIT and halts.
Wait a minute!

1. \( e \) sends VOTE-REQ to all participants.
2. When participant \( p_i \) receives a VOTE-REQ, it responds by sending a vote to \( e \).
   - If vote is No, then \( p_i \) decides ABORT and halts.
3. \( e \) collects vote from all participants.
   - If all votes are Yes, then \( e \) sends PRECOMMIT to all participants.
   - If any vote is No, then \( e \) sends ABORT to all participants.
4. If \( p_i \) receives PRECOMMIT then it sends ACK to \( e \).
5. \( e \) collects ACKs from all participants.
   - If all ACKs have been received, \( e \) decides COMMIT.
   - Otherwise, \( e \) sends COMMIT to all participants.
6. When \( p_i \) receives COMMIT, it sets \( \text{decide COMMIT} \) and halts.

Messages are known to the recipient before they are sent...so, why are they sent?
They inform the recipient of the protocol's progress.

When \( e \) receives ACK from \( p_i \), it knows \( p_i \) is not uncertain.
When \( p_i \) receives COMMIT, it knows no participant is uncertain, so it can commit.
Timeout Actions

Processes are waiting on steps 2, 3, 4, 5, and 6

Step 2 \( p_i \) is waiting for VOTE-REQ from coordinator
   Exactly as in 2PC

Step 3 Coordinator is waiting for vote from participants
   Exactly as in 2PC

Step 4 \( p_i \) waits for PRECOMMIT
   Run some Termination protocol

Step 5 Coordinator waits for ACKs
   Coordinator sends COMMIT

Step 6 \( p_i \) waits for COMMIT

Participant knows what is going to receive...

Exactly as in 2PC

Coordinator sends COMMIT
Termination protocol: Process states

At any time while running 3 PC, each participant can be in exactly one of these 4 states:

- **Aborted**: Not voted, voted NO, received ABORT
- **Uncertain**: Voted YES, not received PRECOMMIT
- **Committable**: Received PRECOMMIT, not COMMIT
- **Committed**: Received COMMIT

Termination protocol

- When $P_i$ times out, it starts an election protocol to elect a new coordinator.
- The new coordinator sends STATE-REQ to all processes that participated in the election.
- The new coordinator collects the states and follows a termination rule.

<table>
<thead>
<tr>
<th>Process State</th>
<th>Aborted</th>
<th>Uncertain</th>
<th>Committable</th>
<th>Committed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aborted</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Uncertain</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Committable</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Committed</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

Termination protocol and failures

Processes can fail while executing the termination protocol...

- If $c$ times out on $p$, it can just ignore $p$.
- If $c$ fails, a new coordinator is elected and the protocol is restarted (election protocol to follow).
- Total failures will need special care...
Recovering $p$

- If $p$ fails before sending YES, decide ABORT
- If $p$ fails after having decided, follow decision
- If $p$ fails after voting YES but before receiving decision value
  - $p$ asks other processes for help
  - 3PC is non-blocking: $p$ will receive a response with the decision
- If $p$ has received PRECOMMIT
  - Still needs to ask other processes (cannot just COMMIT)

The election protocol

- Processes agree on linear ordering (e.g. by pid)
- Each $p$ maintains set $UP_p$ of all processes that $p$ believes to be operational
- When $p$ detects failure of $c$, it removes $c$ from $UP_p$ and chooses smallest $q$ in $UP_p$ to be new coordinator
- If $q = p$, then $p$ is new coordinator
- Otherwise, $p$ sends UR-ELECTED to $q$

Recovering $p$

- If $p$ fails before sending YES, decide ABORT
- If $p$ fails after having decided, follow decision
- If $p$ fails after voting YES but before receiving decision value
  - $p$ asks other processes for help
  - 3PC is non-blocking: $p$ will receive a response with the decision
- If $p$ has received PRECOMMIT
  - Still needs to ask other processes (cannot just COMMIT)

A few observations

- What if $p'$, which has not detected the failure of $c$, receives a STATE-REQ from $q$?
A few observations

What if $p'$, which has not detected the failure of $c$, receives a STATE-REQ from $q$?
- it concludes that $c$ must be faulty
- it removes from $UP_{p'}$ every $q' < q$

Total failure

Suppose $p$ is the first process to recover, and that $p$ is uncertain
- Can $p$ decide ABORT?

Some processes could have decided COMMIT after $p$ crashed!
Total failure

- Suppose \( p \) is the first process to recover, and that \( p \) is uncertain
- Can \( p \) decide ABORT?
  - Some processes could have decided COMMIT after \( p \) crashed!
- \( p \) is blocked until some \( q \) recovers s.t. either
  - \( q \) can recover independently
  - \( q \) is the last process to fail—then \( q \) can simply invoke the termination protocol

Determining the last process to fail

- Suppose a set \( R \) of processes has recovered
- Does \( R \) contain the last process to fail?
  - \( p \) is blocked until some \( q \) recovers s.t. either
    - \( q \) can recover independently
    - \( q \) is the last process to fail
  - so the last process to fail must be in
  \[ \bigcap_{p \in R} \text{UP}_p \]

\( R \) contains the last process to fail if
\[ \bigcap_{p \in R} \text{UP}_p \subseteq R \]