TWO-PHASE COMMIT

George Porter
May 9 and 11, 2018

ATTRIBUTION

• These slides are released under an Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0) Creative Commons license
• These slides incorporate material from:
  • Tanenbaum and Van Steen, Dist. Systems: Principles and Paradigms
  • Kyle Jamieson, Princeton University (also under a CC BY-NC-SA 3.0 Creative Commons license)
ANNOUNCEMENTS

Outline

1. Fault tolerance in a nutshell
2. Safety and liveness
3. Two-phase commit
4. Two-phase commit failure scenarios
WHAT IS FAULT TOLERANCE?

• Building **reliable** systems from **unreliable** components

• Three basic steps

  1. **Detecting errors**: discovering the presence of an error in a data value or control signal
  2. **Containing errors**: limiting how far errors propagate
  3. **Masking errors**: designing mechanisms to ensure a system operates correctly despite error (and possible correct error)

WHY IS FAULT TOLERANCE HARD?

• Say **one bit** in a DRAM fails...

• ...it **flips a bit** in a memory address the kernel is writing to...

• ...causes big memory error elsewhere, or a **kernel panic**...

• ...program is running one of many distributed file system storage servers...

• ...a client **can’t read from FS**, so it hangs.
### SO WHAT TO DO?

1. **Do nothing**: silently return the failure

2. **Fail fast**: detect the failure and report at interface
   - Ethernet station jams medium on detecting collision

3. **Fail safe**: transform incorrect behavior or values into acceptable ones
   - Failed traffic light controller switches to blinking-red

4. **Mask the failure**: operate despite failure
   - Retry op for transient errors, use error-correcting code for bit flips, replicate data in multiple places

### MASKING FAILURES

- We mask failures on **one server** via
  - Atomic operations
  - Logging and recovery
- In a distributed system with **multiple servers**, we might replicate some or all servers
- But if you have some replicated servers
  - You’re going to want a way to keep them all consistent in a fault tolerant way
Outline

1. Fault tolerance in a nutshell
2. Safety and liveness
3. Two-phase commit
4. Two-phase commit failure scenarios

REASONING ABOUT FAULT TOLERANCE

• This is hard!
  • How do we design fault-tolerant systems?
  • How do we know if we’re successful?
• Often use “properties” that hold true for every possible execution
• We focus on safety and liveness properties
SAFETY

• “Bad things” don’t happen
  • No stopped or deadlocked states
  • No error states
• Examples
  • Mutual exclusion: two processes can’t be in a critical section at the same time
  • Bounded overtaking: if process 1 wants to enter a critical section, process 2 can enter at most once before process 1

LIVENESS

• “Good things” happen
  • ...eventually
• Examples
  • Starvation freedom: process 1 can eventually enter a critical section as long as process 2 terminates
  • Eventual consistency: if a value in an application doesn’t change, two servers will eventually agree on its value
OFTEN A TRADEOFF

- “Good” and “bad” are application-specific
- Safety is very important in banking transactions
  - May take some time to confirm a transaction
- Liveness is very important in social networking sites
  - See updates right away (what about the “breakup problem”?)

Outline

1. Fault tolerance in a nutshell
2. Safety and liveness
3. Two-phase commit
4. Two-phase commit failure scenarios
MOTIVATION: SENDING MONEY

send_money(A, B, amount) {
    Begin_Transaction();
    if (A.balance - amount >= 0) {
        A.balance = A.balance - amount;
        B.balance = B.balance + amount;
        Commit_Transaction();
    } else {
        Abort_Transaction();
    }
}

SINGLE-SERVER: ACID

- **Atomicity**: all parts of the transaction execute or none (A’s decreases and B’s balance increases)
- **Consistency**: the transaction only commits if it preserves invariants (A’s balance never goes below 0)
- **Isolation**: the transaction executes as if it executed by itself (even if C is accessing A’s account, that will not interfere with this transaction)
- **Durability**: the transaction’s effects are not lost after it executes (updates to the balances will remain forever)
DISTRIBUTED TRANSACTIONS?

• Partition databases across multiple machines for scalability (A and B might not share a server)
• A transaction might touch more than one partition
• How do we guarantee that all of the partitions commit the transactions or none commit the transactions?

TWO-PHASE COMMIT (2PC)

• **Goal:** General purpose, distributed agreement on some action, with failures
  • Different entities play different roles in the action
• **Running example:** Transfer money from A to B
  • Debit at A, credit at B, tell the client “okay”
  • Require *both* banks to do it, or *neither*
  • Require that *one bank never act alone*
STRAW MAN PROTOCOL

1. C → TC: “go!”

 Client C

 Transaction Coordinator TC

 Bank A B

STRAW MAN PROTOCOL

1. C → TC: “go!”

 2. TC → A: “debit $20!”

  TC → B: “credit $20!”

  TC → C: “okay”

• A, B perform actions on receipt of messages
REASONING ABOUT THE STRAW MAN PROTOCOL

What could possibly go wrong?

1. Not enough money in A’s bank account?
2. B’s bank account no longer exists?
3. A or B crashes before receiving message?
4. The best-effort network to B fails?
5. TC crashes after it sends debit to A but before sending to B?

SAFETY VERSUS LIVENESS

• Note that TC, A, and B each have a notion of committing
• We want two properties:
  1. Safety
     • If one commits, no one aborts
     • If one aborts, no one commits
  2. Liveness
     • If no failures and A and B can commit, action commits
     • If failures, reach a conclusion ASAP
A **CORRECT ATOMIC COMMIT PROTOCOL**

1. **C → TC:** “go!”

2. **TC → A, B:** “prepare!”
A CORRECT ATOMIC COMMIT PROTOCOL

1. C → TC: “go!”

2. TC → A, B: “prepare!”

3. A, B → P: “yes” or “no”

4. TC → A, B: “commit!” or “abort!”
   - TC sends commit if both say yes
   - TC sends abort if either say no
A **CORRECT** ATOMIC COMMIT PROTOCOL

1. **C → TC:** “go!”
2. **TC → A, B:** “prepare!”
3. **A, B → P:** “yes” or “no”
4. **TC → A, B:** “commit!” or “abort!”
   - TC sends **commit** if both say yes
   - TC sends **abort** if either say no
5. **TC → C:** “okay” or “failed”
   - A, B commit on receipt of commit message

---

**REASONING ABOUT ATOMIC COMMIT**

- **Why is this correct?**
  - Neither can commit unless both agreed to commit
- **What about performance?**
  1. **Timeout:** I’m up, but didn’t receive a message I expected
     - Maybe other node crashed, maybe network broken
  2. **Reboot:** Node crashed, is rebooting, must clean up
TIMEOUTS IN ATOMIC COMMIT

Where do hosts wait for messages?

1. **TC** waits for “yes” or “no” from A and B
   - TC hasn’t yet sent any commit messages, so can safely abort after a timeout
   - But this is conservative: might be network problem
     - We’ve preserved correctness, sacrificed performance

2. A and B wait for “commit” or “abort” from TC
   - If it sent a no, it can safely abort (why?)
   - If it sent a yes, can it unilaterally abort?
   - Can it unilaterally commit?
   - A, B could wait forever, but there is an alternative...

SERVER TERMINATION PROTOCOL

• Consider Server B (Server A case is symmetric) waiting for commit or abort from TC
  - Assume B voted yes (else, unilateral abort possible)

• B → A: “status?” A then replies back to B. Four cases:
  - (No reply from A): no decision, B waits for TC
  - Server A received commit or abort from TC: Agree with the TC’s decision
  - Server A hasn’t voted yet or voted no: both abort
    - TC can’t have decided to commit
  - Server A voted yes: both must wait for the TC
    - TC decided to commit if both replies received
    - TC decided to abort if it timed out
REASONING ABOUT THE SERVER TERMINATION PROTOCOL

• What are the liveness and safety properties?
  • Safety: if servers don’t crash, all processes will reach the same decision
  • Liveness: if failures are eventually repaired, then every participant will eventually reach a decision

• Can resolve some timeout situations with guaranteed correctness

• Sometimes however A and B must block
  • Due to failure of the TC or network to the TC

• But what will happen if TC, A, or B crash and reboot?

HOW TO HANDLE CRASH AND REBOOT?

• Can’t back out of commit if already decided
  • TC crashes just after sending “commit!”
  • A or B crash just after sending “yes”

• If all nodes knew their state before crash, we could use the termination protocol...
  • Use write-ahead log to record “commit!” and “yes” to disk
RECOVERY PROTOCOL WITH NON-VOLATILE STATE

• If everyone rebooted and is reachable, TC can just check for commit record on disk and resend action
• TC: If no commit record on disk, abort
  • You didn’t send any “commit!” messages
• A, B: If no yes record on disk, abort
  • You didn’t vote “yes” so TC couldn’t have committed
• A, B: If yes record on disk, execute termination protocol
  • This might block

TWO-PHASE COMMIT

• This recovery protocol with non-volatile logging is called Two-Phase Commit (2PC)
• Safety: All hosts that decide reach the same decision
  • No commit unless everyone says “yes”
• Liveness: If no failures and all say “yes” then commit
  • But if failures then 2PC might block
  • TC must be up to decide
• Doesn’t tolerate faults well: must wait for repair
Outline

1. Fault tolerance in a nutshell
2. Safety and liveness
3. Two-phase commit
4. Two-phase commit failure scenarios

WHAT IF PARTICIPANT FAILS BEFORE SENDING RESPONSE?
WHAT IF PARTICIPANT FAILS AFTER SENDING VOTE

WHAT IF PARTICIPANT LOST A VOTE?
WHAT IF COORDINATOR FAILS BEFORE SENDING PREPARE?

WHAT IF COORDINATOR FAILS AFTER Sending PREPARE?
WHAT IF COORDINATOR FAILS AFTER RECEIVING VOTES

WHAT IF COORDINATOR FAILS AFTER SENDING DECISION?
DO WE NEED THE COORDINATOR?

UC San Diego