Inconsistency in Distributed Systems

CS 475, Fall 2019
Concurrent & Distributed Systems
Recurring Problem: Replication

- Replication solves some problems, but creates a huge new one: consistency

OK, we obviously need to actually do something here to replicate the data… but what?
Sequentially Consistent DSM

CPU 1
- thread0()
- CPU Cache
- 7ns
- 100ns
- FIFO queue
- 1s?

CPU 2
- thread1()
- CPU Cache
- X

Main Memory
Ivy Architecture

Each node keeps a cached copy of each piece of data it reads.

If some data doesn’t exist locally, request it from remote node.
Sequential Consistency

Set A=5

“OK”!

Read A

“5”!

Set A=5

“OK!”
Our protocol for sequential consistency does NOT guarantee that the system will be available!
Consistent + Available

Set A=5

“OK”!

“5”!

Set A=5

Assume replica failed

Read A

A

B

5

7

6

7

5

6

7

A

B

A
Still broken...

Set $A = 5$

"OK"!

Read $A$

"6"!

Assume replica failed

Set $\text{replica} = 5$

5 7 6 7
Network Partitions

- The communication links between nodes may fail arbitrarily.
- But other nodes might still be able to reach that node.

Set A = 5

"OK"!

Read A

"6"!

Assume replica failed
HW3 Discussion

Go to socrative.com and select “Student Login” Room: CS475; ID is your G-Number

1. How fair do you think this assignment was?
2. How difficult did you think this assignment was?
3. How long did you spend on this assignment?

Reminder: If you are not in class, you may not complete the activity. If you do anyway, this will constitute a violation of the honor code.
Today

- Consistency in distributed systems - can we have it all? If not, what can we get?
- Relaxed consistency models
- Reminders:
  - HW3 graded by end of week
  - HW4 is out!
CAP Theorem

• Pick two of three:
  • Consistency: All nodes see the same data at the same time (sequential consistency)
  • Availability: Individual node failures do not prevent survivors from continuing to operate
  • Partition tolerance: The system continues to operate despite message loss (from network and/or node failure)
• You can not have all three, ever
CAP Theorem vs FLP

- FLP: Can not guarantee both liveness and agreement assuming messages may be delayed but are eventually delivered
- CAP: Can not guarantee consistency, availability, partition-tolerance assuming messages may be dropped
CAP Theorem

• C+A: Provide strong consistency and availability, assuming there are no network partitions
• C+P: Provide strong consistency in the presence of network partitions; minority partition is unavailable
• A+P: Provide availability even in presence of partitions; no sequential consistency guarantee, maybe can guarantee something else
Relaxing Consistency

- We can relax two design principles:
  - How stale reads can be
  - The ordering of writes across the replicas
### Allowing Stale Reads

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
</tr>
</thead>
<tbody>
<tr>
<td>W(X)</td>
<td>0</td>
<td>W(X) 1</td>
<td></td>
</tr>
<tr>
<td>R(X)</td>
<td></td>
<td>R(X)</td>
<td>W(X) 0</td>
</tr>
<tr>
<td>R(X)</td>
<td></td>
<td>R(X)</td>
<td>R(X)</td>
</tr>
<tr>
<td>R(X)</td>
<td></td>
<td>R(X)</td>
<td>R(X)</td>
</tr>
</tbody>
</table>
Allowing Stale Reads

class MyObj {
    int x = 0;
    int y = 0;

    void thread0()
    {
        x = 1;
        if(y==0)
            System.out.println("OK");
    }

    void thread1()
    {
        y = 1;
        if(x==0)
            System.out.println("OK");
    }
}

Java's memory model is “relaxed” in that you can have stale reads (if not volatile).
Relaxing Consistency

- Intuition: less constraints means less coordination overhead, less prone to partition failure

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
</tr>
</thead>
<tbody>
<tr>
<td>W(X)</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>R(X) [0,1]</td>
<td></td>
<td>R(X) [0,1]</td>
<td></td>
</tr>
<tr>
<td>R(X) [0,1]</td>
<td></td>
<td></td>
<td>R(X) [0,1]</td>
</tr>
</tbody>
</table>
Naïve DSM

- Assume each machine has a complete copy of memory
- Reads from local memory
- Writes broadcast update to other machines, then immediately continue

```java
class Machine1 {
    DSMInt x = 0;
    DSMInt y = 0;
    
    static void main(String[] args) {
        x = 1;
        if(y==0) System.out.println("OK");
    }
}

class Machine2 {
    DSMInt x = 0;
    DSMInt y = 0;
    
    static void main(String[] args) {
        y = 1;
        if(x==0) System.out.println("OK");
    }
}
```
Naïve DSM

- Assume each machine has a complete copy of memory
- Reads from local memory
- Writes broadcast update to other machines, then immediately continue

class Machine1 {
    DSMInt x = 1;
    DSMInt y = 0;

    static void main(String[] args) {
        x = 1;
        if(y==0)
            System.out.println("OK");
    }
}

class Machine2 {
    DSMInt x = 0;
    DSMInt y = 1;

    static void main(String[] args) {
        y = 1;
        if(x==0)
            System.out.println("OK");
    }
}
Naïve DSM

- Assume each machine has a complete copy of memory
- Reads from local memory
- Writes broadcast update to other machines, then immediately continue

```java
class Machine1 {
    DSMInt x = 0;
    DSMInt y = 0;
    static void main(String[] args) {
        x = 1;
        if (y == 0)
            System.out.println("OK");
    }
}

class Machine2 {
    DSMInt x = 0;
    DSMInt y = 0;
    static void main(String[] args) {
        y = 1;
        if (x == 0)
            System.out.println("OK");
    }
}
```

Is this correct?
Naïve DSM

• It definitely is not sequentially consistent
• Are there any guarantees that it provides though?
  • Reads can be stale
  • Writes can be re-ordered
  • Not really.
• Can we come up with something more clever though with SOME guarantee?
  • (Not as is, but with some modifications maybe it’s…)
Causal Consistency

- An execution is **causally-consistent** if all **causally-related** read/write operations are executed in an order that reflects their causality.
- Reads are fresh ONLY for writes that they are dependent on.
- Causally-related writes appear in order, but not in order to others.
- Concurrent writes can be seen in different orders by different machines.
- Compare to sequential consistency: **every machine** must see the same order of operations!
## Causal Consistency

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W(X)a</td>
<td>R(X)a</td>
<td>R(X)a</td>
<td>R(X)a</td>
</tr>
<tr>
<td></td>
<td>W(X)c</td>
<td>W(X)b</td>
<td>R(X)c</td>
<td>R(X)b</td>
</tr>
</tbody>
</table>

**Causally Consistent.** $W(X)\ b$ and $W(X)\ c$ are not related, hence could have happened one either order. $W(X)a$ and $W(X)b$ ARE causally related and must occur in this order.
Causal Consistency

NOT Causally Consistent. X couldn’t have been b after it was a

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W(X)a</td>
<td>R(X)a</td>
<td>R(x)b</td>
<td>R(x)a</td>
</tr>
<tr>
<td></td>
<td></td>
<td>W(X)b</td>
<td>R(x)</td>
<td>R(x)b</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R(x)b</td>
<td>R(x)a</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R(x)</td>
<td>R(x)b</td>
</tr>
</tbody>
</table>

Causally Consistent. X can be a or b concurrently
Why Causal Consistency?

- It is clearly weaker than sequential consistency
  - (Note that anything that is sequentially consistent is also causally consistent)
- Many more operations for concurrency
  - Parallel (non-dependent) operations can occur in parallel in different places
    - Sequential would enforce a global ordering
  - E.g. if W(X) and W(Y) occur at the same time, and without dependencies, then they can occur without any locking
- Still requires some perhaps complicated implementation - each client must know what is related to what.
Eventual Consistency

• Allow stale reads, but ensure that reads will **eventually** reflect the previously written values
• Eventually: milliseconds, seconds, minutes, hours, years…
• Writes are NOT ordered as executed
  • Allows for conflicts. Consider: Dropbox
• Git is eventually consistent
Eventual Consistency

- More concurrency than strict, sequential or causal
  - These require **highly available** connections to send messages, and generate lots of chatter
- Far looser requirements on network connections
  - Partitions: OK!
  - Disconnected clients: OK!
  - Always available!
- Possibility for conflicting writes :(
Each node keeps a cached copy of each piece of data it reads. If some data doesn't exist locally, request it from a remote node.

Write X=1

invalidate x

All of these messages...
All of the clients must always be online!
Relax!

Read X

cached data
x=0

cached data
x=1

Read X

read x

read x
Sequential vs Eventual Consistency

• Sequential: “Pessimistic” concurrency control
  • Assume that everything could cause a conflict, decide on an update order as things execute, then enforce it
• Eventual: “Optimistic” concurrency control
  • Just do everything, and if you can’t resolve what something should be, sort it out later
  • Can be tough to resolve in general case
When everything can talk, it’s easy to synchronize, right?

Goal: Everything eventually becomes synchronized. No lost updates (don’t replace new version with old)
Eventual Consistency: Distributed Filesystem

When everything can talk, it's easy to synchronize, right?

Goal: Everything eventually becomes synchronized.
No lost updates (don’t replace new version with old)

Fix: Add coordinating sync server
Eventual Consistency: Distributed Filesystem

- Role of the sync server:
  - Resolve conflicting changes, report conflicts to user
  - Do not allow sync between clients
  - Detect if updates are sequential
  - Enforce ordering constraints
Detecting Conflicts

Do we just use timestamps?

$t=0$
write $x = a$

$t=1$
write $x = b$
Detecting Conflicts

Do we just use timestamps?

\[ t=0 \]
write \( x = a \)

\[ t=1 \]
write \( x = b \)

NO, what if clocks are out of sync?
NO does not actually detect conflicts
Detecting Conflicts

Solution: Track version history on clients

\[ v=0 \]
write \( x = a \)

\[ v=0 \]
write \( x = b \)

Still doesn’t tell us what to do with a conflict
Client-Centric Consistency

• What can we guarantee in disconnected operation?
• Monotonic-reads: any future reads will return the same or newer value (never older)
• Monotonic-writes: A processes’ writes are always processed in order
• Read-you-writes
• Writes follow reads
Eventually Consistent +
Available + Partition Tolerant

Set $A=5$

"OK"!

Read $A$

"6"!

Set $A=5$

Assume replica failed

5

7

5

7
Choosing a consistency model

- **Strict Consistency**
  - Read always returns value from latest write
- **Sequential Consistency**
  - All nodes see operations in some sequential order
  - Operations of each process appear in-order in this sequence
- **Causal Consistency**
  - All nodes see causally related writes in same order
  - But concurrent writes may be seen in different order on different machines
- **Eventual Consistency**
  - All nodes will learn eventually about all writes, in the absence of updates
Example: Facebook

• Problem: >1 billion active users
• Solutions: Thousands of servers across the world
• What kind of consistency guarantees are reasonable? Need 100% availability!
• If I post a story on my news feed, is it OK if it doesn’t immediately show up on yours?
• Two users might not see the same data at the same time
• Now this is “solved” anyway because there is no “sort by most recent first” option anyway
Example: Web-based mail client

• Problem: entire email database is replicated to $N$ remote servers for redundancy. Want to have replication *but* not impose significant availability problems.

• Consider: Saving an email to your “sent folder” vs marking an email as read

• Solutions:
  • Saving mail to sent folder is important! People hate when they think they sent something but it got lost! $\rightarrow$ Make them wait for 2PC replication
  • Marking an email as read is not as important! It would be annoying if you couldn’t browse emails while that message was being updated $\rightarrow$ perform replication in background (no strong consistency)
Example: Airline Reservations

- Reservations and flight inventory are managed by a GDS (Global Distribution System), who acts as a middle broker between airlines, ticket agencies and consumers [Except for Southwest and Air New Zealand and other oddballs]
- GDS needs to sell as many seats as possible within given constraints
- If I have 100 seats for sale on a flight, does it matter if reservations for flights are reconciled immediately?
- If I have 5 seats for sale on a flight, does it matter if reservations are reconciled immediately?
Example: Airline Reservations

• Result: Reservations can be made using either a strong consistency model or a weak, eventual one
• Most reservations are made under the normal strong model (reservation is confirmed immediately)
• GDS also supports “Long Sell” - issue a reservation without confirmed availability, need to eventually reconcile it
• Long sells require the seller to make clear to the customer that even though there’s a confirmation number it’s not confirmed!
Fail: Airline Reservations

What I learned: International flights booked through united.com are not tickets until they are confirmed internally a couple days later, even though they bill your credit card and give a confirmation number.

David Darais @daviddarais
@united just found out the international flight I booked over a month ago was canceled 3 days after I booked, but with no email or text to notify me. thought I was flying today but my flight doesn’t exist and phone support isn’t helping. Help??

2:43 PM - 21 Oct 2019

9 Retweets 17 Likes

David Darais @daviddarais · Oct 21
If confirmation fails in “rare cases” your flight is quietly canceled with no notification; no email or text. We tried to check in for our international flight today and saw a canceled flight; called United.
Filesystem consistency

- What consistency guarantees do a filesystem provide?
  - read, write, sync, close
  - On sync, guarantee writes are persisted to disk
  - Readers see most recent
  - What does a network file system do?
Network Filesystem Consistency

- How do you maintain these same semantics?
- (Cheat answer): Very, very expensive
  - EVERY write needs to propagate out
  - EVERY read needs to make sure it sees the most recent write
- Oof. Just like Ivy.
Consistency Takeaways

- Strong consistency (sequential or strict) comes at a tradeoff: performance, availability
- Weaker consistency also has a tradeoff (weaker consistency)
- But: applications can make these design choices clear to end-users
  - Facebook
  - Dropbox
- Next week: examples of two systems that involve replication and handle consistency differently: DNS, NFS
• Aside from consistency semantics, we can choose **what to replicate**: state vs operations
• Sending operations (notification of an update)
  • An “invalidation” protocol (like Ivy). Works well when data is not often shared.
• Sending update operations
  • Works well when data is read more than written
  • With big files, can choose to only send diffs to preserve bandwidth
• When do we send updates?
  • “Pull” based - replicas/clients poll for updates (caches)
  • “Push” based - server pushes updates to clients
HW4 - Push-based replication

- Each KVStore client will have the entire dataset cached locally
- When updating values, the update will be propagated to each replica
• Each KVStore client will have the entire dataset cached locally
• When updating values, the update will be propagated to each replica

HW4 - Push-based replication
Ivy vs HW4

- Ivy never copies the actual values until a replica reads them (unlike HW4)
  - Invalidate messages are probably smaller than the actual data!
- Ivy only sends update (invalidate) messages to replicas who have a copy of the data (unlike HW4)
  - Maybe most data is not actively shared
- Ivy requires the lock server to keep track of a few more bits of information (which replica has which data)
  - With near certainty Ivy is a lot faster :)

Ivy vs HW4
This work is licensed under a Creative Commons Attribution-ShareAlike license

- This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/4.0/
- You are free to:
  - Share — copy and redistribute the material in any medium or format
  - Adapt — remix, transform, and build upon the material
  - for any purpose, even commercially.
- Under the following terms:
  - Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
  - ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.
  - No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.