CS 138: Epidemics and Bayou
Topics

- Client-based consistency
- Epidemic algorithms
- Bayou and Weak Consistency
Room Reservation

- Room database maintained on one or more servers
- Clients Gossip to make updates
  - everything is linearizable
- Problem solved …
Room Reservation (2)

- 138 TAs schedule room for help session
- 167 TAs also schedule room for help session
- But it’s spring break …
  - 138 TAs are on flight to Rio de Janeiro
    - at least one has copy of reservation DB
  - 167 TAs are on flight to Tahiti
    - at least one has copy of reservation DB
  - can’t detect conflicts en route
    - cheap airfares — no internet on plane
  - Gossip can’t cope
Weak Consistency

• Replicas are eventually consistent with one another
  – assuming no conflicts
Synchronization of Content

- Consensus
  - requires quorum
- Simpler approach
  - each node periodically contacts a random node
    - update it (or vice versa)
  - doesn’t require quorum
  - does it work?
Anti-entropy works as follows: periodically, a server P randomly chooses a server Q and then the two exchange databases. Three modes of anti-entropy are possible. In push mode, P pushes its database to Q, which adds to its database everything in P that it doesn’t have. In the pull mode, the reverse is done. Push-pull mode works in both directions.

Epidemics

- Established theory of how diseases spread by epidemics
  - adapt it
  - primary difference: infection is good
We look at the fundamentals of epidemic theory in order to analyze the basic performance of the anti-entropy algorithms. Assume that any server can infect any other server with equal probability. Also assume that k members are already infected and infections occur in rounds: at each round, each server randomly picks another server.
Time to Infect

- Number of rounds necessary to infect whole population grows $O(\log n)$
  - for push: $\log_2(n) + \ln(n) + O(1)$ (large $n$)
If we compare the performance of the pull and push approaches; we observe that (for a very large number of servers) the pull approach converges (i.e., infects the entire population) faster than push does. Push-pull performs similar to pull.

---

**Push vs. Pull**

- Which is better initially?
  - push
- If number of infected members is large; pull is better
  - let $p_i$ be the probability of a site's remaining susceptible after the $i$th cycle of anti-entropy
    - for pull, a site remains susceptible after the $i+1$st cycle if it was susceptible after the $i$th cycle and it contacted a susceptible site in the $i+1$st cycle.
      
      
      $p_{i+1} = p_i^2$

      quickly goes to zero for small $p_i$
    - for push, a site remains susceptible after the $i+1$st cycle if it was susceptible after $i$th cycle and no infectious site chose to contact it in the $i+1$st cycle.

      $p_{i+1} = p_i(1 - \frac{1}{n})^{n(1-p_i)} \approx \frac{p_i}{e}$
Exchange Updates, Not Databases

- Each node keeps list of updates (list of “infections”)
  - propagates these to others at each contact
    - receiver adds updates to its list
  - when do items get removed from list?
A form of non-simple epidemic is called *rumor mongering*.
The math is taken from epidemic theory. The final equation tells us how many are still susceptible when there are no more infectious nodes. Thus these will never get the updates.
Rumor Mongering with Anti-Entropy

- Spread updates using rumor mongering
  - some nodes will not get updates
  - but exchanges are cheap
- Run anti-entropy infrequently
  - guarantees all nodes will get updates
  - but exchanges are expensive
  - if susceptible node encountered, restart rumor mongering
Recall that we have no time stamps, so it cannot be determined if an update
is new or old.
Death Certificates

- Solution: hold on to fact that x has been deleted
  - propagate death certificates
  - like updates, but record a void

- How long do nodes hold on to death certificates?
  - a finite period
    - never long enough …

- Retain a few dormant death certificates
  - will be repropagated as death certificates if old update shows up
There are many scenarios in which conventional version vector-based techniques fail to detect the conflicts correctly. In some applications, concurrent writes to the same object may not conflict as expected. Furthermore, concurrent writes to different objects may conflict. Bayou handles such cases by having the application define its own notion of conflict.
Papers on Bayou include:
Conflict detection is accomplished by using dependency checks — each write operation also supplies a dependency check that specifies a simple condition defined over the state of the database. The dependency check also defines what the application expects to see as a result when the condition gets evaluated. If indeed the expected result is observed then Bayou decides that there is no conflict, else it decides that there is a conflict.

If there is a conflict, then a merge procedure (i.e., a conflict resolution procedure), also supplied by the application, is executed to resolve the conflict. The merge procedure produces a new alternate update that is acceptable by the application and that would not create a conflict with the current database state (i.e., an update for which the dependency check will produce the expected result).
A Bayou “write”

- Processed at each replica:

```c
Bayou_Write(update, dep_check, mergeproc) {
    IF (DB_EVAL(dep_check.query) ≠
        dep_check.expected_result)
        resolved_update = EXECUTE(mergeproc);
    ELSE
        resolved_update = update;
    DB_EXEC(resolved_update);
}
```

The slide shows the basic format of a Bayou write.
The slide shows an example Bayou write, including the dependency check and the merge procedure.
Bayou

- Updates propagated via epidemic protocols
  - some combination of rumor mongering and anti-entropy
- Weak consistency
  - requirements given from the client's perspective
  - client can choose desired (weak) consistency model
    - (next several slides)
Monotonic Reads

• A data store is said to provide monotonic-read consistency if the following condition holds:
  – if a process reads the value of a data item $x$ any successive read operation on $x$ by that process will always return that same value or a more recent value
• Example (of its not holding)
  – you read your calendar at one server, then go to another and don't see a recent update that was present at the first server, then go to another server and see it, then go to another and don't see it …

Note that “process” here means a client who is connecting to multiple servers.
Monotonic Writes

- In a monotonic-write consistent store, the following condition holds:
  - a write operation by a process on a data item x is completed before any successive write operation on x by the same process

- Example
  - you update a header file in a source-code directory on one server, then go to another and update a source-code file; someone else now tries to compile
Read Your Writes

• A data store is said to provide read-your-writes consistency, if the following condition holds
  – the effect of a write operation by a process on data item x will always be seen by a successive read operation on x by the same process
• Example
  – you change your password, then move to another server and your new password hasn’t taken effect
Writes Follow Reads

• A data store is said to provide writes-follow-reads consistency, if the following holds:
  – a write operation by a process on a data item x following a previous read operation on x by the
    same process is guaranteed to take place on
    the same or a more recent value of x that was
    read

• Example:
  – consider a loosely connected bulletin-board
    service (such as Usenet). You would like it to
    be the case that responses to queries are seen
    after the queries at all sites
We sketch an approach to an implementation of the constraints. Our first few ideas aren’t entirely practical, but will lead to an approach that is.
If a server can't be found that satisfies the restriction of the second bullet, then the application must be notified that read-your-writes cannot be maintained.
Monotonic Reads

- Before each read from server $S$ at time $t$, client must check that
  read set $\subseteq DB(S, t)$
- After each read, add to read set the writes that the read depended on
Anti-Entropy Constraints

- Required of servers to handle *monotonic writes* and *writes follow reads*
  - new writes accepted by a server from a client are ordered after existing writes
  - during anti-entropy exchanges:
    - if server S1 sends write W2 to server S2, then any W1 ordered before W2 on S1 is also sent to S2
Writes Follow Reads

- After each read, add to read set the writes that the read depended on
- Before each write at time $t$, client must check that
  
  $\text{read set} \subseteq \text{DB}(S, t)$
Monotonic Writes

- Before each write at time t, client must check that
  \[ \text{write set} \subseteq \text{DB}(S, t) \]
- After each write is accepted by server, client adds write ID to its write set
A Practical Implementation

- Implement write IDs as <server, logical time>
- Servers maintain version vectors (essentially vector clocks)
  - \( V_s[x] \) = logical time of most recent (by logical time) write ID received from server \( x \) by server \( S \)
- Clients maintain two session vectors, updated like vector clocks
  - read vector: writes relevant to session reads
  - write vector: session writes
Anti-Entropy Implementation

- Each server keeps log of all writes, ordered by write ID
- On anti-entropy exchange (sender S; receiver R):
  - S sends R all writes unknown to R
    - uses version vectors
Ordering

- Initial write order is tentative
- Designated primary server determines total order of writes: *commit sequence number* (CSN)
  - write ID = <CSN, server ID, logical time>
    - CSN = $\infty$ initially (while tentative)
  - primary propagates updated write IDs
  - once received, order is permanent (and total)

```
c_0 | c_1 | ... | c_n | t_0 | t_1 | t_2 | ... | t_i | t_{i+1}
```

committed    tentative
Log Truncation

- When can it be done?

```
c_0  c_1  \cdots  c_n  t_0  t_1  t_2  \cdots  t_i  t_{i+1}
```

comitted \hspace{1cm} tentative

This part is fixed and represented by database contents

This part is not fixed: its order can change