# Spanner

COS 418/518: Distributed Systems
Lecture 17

Mike Freedman & Wyatt Lloyd

Slides adapted from the Spanner OSDI talk

## Why Google Built Spanner

2005 – BigTable [OSDI 2006]
- Eventually consistent across datacenters
- Lesson: "don't need distributed transactions"

2008? – MegaStore [CIDR 2011]
- Strongly consistent across datacenters
- Option for distributed transactions
  - Performance was not great…

2011 – Spanner [OSDI 2012]
- Strictly Serializable Distributed Transactions
- "We wanted to make it easy for developers to build their applications"

**Spanner: Google's Globally-Distributed Database**

**OSDI 2012**

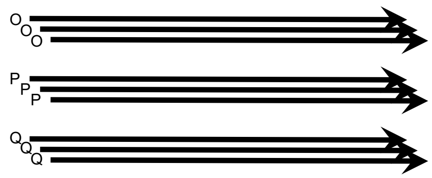
3

## Google's Setting

- Dozens of datacenters (zones)
- Per zone, 100-1000s of servers
- Per server, 100-1000 shards (tablets)
- Every shard replicated for fault-tolerance (e.g., 5x)

4

## Scale-out vs. Fault Tolerance




- Every shard replicated via MultiPaxos
- So every "operation" within transactions across tablets actually a replicated operation within Paxos RSM
- Paxos groups can stretch across datacenters!

5

## Read-Only Transactions

- Transactions that only read data
  - Predeclared, i.e., developer uses READ_ONLY flag / interface

- Reads are dominant operations
  - e.g., FB's TAO had 500 reads : 1 write [ATC 2013]
  - e.g., Google Ads (F1) on Spanner from 1? DC in 24h:
    21.5B reads
    31.2M single-shard transactions
    32.1M multi-shard transactions

## Make Read-Only Txns Efficient

- Ideal: Read-only transactions that are non-blocking
  - Arrive at shard, read data, send data back
  - Impossible with Strict Serializability ("SNOW theorem" later)

- Goal 1: Lock-free read-only transactions

- Goal 2: Non-blocking stale read-only txns

**Disruptive idea:**

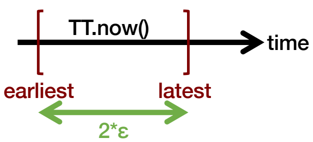
Do clocks **really** need to be **arbitrarily** unsynchronized?

Can you engineer some max divergence?

8

## TrueTime

- "Global wall-clock time" with bounded uncertainty
  - ε is worst-case clock divergence
  - Timestamps become intervals, not single values




- Consider event $e_{now}$ which invoked tt = TT.now():
  - Guarantee: tt.earliest <= $t_{abs}(e_{now})$ <= tt.latest

9

## TrueTime for Read-Only Txns

- Assign all transactions a wall-clock commit time (s)
  - All replicas of all shards track how up-to-date they are with $t_{safe}$:
  - all transactions with s < $t_{safe}$ have committed on this machine
- $t_{safe} = \min(t^{Paxos}_{safe}, t^{TM}_{safe})$

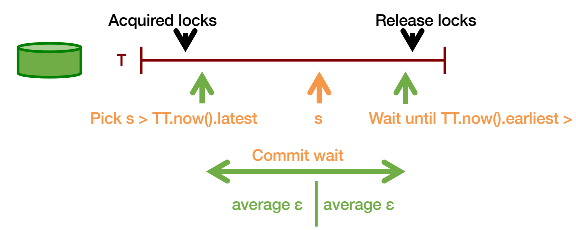
tion manager has a safe time $t^{TM}_{safe}$. $t^{Paxos}_{safe}$ is simpler: it is the timestamp of the highest-applied Paxos write. Because timestamps increase monotonically and writes are applied in order, writes will no longer occur at or below $t^{Paxos}_{safe}$ with respect to Paxos.

pared transaction's timestamp. Every participant leader (for a group $g$) for a transaction $T_i$ assigns a prepare timestamp $s^{prepare}_{i,g}$ to its prepare record. The coordinator leader ensures that the transaction's commit timestamp $s_i >= s^{prepare}_{i,g}$ over all participant groups $g$. Therefore, for every replica in a group $g$, over all transactions $T_i$ prepared at $g$, $t^{TM}_{safe} = min_i(s^{prepare}_{i,g}) - 1$ over all transactions prepared at $g$.

## TrueTime for Read-Only Txns

- Assign all transactions a wall-clock commit time (s)
  - All replicas of all shards track how up-to-date they are with $t_{safe}$:
  - all transactions with s < $t_{safe}$ have committed on this machine
- Goal 1: Lock-free read-only transactions
  - Current time ≤ TT.now.latest()
  - $s_{read}$ = TT.now.latest()
  - wait until $s_{read} < t_{safe}$
  - Read data as of $s_{read}$
- Goal 2: Non-blocking stale read-only txns
  - Similar to above, except explicitly choose time in the past
  - (Trades away consistency for better perf, e.g., lower latency)

## Timestamps and TrueTime

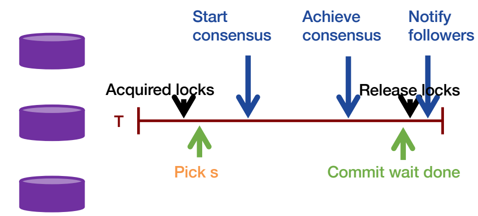



12

## Commit Wait

- Enables efficient read-only transactions
- Cost: 2ε extra latency
- Reduce/eliminate by overlapping with:
  - Replication
  - Two-phase commit

## Commit Wait and Replication




**Sufficient for single-shard transactions!**

14

## Client-Driven Transactions for Multi-Shard Transactions

### Client (via 2PL w/ 2PC) :

1. Issues reads to leader of each shard group, which acquires read locks and returns most recent data
2. Locally performs writes
3. Chooses coordinator from set of leaders, initiates commit
4. Sends commit message to each leader, include identity of coordinator and buffered writes
5. Waits for commit from coordinator

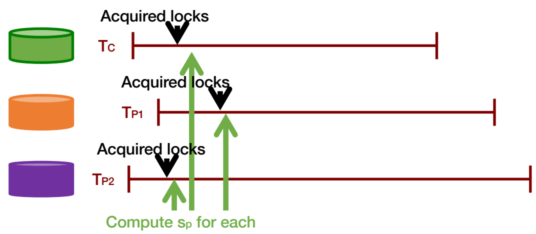
15

## Commit Wait and 2PC

- On commit msg from client, leaders acquire local write locks
- If non-coordinator:
  - Choose prepare ts > previous local timestamps
  - Log prepare record through Paxos
  - Notify coordinator of prepare timestamp
- If coordinator:
  - Wait until hear from other participants
  - Choose commit timestamp >= prepare ts, > local ts
  - Logs commit record through Paxos
  - Wait commit-wait period
  - Sends commit timestamp to replicas, other leaders, client
- All apply at commit timestamp and release locks

16

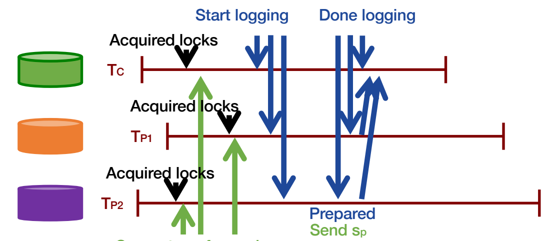
## Commit Wait and 2PC




1. Client issues reads to leader of each shard group, which acquires read locks and returns most recent data

17

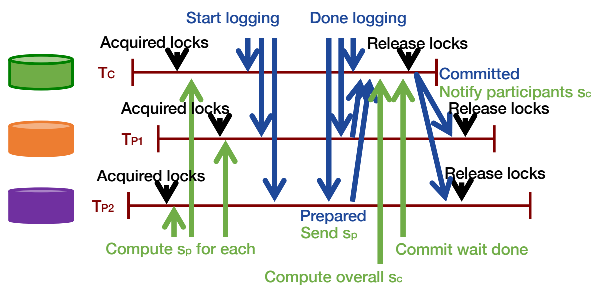
## Commit Wait and 2PC




2. Locally performs writes
3. Chooses coordinator from set of leaders, initiates commit
4. Sends commit msg to each leader, incl. identity of coordinator

18

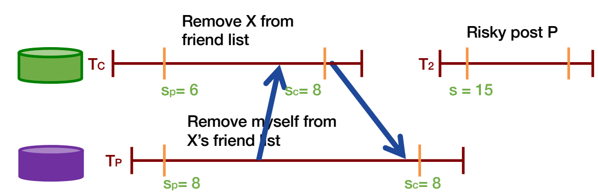
## Commit Wait and 2PC




5. Client waits for commit from coordinator

19

## Example




| Time | <8 | 8 | 15 |
|---|---|---|---|
| My friends | [X] | [] | |
| My posts | | | [P] |
| X's friends | [me] | [] | |

20

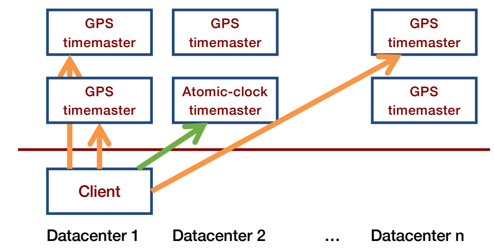
**Disruptive idea:**

Do clocks **really** need to be arbitrarily unsynchronized?

**Can you engineer some max divergence?**



## TrueTime Architecture

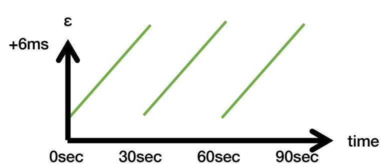



**Compute reference [earliest, latest] = now ± ε**



## TrueTime Implementation

| | | |
|---|---|---|
| now | = reference now | + local-clock offset |
| ε | = reference ε | + worst-case local-clock drift |
| | = 1ms | + 200 μs/sec |




- What about faulty clocks?
  - Bad CPUs 6x more likely in 1 year of empirical data



## Spanner

- Make it easy for developers to build apps!
- Reads dominant, make them lock-free
- TrueTime exposes clock uncertainty
  - Commit wait ensures transactions end after their commit time
  - Read at TT.now.latest()
- Globally-distributed database
  - 2PL w/ 2PC over Paxos!