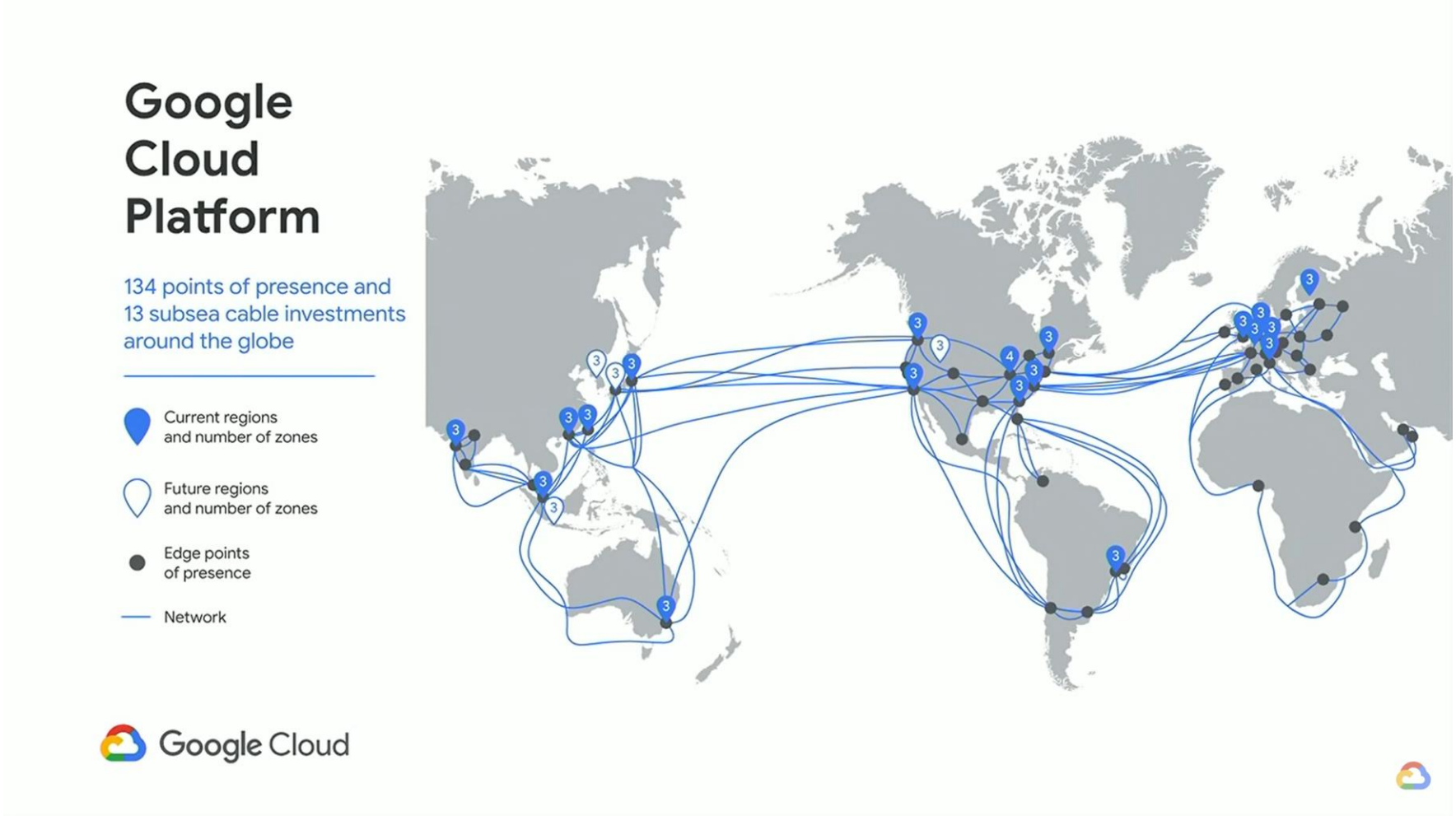
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

HY590.45
Modern Topics in Scalable Storage Systems

Kostas Magoutis
magoutis@csd.uoc.gr
http://www.csd.uoc.gr/~hy590-45

Google network and infrastructure sites

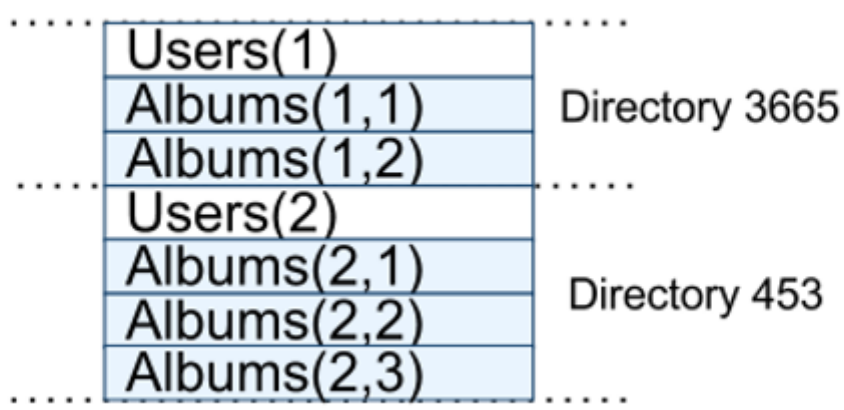


Source: Spanner Internals Part 1: What Makes Spanner Tick? (Cloud Next '19)

Spanner schema

```
CREATE TABLE Users {
  uid INT64 NOT NULL, email STRING
} PRIMARY KEY (uid), DIRECTORY;

CREATE TABLE Albums {
  uid INT64 NOT NULL, aid INT64 NOT NULL,
  name STRING
} PRIMARY KEY (uid, aid),
  INTERLEAVE IN PARENT Users ON DELETE CASCADE;
```



Logical data layout

Albums

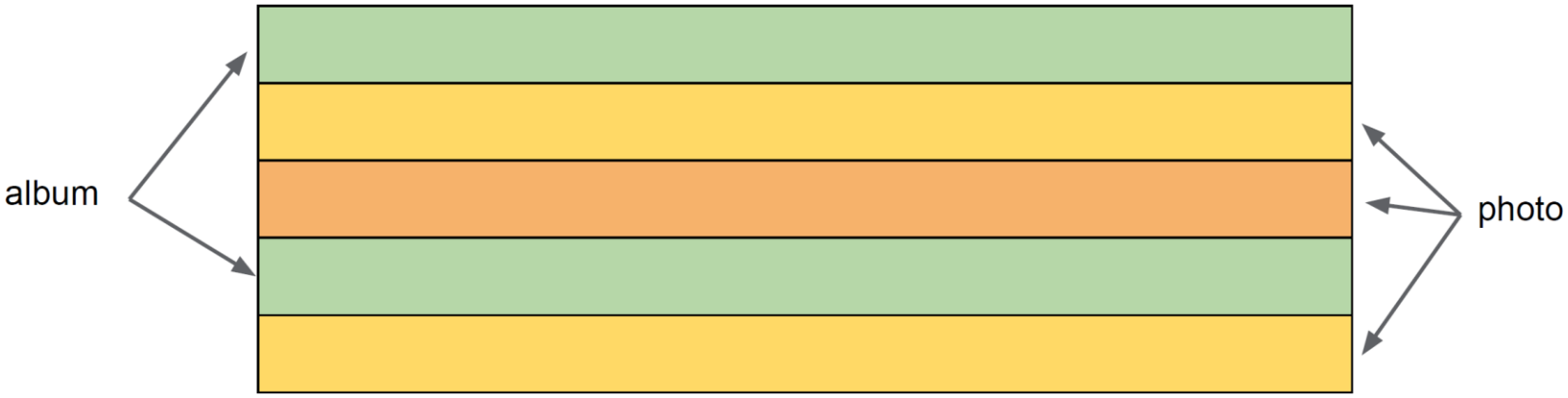
user_id	album_id	name
1	1	Maui
1	2	St. Louis

Photos

user_id	album_id	photo_id	title
1	1	2	Beach
1	1	5	Snorkeling
1	2	3	Gateway Arch

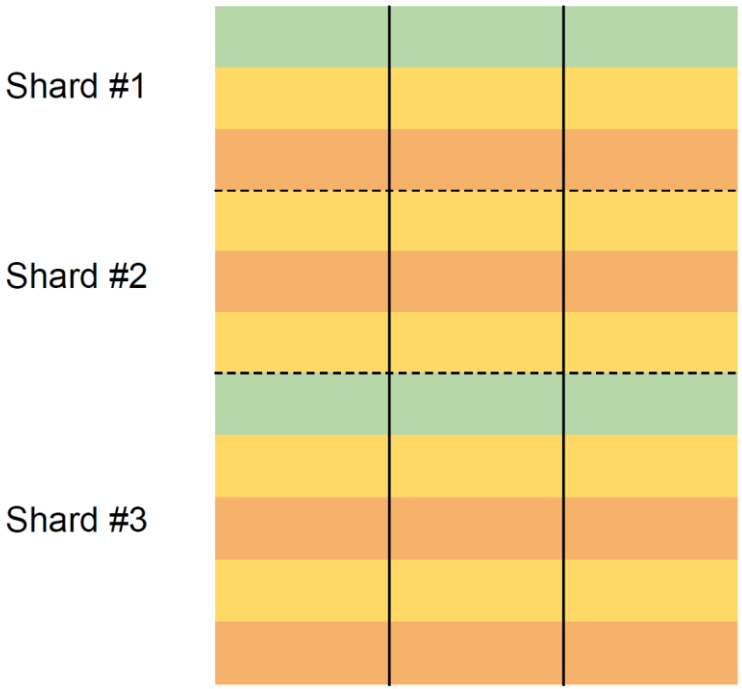
Source: Sebastian Kanthak (Google), Strange Loop Conference, St. Louis, MO, 2013

Physical data layout: interleaved tables



Source: Sebastian Kanthak (Google), Strange Loop Conference, St. Louis, MO, 2013

Sharding

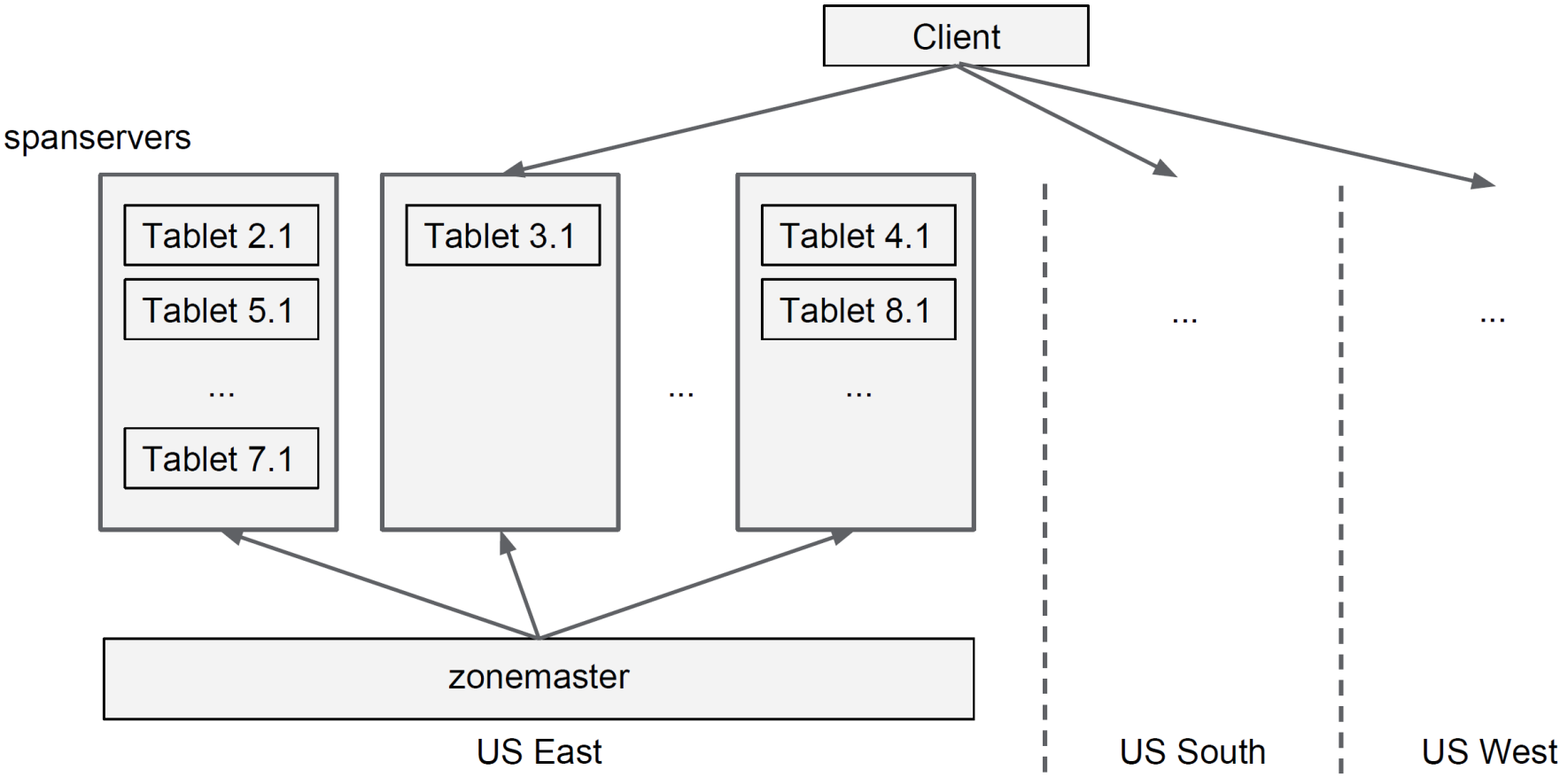


Cross-shard support for

- Transactions (read/write)
- Consistent (snapshot) reads

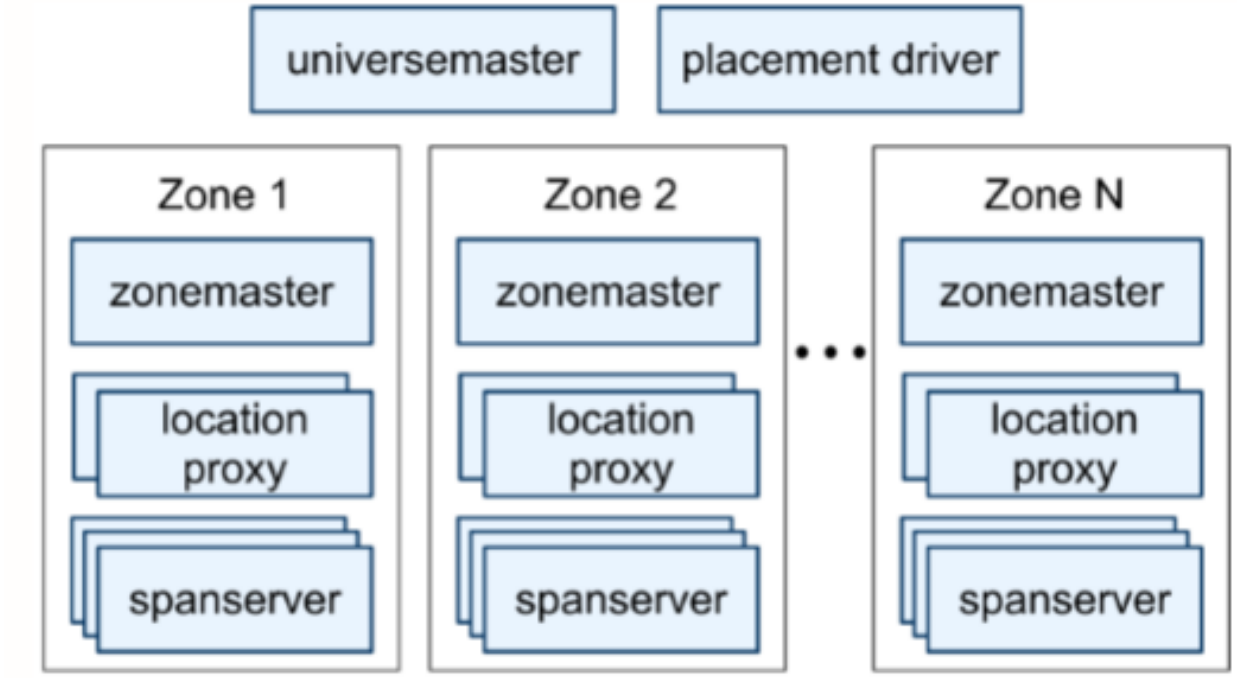
Source: Sebastian Kanthak (Google), Strange Loop Conference, St. Louis, MO, 2013

Serving structure

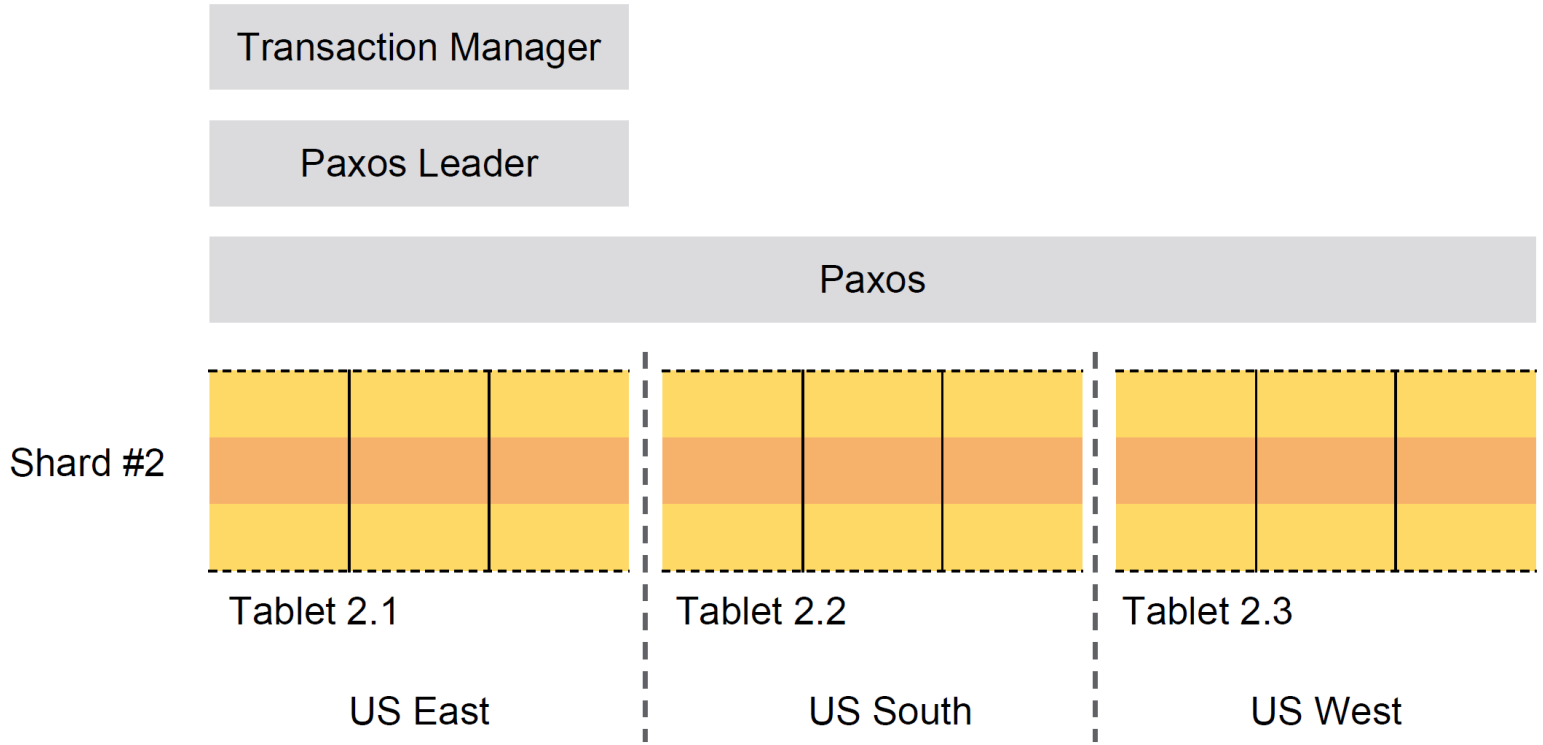


Source: Sebastian Kanthak (Google), Strange Loop Conference, St. Louis, MO, 2013

Spanner server organization

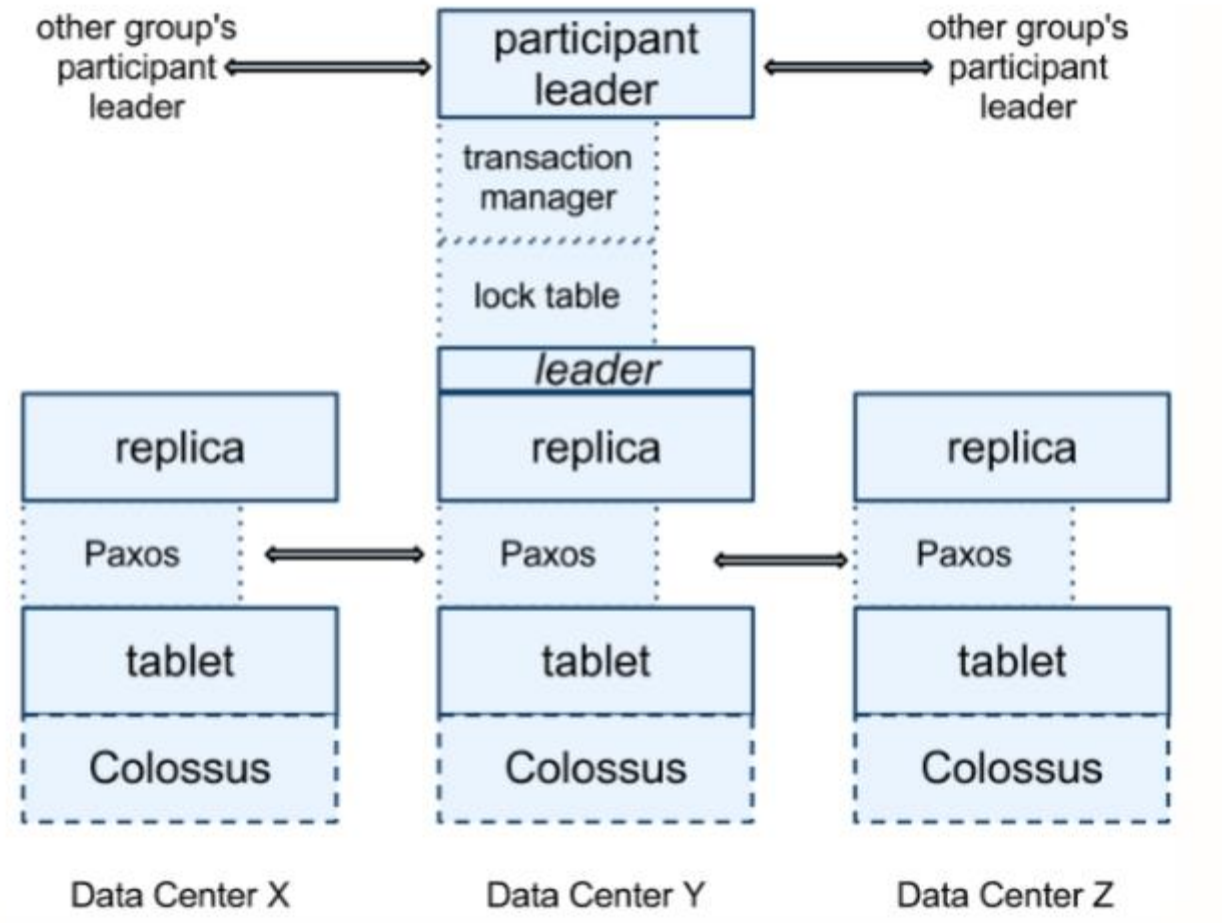


Replication

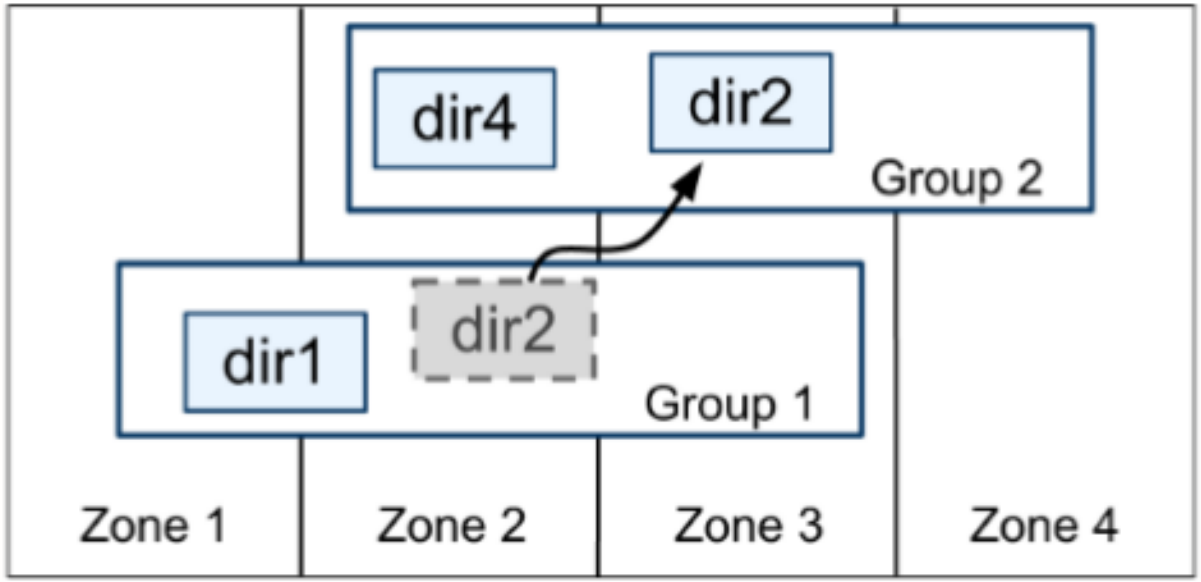


Source: Sebastian Kanthak (Google), Strange Loop Conference, St. Louis, MO, 2013

Spanner software stack



Directories (“buckets”) can be moved across groups



Key techniques

- Paxos for data replication of each tablet across zones
- Two-phase locking (2PL) for serializability
  - Transactions should acquire all locks they need before starting

Lock type	read-lock	write-lock
read-lock		X
write-lock	X	X

  - For performance: support read-only transactions without locks
  - Multi-value concurrency control (MVCC): timestamps/snapshots
  - Timestamps consistent with externally visible order
- Two-phase commit (2PC) for cross-table atomicity

Example: Ad System

Campaigns

campgain_id	keyword	bid
4	strange loop	$2.00

On US server

On EU server

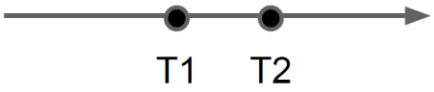
Impressions

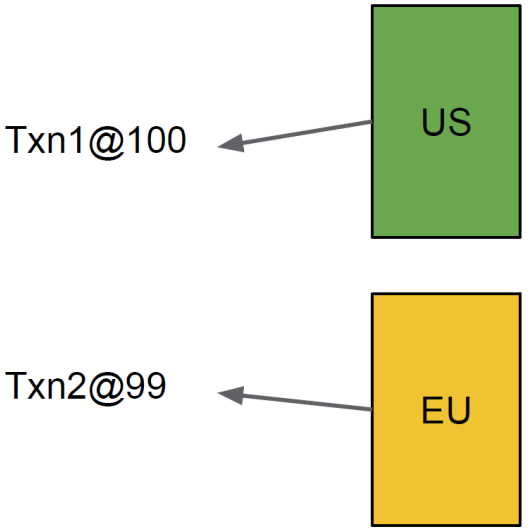
region	time	campaign_id	cost
US	2013/09/20-07...	4	$1.50
US			
...			
EU	2013/09/20-06...	4	$0.50
EU			

Source: Sebastian Kanthak (Google), Strange Loop Conference, St. Louis, MO, 2013

What can go wrong if using local timestamps

- Txn 1 creates a new ad on US server
- Ad serving system notified
- Ad server in Europe
- User clicks on ad
- Txn 2 logs click on EU server



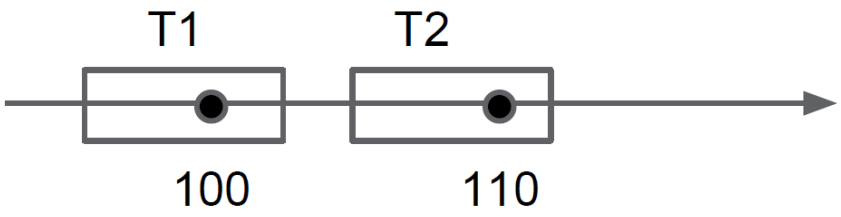


Invariant: Any snapshot that contains txn 2 should also contain txn 1

Source: Sebastian Kanthak (Google), Strange Loop Conference, St. Louis, MO, 2013

External consistency

- Assume T1 commits before T2 starts according to global wall-clock
- T1 should be serialized before T2
- T2's commit timestamp should be > T1's commit timestamp
- Must apply even if T1 and T2 do not conflict

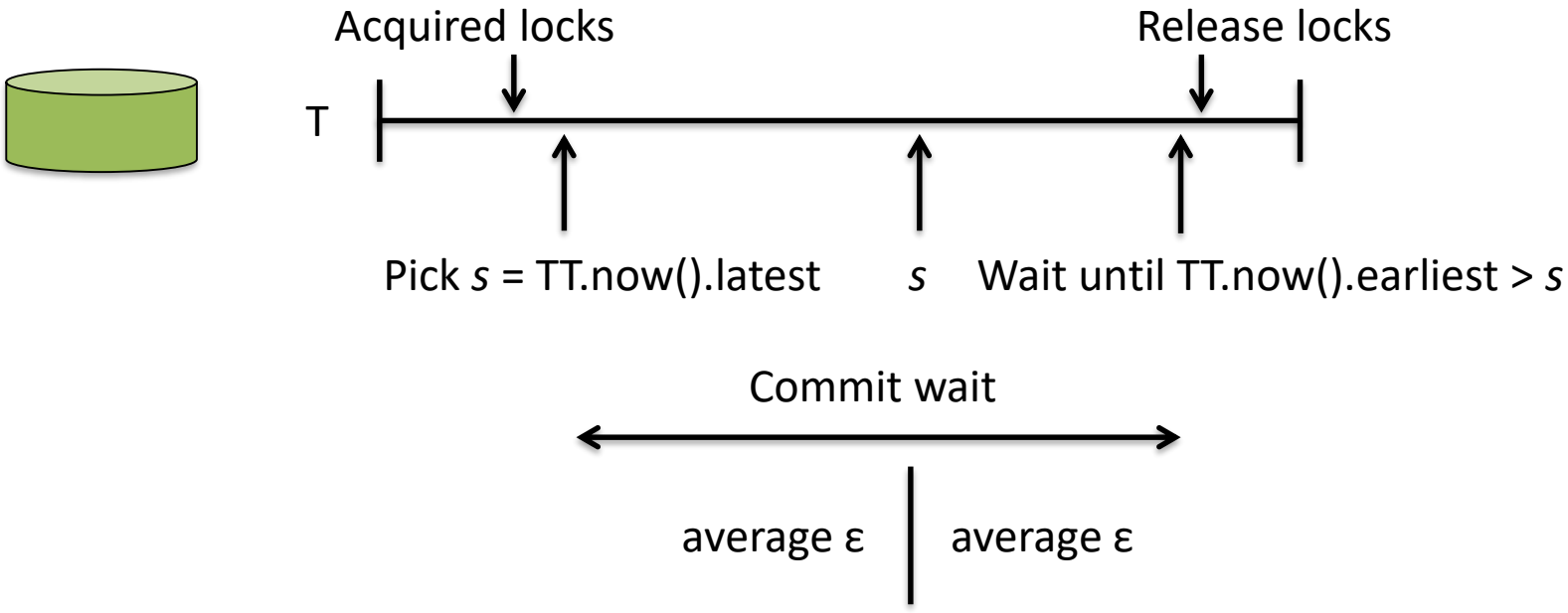


Source: Sebastian Kanthak (Google), Strange Loop Conference, St. Louis, MO, 2013

TrueTime API

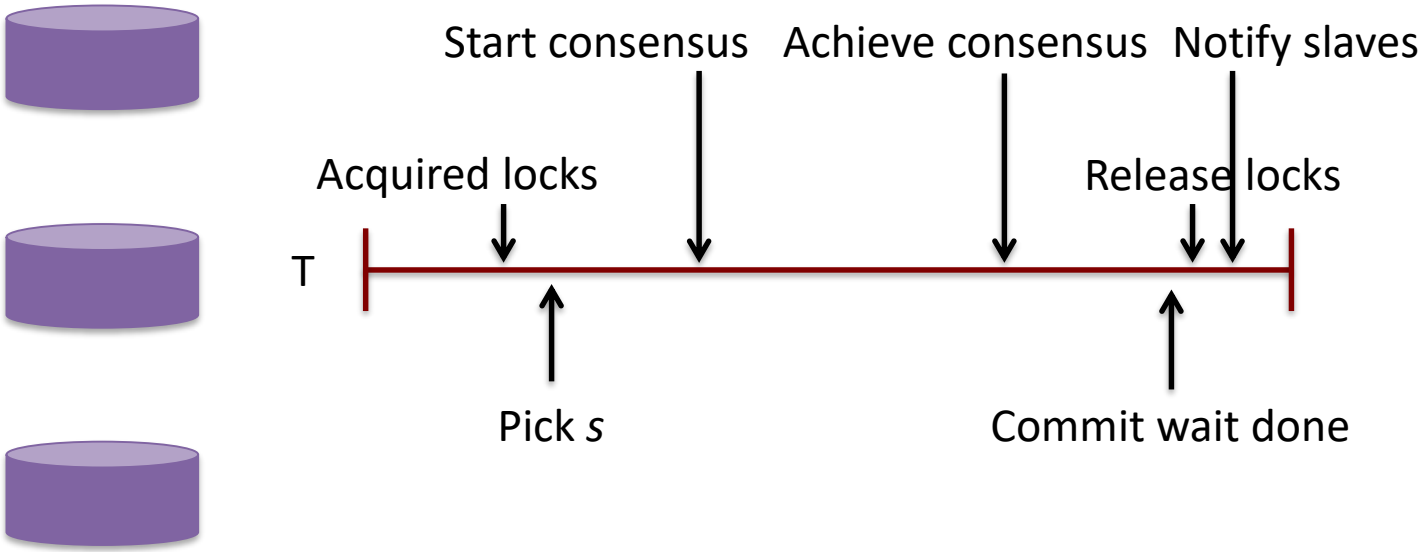
Method	Returns
$TT.now()$	*TTinterval*: $[earliest, latest]$
$TT.after(t)$	true if t has definitely passed
$TT.before(t)$	true if t has definitely not arrived

Picking commit timestamps



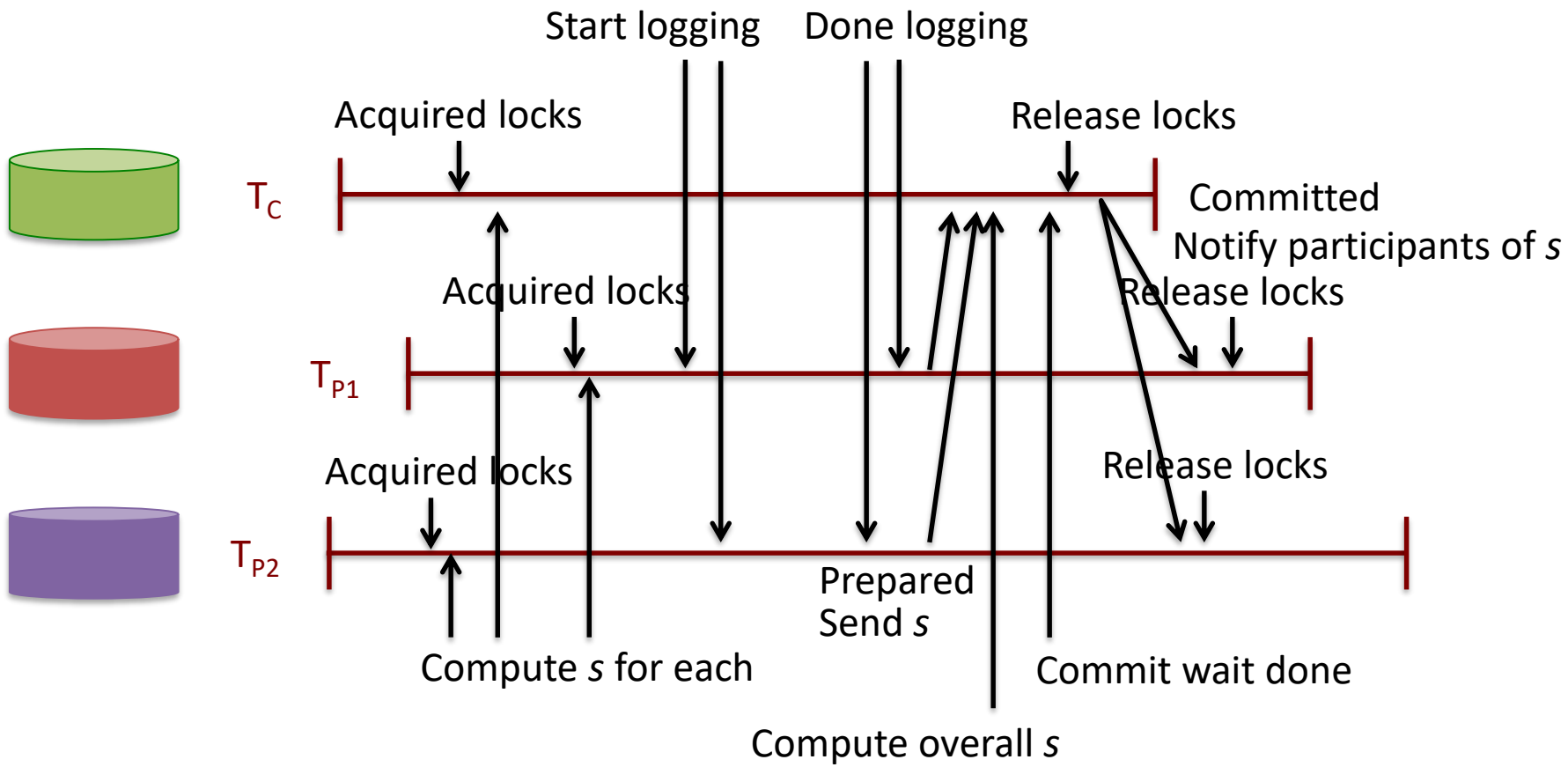
Source: Wilson Hsieh, Spanner: Google's Globally-Distributed Database, OSDI'12

Commit-wait and replication



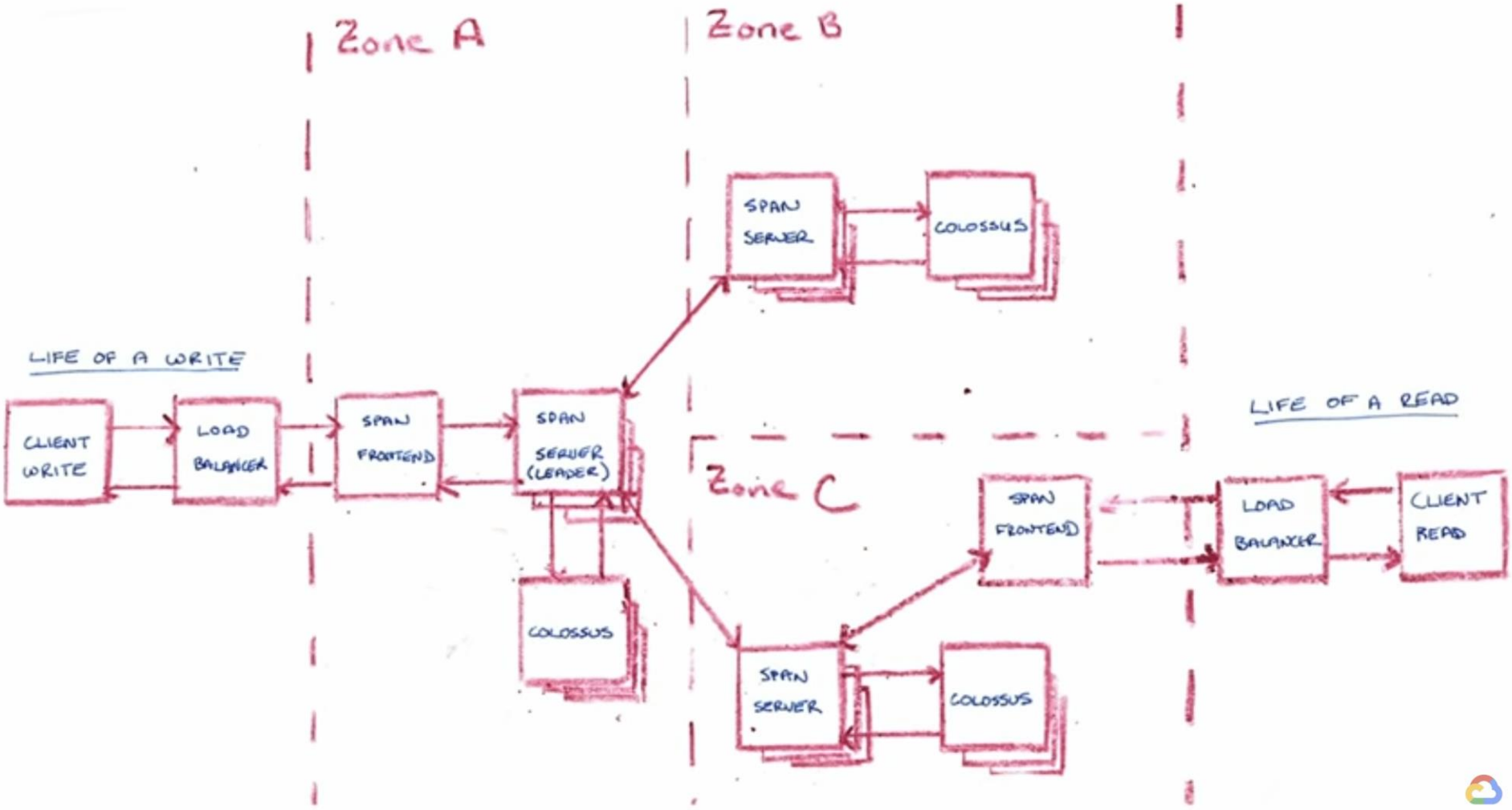
Source: Wilson Hsieh, Spanner: Google's Globally-Distributed Database, OSDI'12

Commit-wait and 2PC



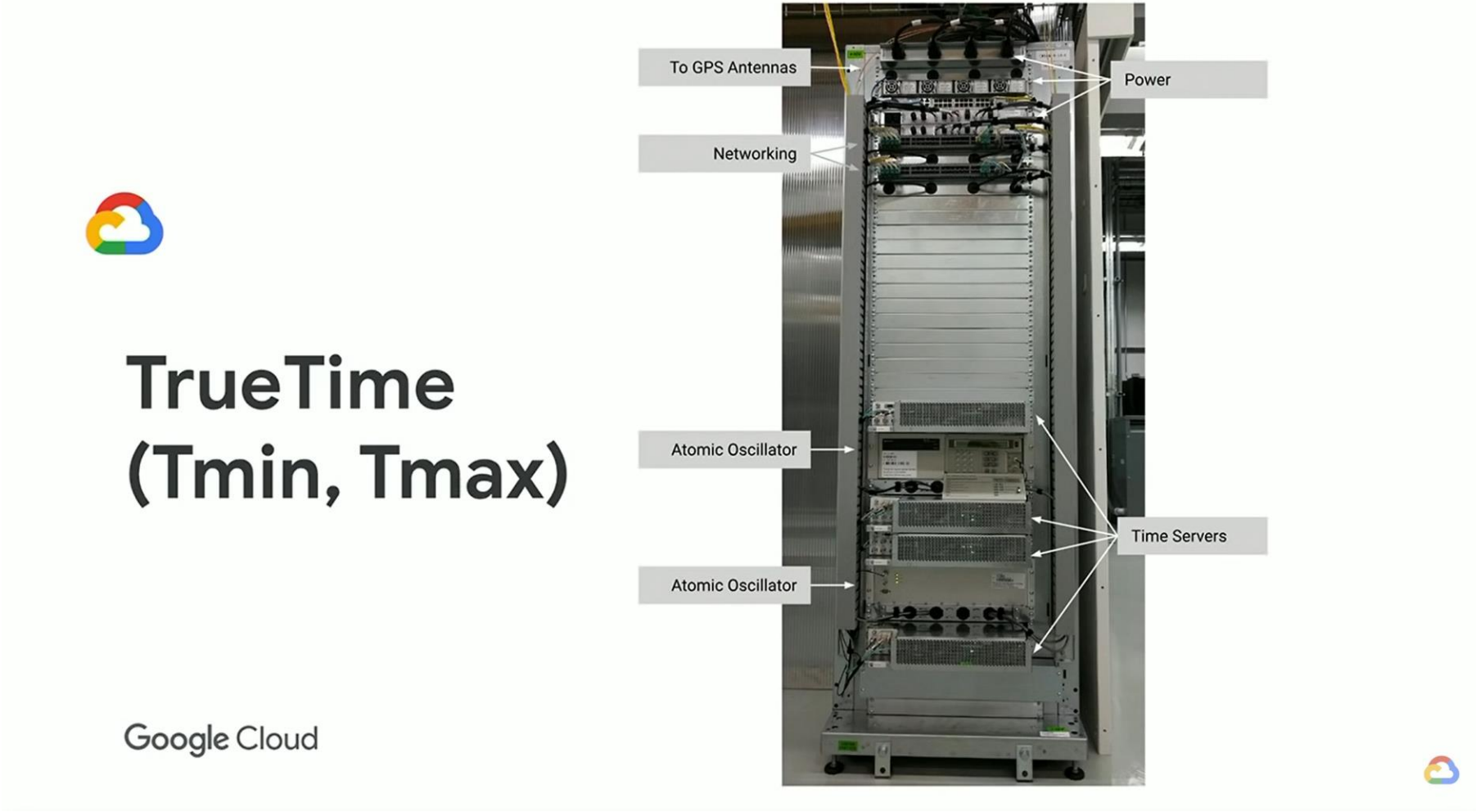
Source: Wilson Hsieh, Spanner: Google's Globally-Distributed Database, OSDI'12

Life of a read / write



Source: Andrew Fikes, Spanner Internals Part 1: What Makes Spanner Tick? (Cloud Next '19)

TrueTime servers



Source: Spanner Internals Part 1: What Makes Spanner Tick? (Cloud Next '19)