CockroachDB: The Resilient Geo-Distributed SQL Database

Rebecca Taft, Irfan Sharif, Andrei Matei, Nathan VanBenschoten, Jordan Lewis, Tobias Grieger, Kai Niemi, Andy Woods, Anne Birzin, Raphael Poss, Paul Bardea, Amruta Ranade, Ben Darnell, Bram Gruneir, Justin Jaffray, Lucy Zhang, and Peter Mattis

sigmod2020@cockroachlabs.com
Cockroach Labs, Inc.
CockroachDB: The Resilient Geo-Distributed SQL Database

Rebecca Taft, Irfan Sharif, Andrei Matei, Nathan VanBenschoten, Jordan Lewis, Tobias Grieger, Kai Niemi, Andy Woods, Anne Birzin, Raphael Poss, Paul Bardea, Amruta Ranade, Ben Darnell, Bram Gruneir, Justin Jaffray, Lucy Zhang, and Peter Mattis

sigmod2020@cockroachlabs.com
Cockroach Labs, Inc.
CockroachDB

1. Architecture
2. Transactions
3. Clocks
4. Learnings
CockroachDB -- Architecture

1. Geo-distributed
2. Multi-cloud
3. Monolithic KC-store
4. Transactional
# CockroachDB - Architecture

## Raft Protocol Summary

### Followers
- Respond to RPCs from candidates and leaders.
- Convert to candidate if election timeout elapses without either:
  - Receiving valid AppendsEntries RPC, or
  - Granting vote to candidate

### RequestVote RPC
- Invoked by candidates to gather votes.
- **Arguments:**
  - candidateId
  - term
  - lastLogIndex
  - lastLogTerm
- **Results:**
  - currentTerm, for candidate to update itself
  - voteGranted
- **Implementation:**
  1. If term > currentTerm, currentTerm ← term (step down if leader or candidate)
  2. If term == currentTerm, votedFor is null or candidateId, and candidate’s log is at least as complete as local log, grant vote and reset election timeout

### Candidates
- Increment currentTerm, vote for self
- Reset election timeout
- Send RequestVote RPCs to all other servers, wait for either:
  - Votes received from majority of servers: become leader
  - AppendsEntries RPC received from new leader: step down
- Election timeout elapses without election resolution:
  - Increment term, start new election
- Discover higher term: step down

### Leaders
- Initialize nextIndex for each to last log index + 1
- Send initial empty AppendsEntries RPCs (heartbeat) to each follower; repeat during idle periods to prevent election timeouts
- Accept commands from clients, append new entries to local log
- Whenever last log index ≥ nextIndex for a follower, send AppendsEntries RPC with log entries starting at nextIndex, update nextIndex if successful
- If AppendsEntries fails because of log inconsistency, decrement nextIndex and retry

### AppendEntries RPC
- Invoked by leader to replicate log entries and discover inconsistencies; also used as heartbeat.
- **Arguments:**
  - term
  - leaderId
  - prevLogIndex
  - prevLogTerm
  - entries[]
  - commitIndex
- **Results:**
  - **Argument:**
  - leaderId
  - prevLogIndex
  - prevLogTerm
  - entries[]
  - commitIndex
CockroachDB - Architecture

RocksDB is a log-structured merge tree (LSM)

1. Writes are buffered in RAM in a “memtable”
2. Many megabytes of values are written to a new file at once
3. Files are immutable, and store keys in sorted order
4. “Compaction” creates new files by merging several old files, making things more sorted and removing duplicates

Big writes
CockroachDB -- Transactions

Transaction Start
TS=ta

Uncommitted Write
TS= tb (> ta)

Transaction Read Must wait
CockroachDB -- Transactions

Transaction Start
TS=ta

Committed Write
TS= tb (> ta)

Read
Update TS=tc (>tb)
CockroachDB -- Transactions

Transaction Start
TS=ta

Uncommitted
Write
TS=tb (> ta)

Write
Must Wait

Committed
Write
TS=tb (> ta)
Commits
Possibly at TS=tc

Write
Assign TS=td (>tc)
CockroachDB -- Transactions

Transaction Start
TS=ta

Committed Write
TS= tb (> ta)

Read
Update TS=tc (>tb)
CockroachDB -- Clocks

1 Logical clocks (10 points)

- Mark on the figure Lamport’s logical clock timestamps of the events.
- Mark on the figure vector clock timestamps of the events.
- Write down all the consistent-cuts that pass through event k.
CockroachDB -- Lessons

1. Raft is chatty
2. Pause groups
3. Consolidate heartbeats
4. Joint Consensus
CockroachDB -- Lessons

**Isolation Level Hierarchy**

- **Serializable**
  - **Recoverable Reads**
  - **Cursor Stability**
- **Snapshot Isolation**
  - **Read Committed**
  - **Read Uncommitted**
CockroachDB -- Lessons

**ISOLATION LEVEL HIERARCHY**

- **Serializable**
  - **Repeatable Reads**
  - **Cursor Stability**
  - **Read Committed**
  - **Read Uncommitted**
- **Snapshot Isolation**
CockroachDB -- Lessons
CockroachDB -- Lessons
CockroachDB -- Questions