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

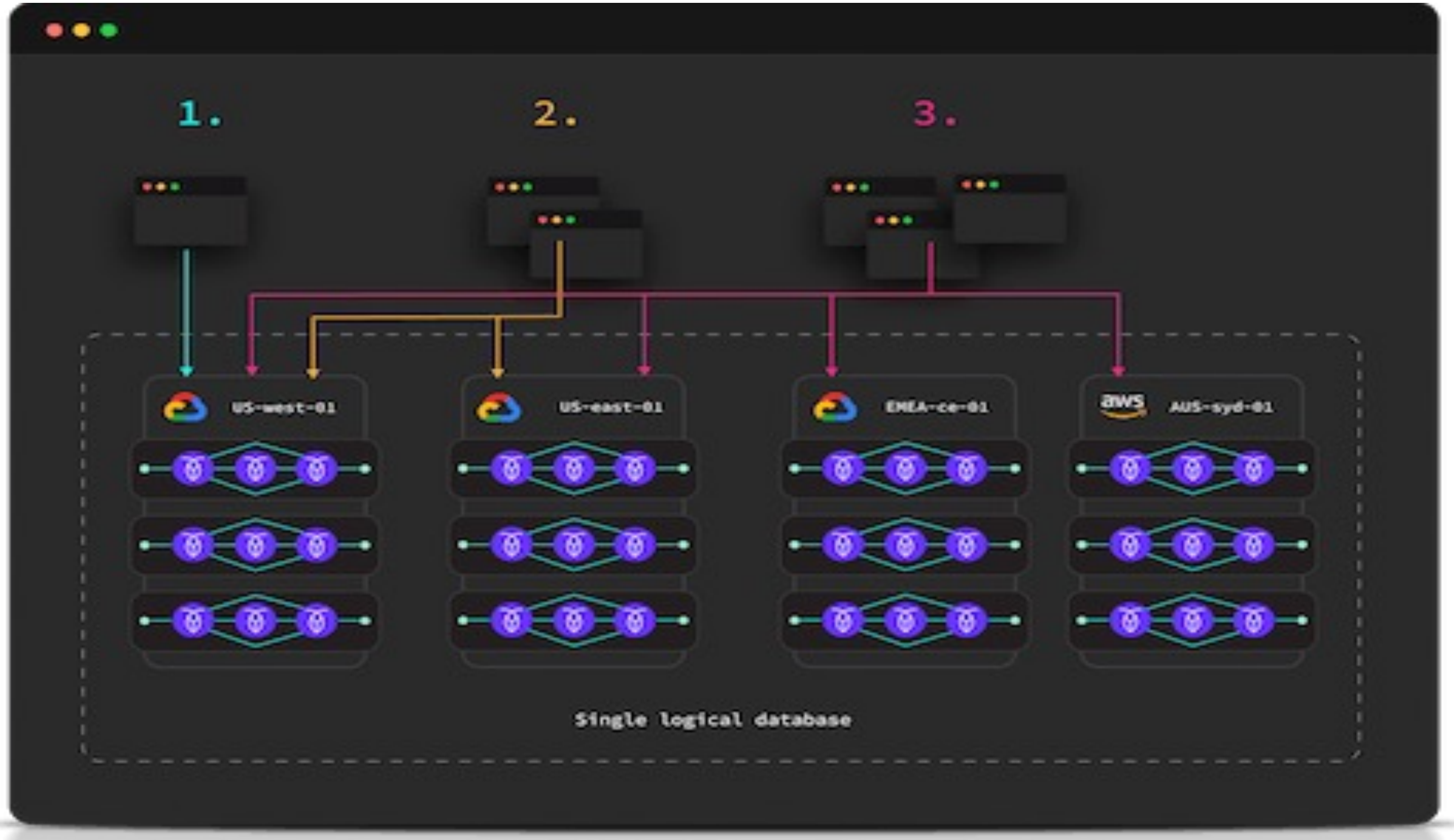
Rebecca Taft, Irfan Sharif, Andrei Matei, Nathan VanBenschoten, Jordan Lewis, Tobias Grieger, Kai Niemi, Andy Woods, Anne Birzin, Raphael Poss, Paul Bardea, Amruta Ranade, Ben Darnell, Bram Gruneir, Justin Jaffray, Lucy Zhang, and Peter Mattis
sigmod2020@cockroachlabs.com
Cockroach Labs, Inc.

CockroachDB

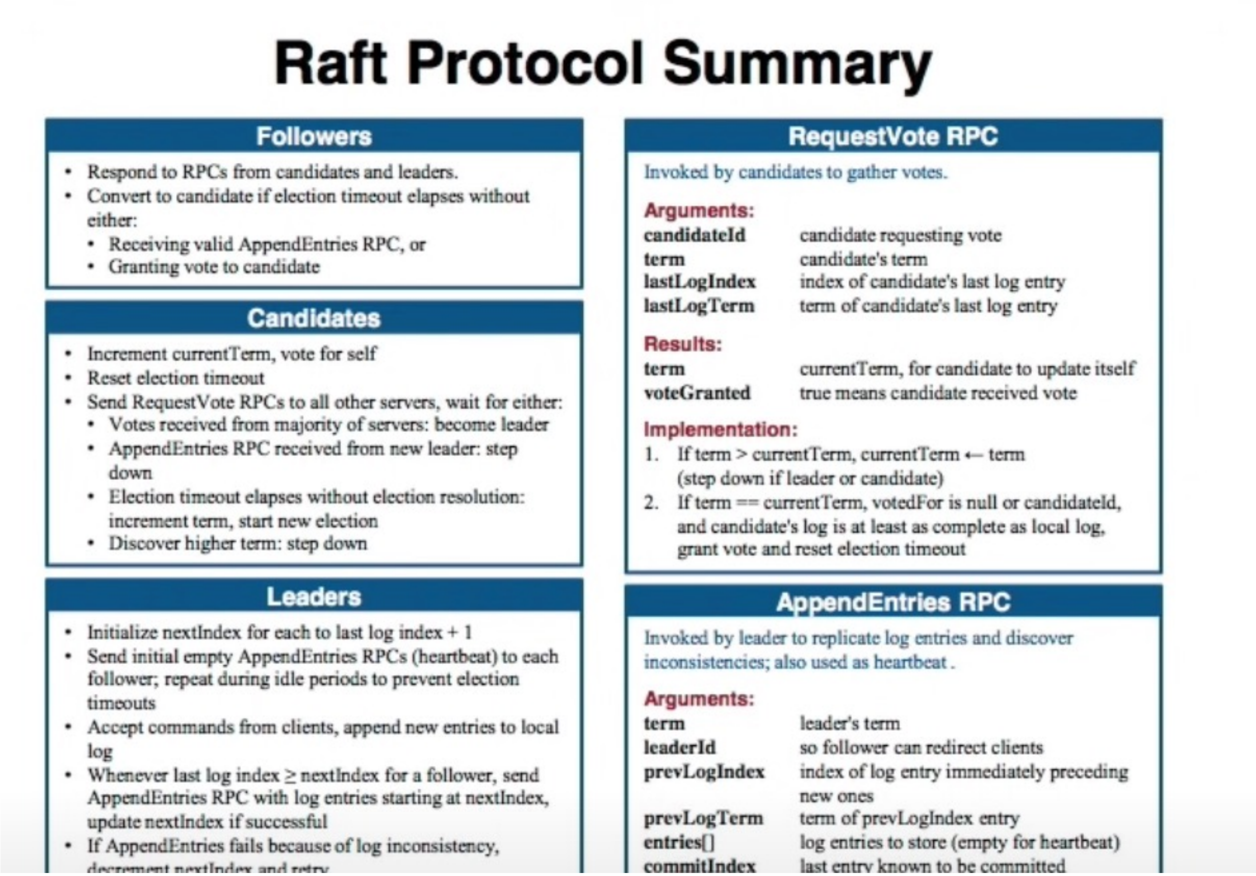
1. Architecture
2. Transactions
3. Clocks
4. Learnings

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 AppendEntries RPC, or
  - Granting vote to candidate

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
  - AppendEntries 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 AppendEntries 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 AppendEntries RPC with log entries starting at nextIndex, update nextIndex if successful
- If AppendEntries fails because of log inconsistency, decrement nextIndex and retry

RequestVote RPC

Invoked by candidates to gather votes.

Arguments:

candidateId	candidate requesting vote
term	candidate's term
lastLogIndex	index of candidate's last log entry
lastLogTerm	term of candidate's last log entry

Results:

term	currentTerm, for candidate to update itself
voteGranted	true means candidate received vote

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

AppendEntries RPC

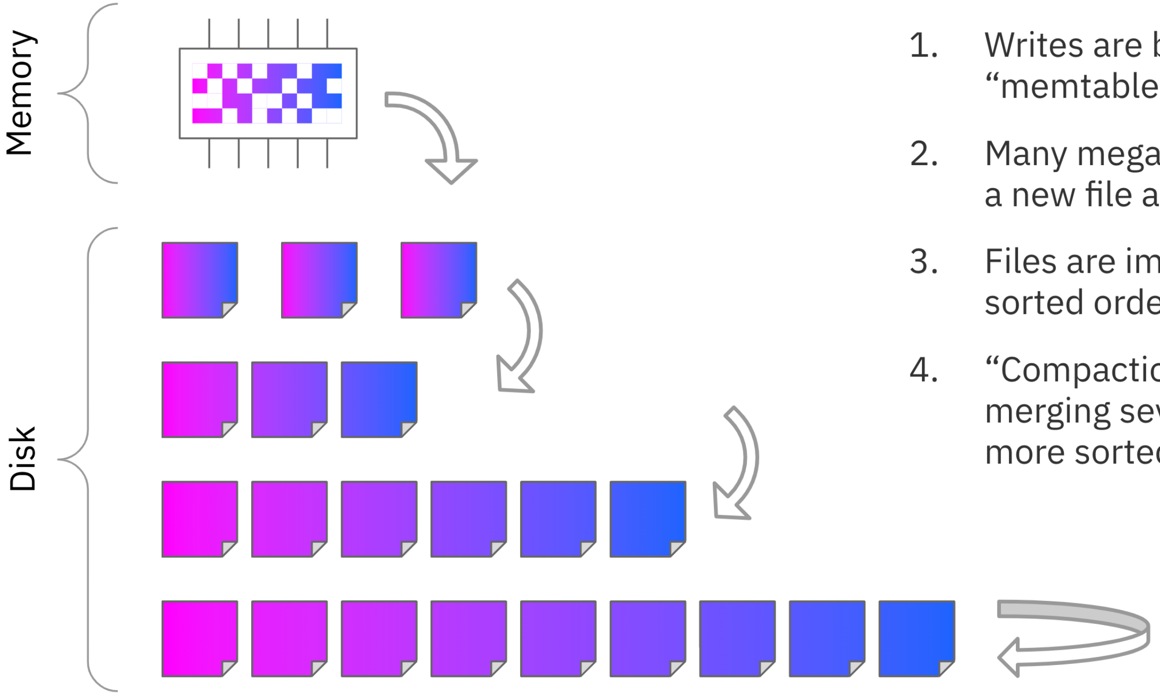
Invoked by leader to replicate log entries and discover inconsistencies; also used as heartbeat .

Arguments:

term	leader's term
leaderId	so follower can redirect clients
prevLogIndex	index of log entry immediately preceding new ones
prevLogTerm	term of prevLogIndex entry
entries[]	log entries to store (empty for heartbeat)
commitIndex	last entry known to be committed

CockroachDB - Architecture

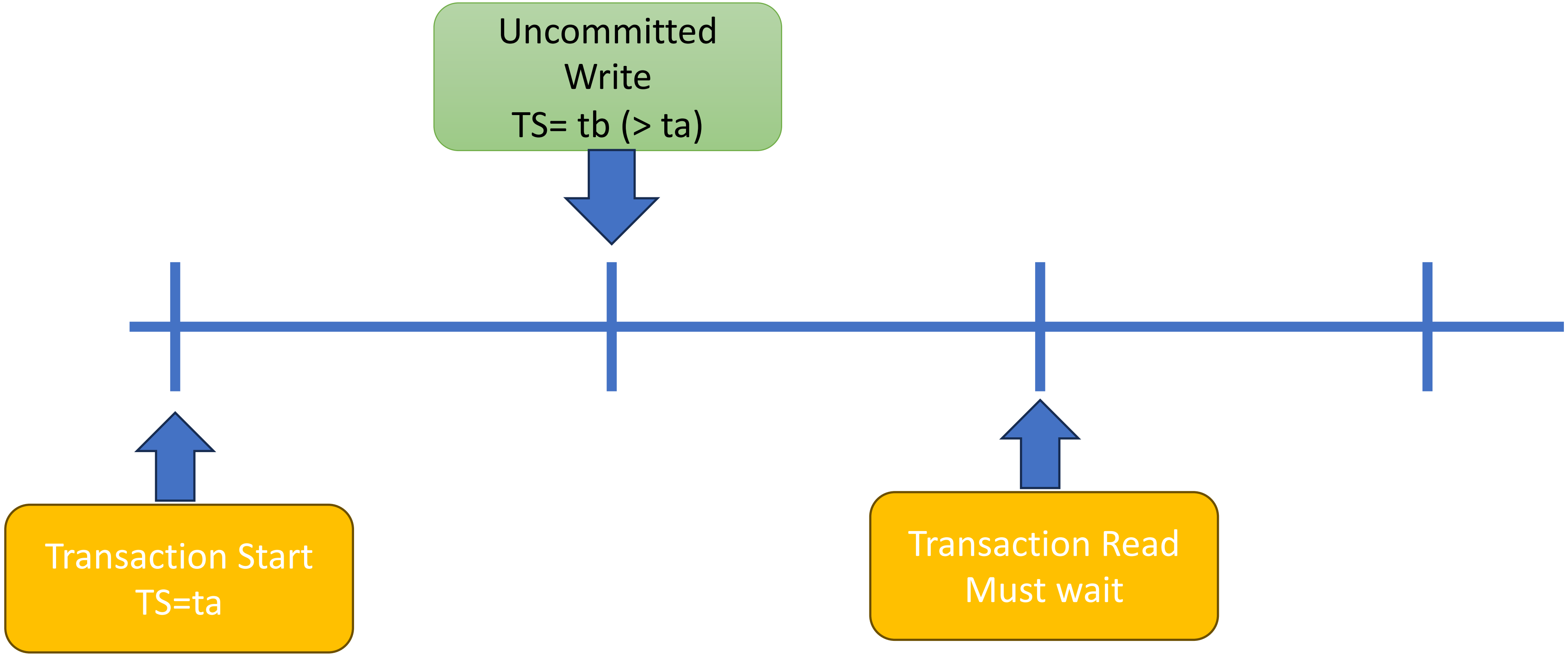
RocksDB is a log-structured merge tree (LSM)



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

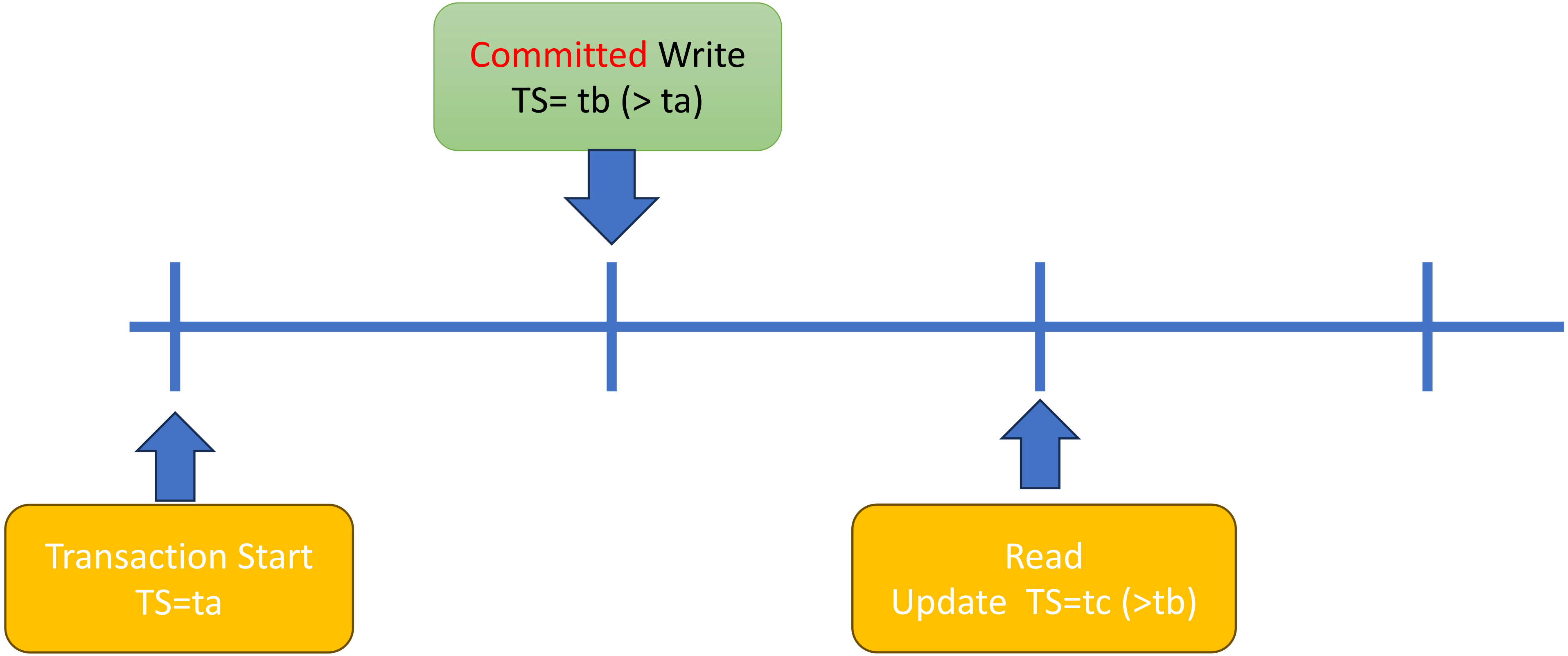


Uncommitted Write TS= tb (> ta)

Transaction Start TS=ta

Transaction Read Must wait

CockroachDB -- Transactions

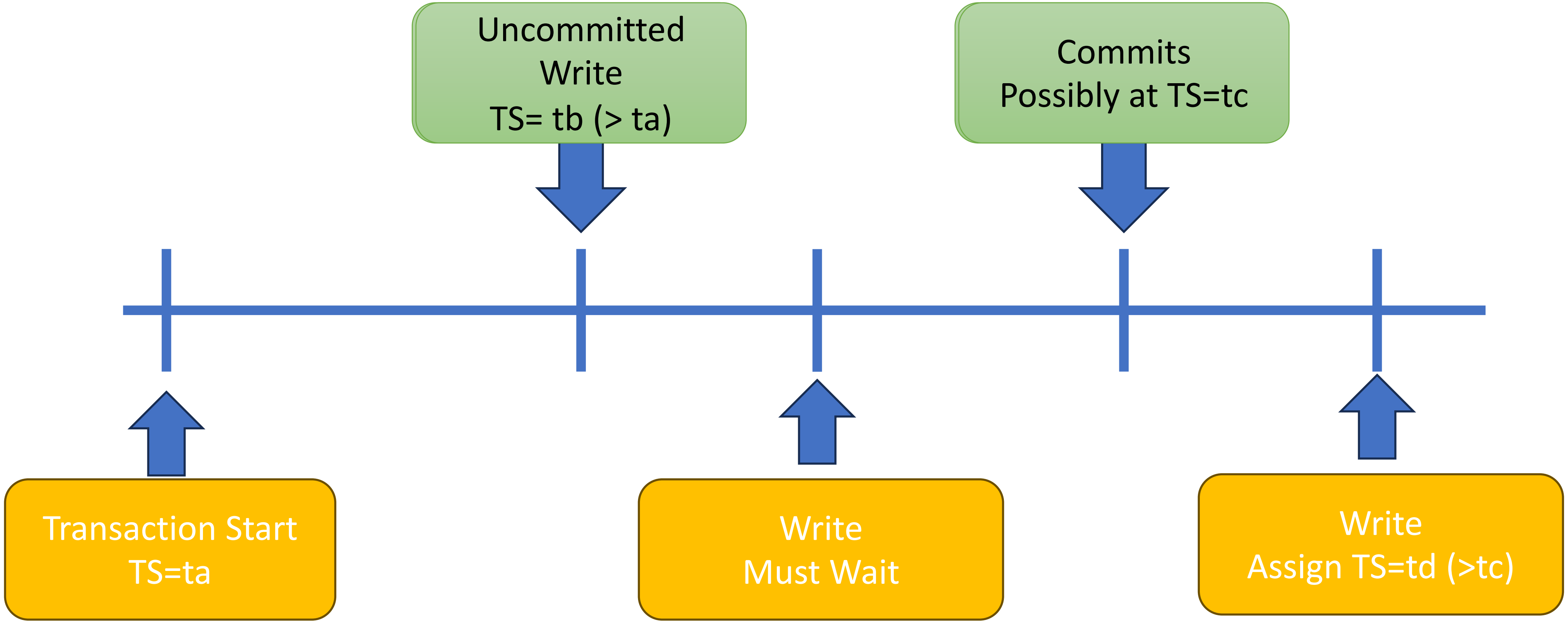


Committed Write
TS= tb (> ta)

Transaction Start
TS=ta

Read
Update TS=tc (>tb)

CockroachDB -- Transactions



Uncommitted Write TS= tb (> ta)

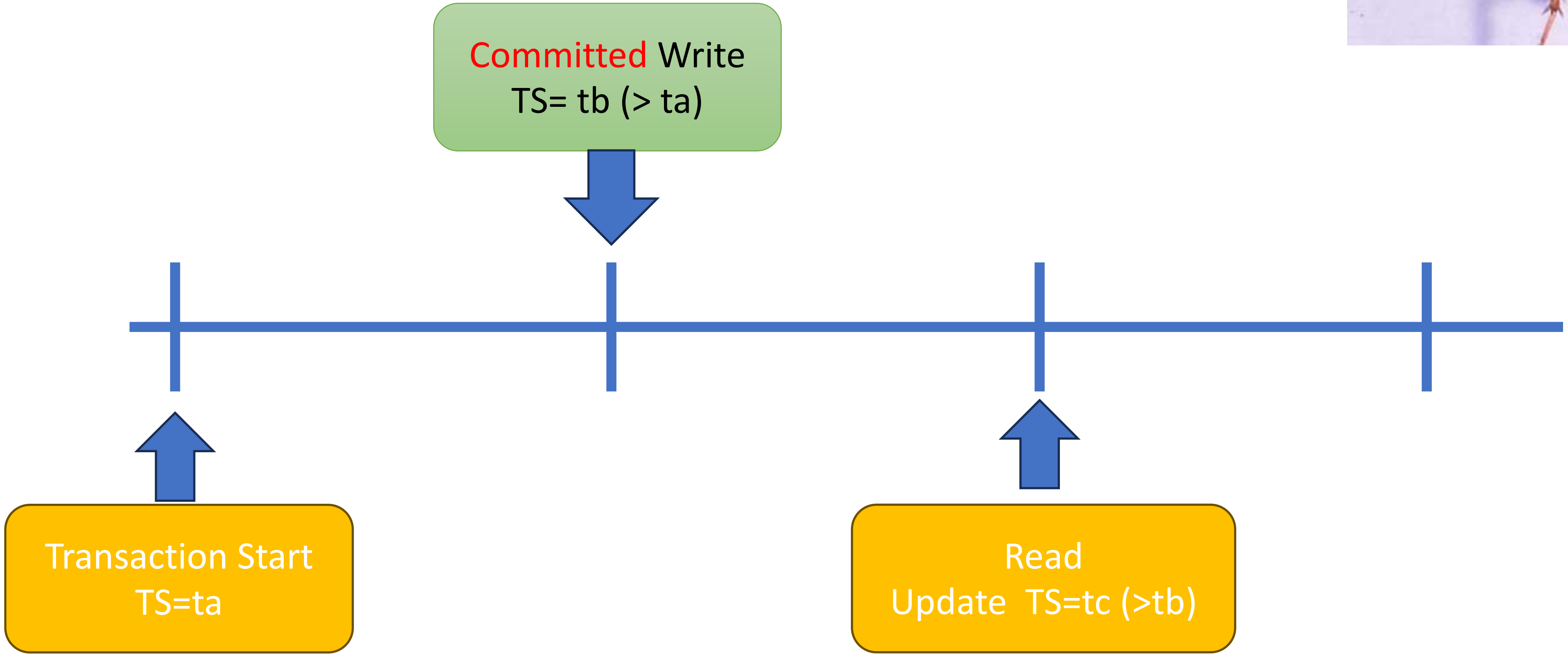
Commits Possibly at TS=tc

Transaction Start TS=ta

Write Must Wait

Write Assign TS=td (>tc)

CockroachDB -- Transactions

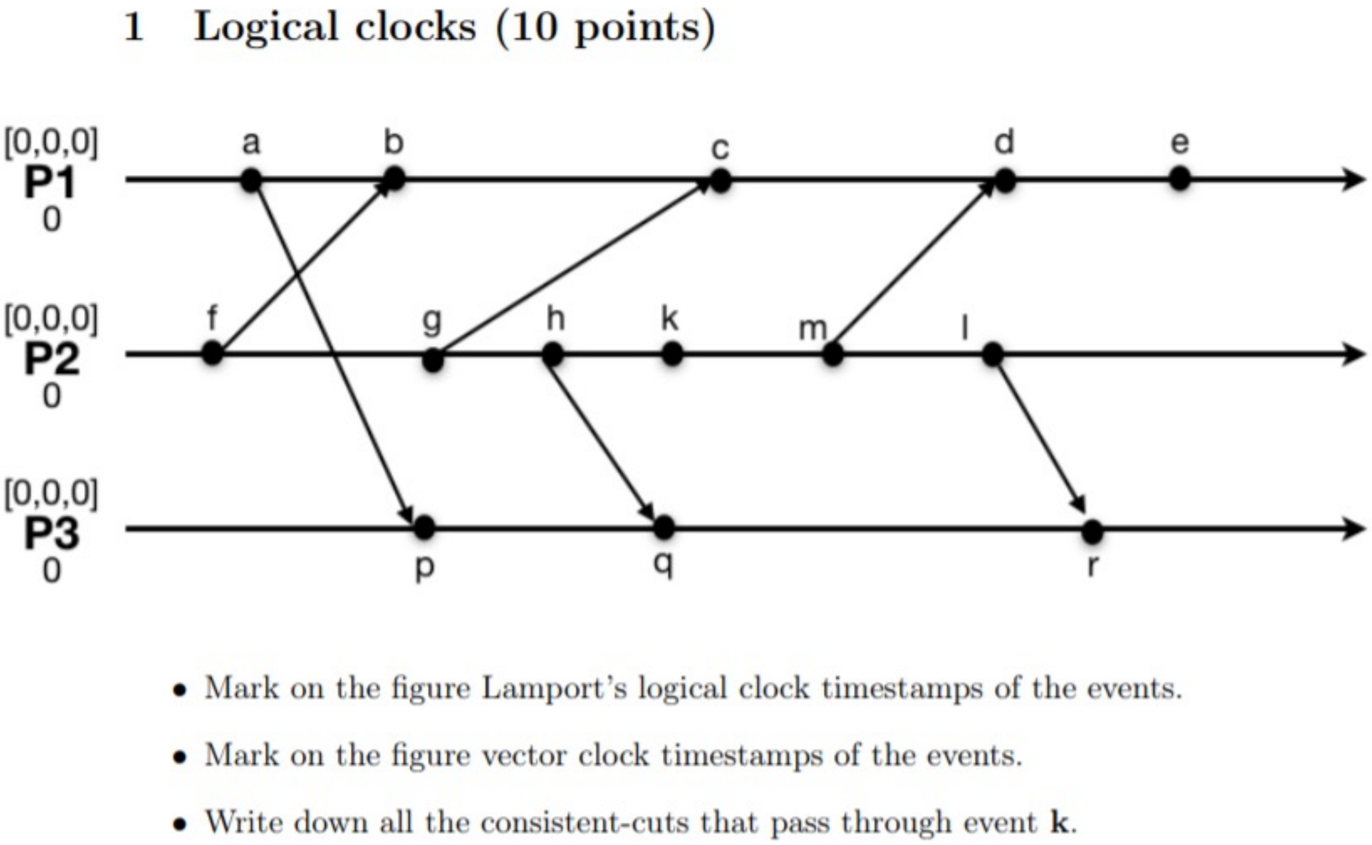


Committed Write
TS= tb (> ta)

Transaction Start
TS=ta

Read
Update TS=tc (>tb)

CockroachDB -- Clocks

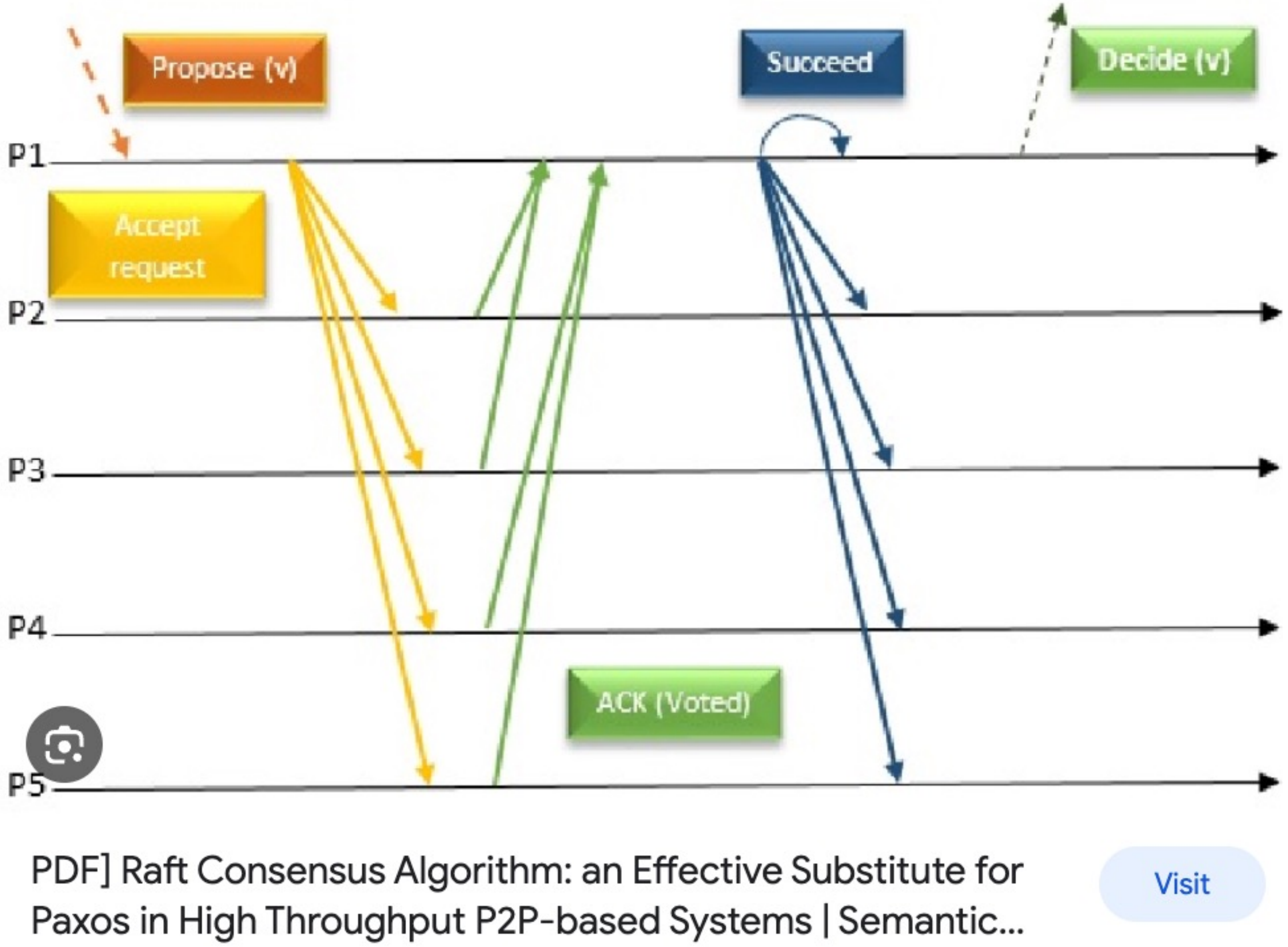


1 Logical clocks (10 points)

- 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**.

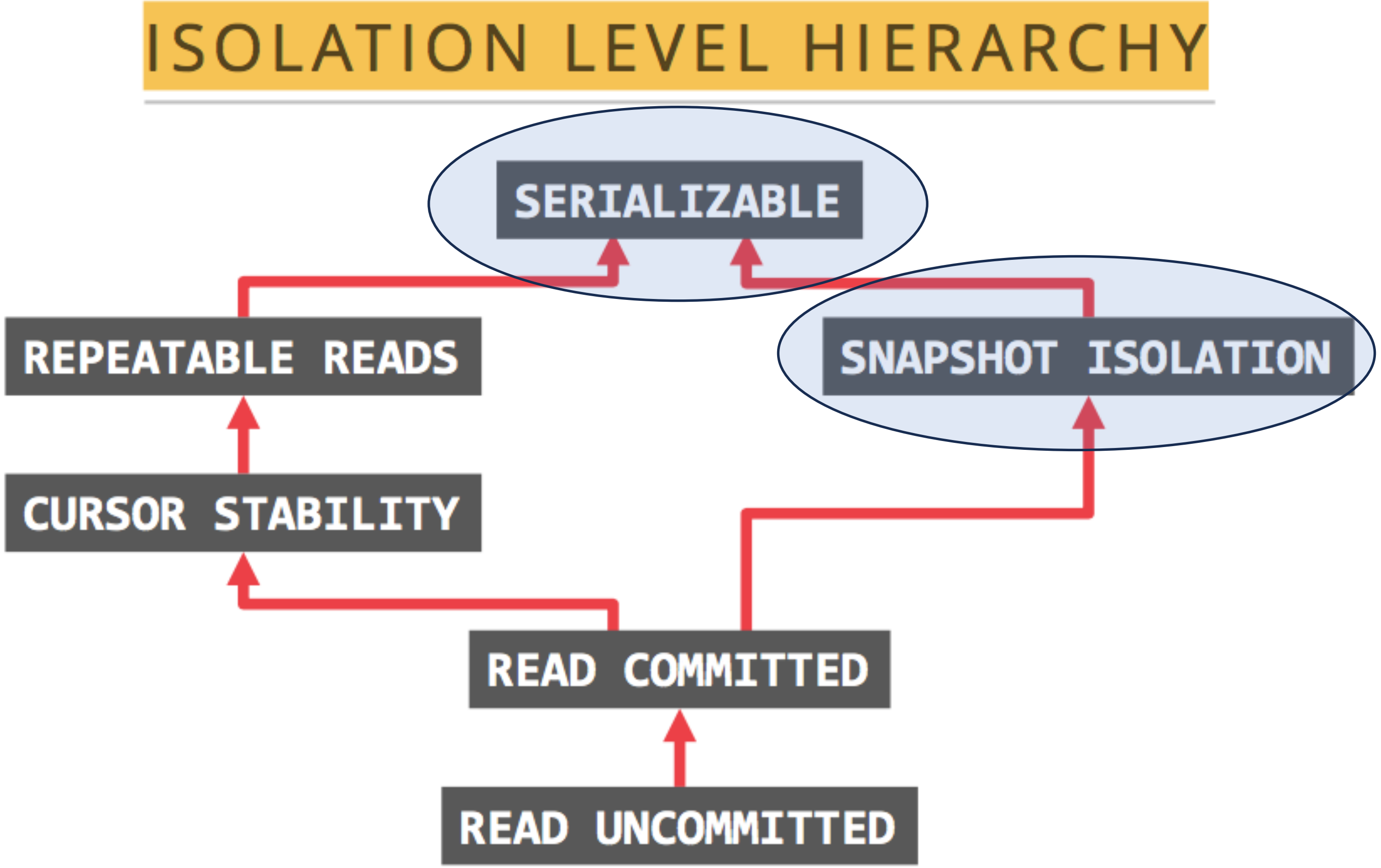
CockroachDB -- Lessons

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

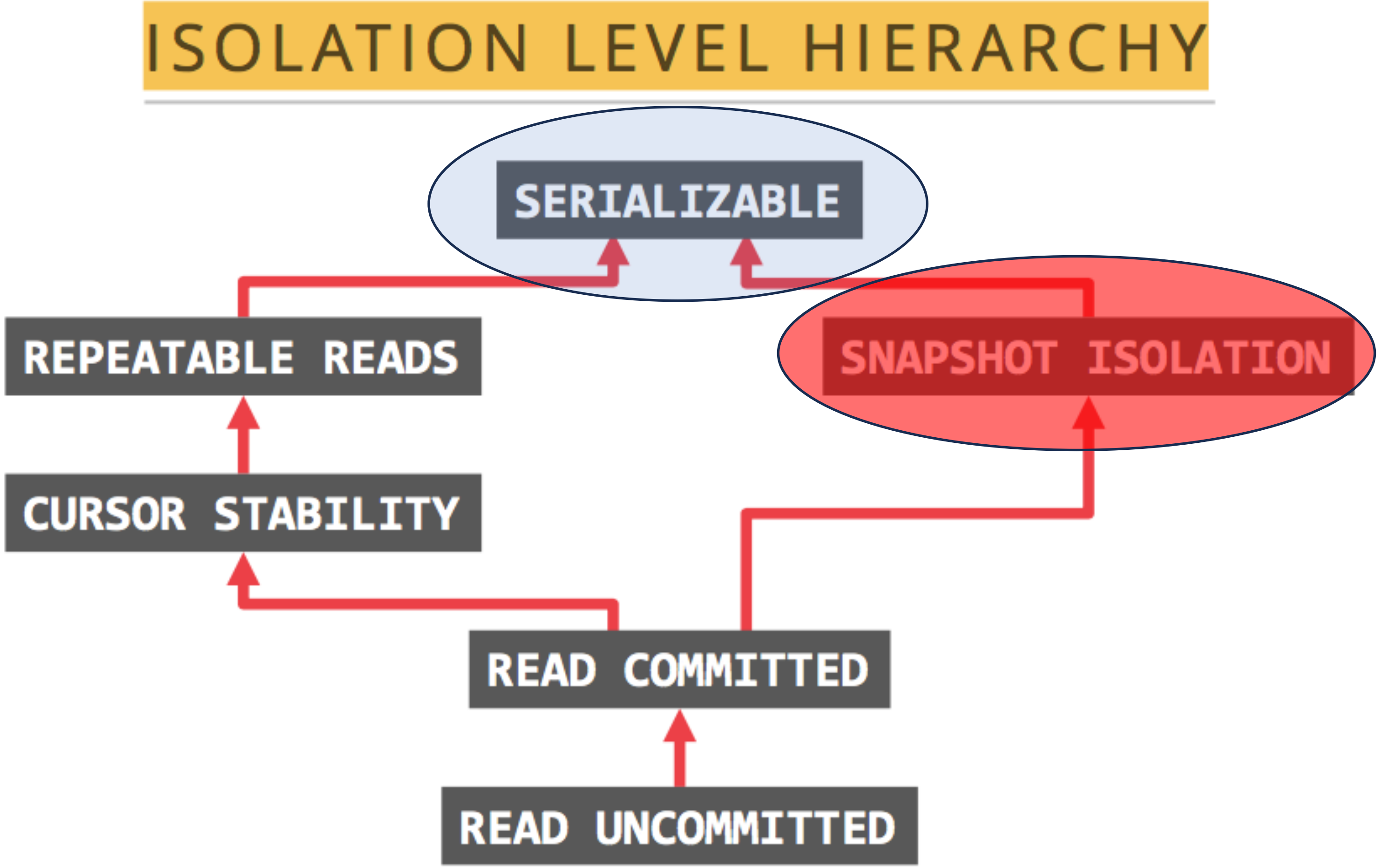


PDF] Raft Consensus Algorithm: an Effective Substitute for Paxos in High Throughput P2P-based Systems | Semantic...

CockroachDB -- Lessons



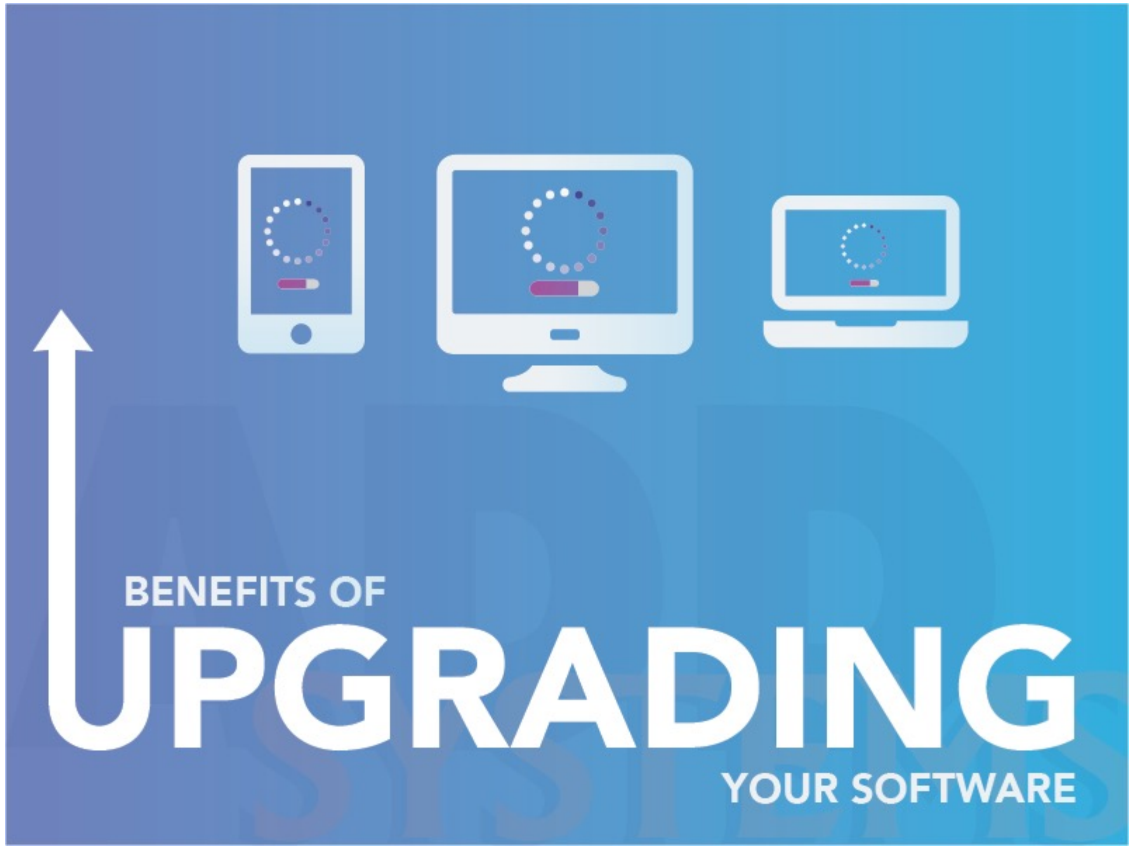
CockroachDB -- Lessons



CockroachDB -- Lessons

PostgreSQL

CockroachDB -- Lessons



CockroachDB -- Questions

CockroachDB