بسم الله الرحمن الرحیم

تکنولوژی کامپیوتر

جلسه‌ی بیست و یکم
کاساندرا

جلسه گذشته

جلسه‌ی جدید

CASSANDRA تاریخچه

Cassandra History

- Summer 2007: start a distributed inbox search on Facebook
- July 2008: Facebook publishes the code under the name Cassandra
- March 2009: The project enters the Apache Incubator
- Sept 2013 – v2.0
- 2021 - 4.0 GA — “Stability First”
- 2024 | 5.0 GA — AI-era Features

Cassandra

- Full multi-primary database replication
- Global availability at low latency
- Scaling out on commodity hardware
- Linear throughput increase with each additional processor
- Online load balancing and cluster growth
- Partitioned key-oriented queries
- Flexible schema

Cassandra

- Distribution & scale-out
  - *Master-less, shared-nothing architecture – every node is equal.*
  - *Virtual nodes (vnodes) slice the ring into hundreds of tiny token ranges, so adding/removing a node only streams a small fraction of data.*
  - *Elastic multi-datacenter replication with rack-aware placement*

Cassandra

- High availability
  - *No single point of failure; automatic fail-over.*
  - *CAP Theorem: AP but tunable consistency*
- Consistency options
  - *Tunable consistency on every read/write (ANY → ALL)*
  - *Lightweight transactions (LWT) with Paxos for linearizable compare-and-set updates*

CASE STUDIES

https://cassandra.apache.org/_/case-studies.html

Ably

Apache Cassandra is trusted to scale at internet level and designed to scale without limits. Which is why, at Ably Realtime, we use Cassandra for our persistent storage of messages.

READ MORE

Activision

Activision built a new system to message players with highly personalised communication. It used large amounts of real-time data and was built with Apache Cassandra

READ MORE

AdStage

AdStage is constantly monitoring performance trends and optimizing campaigns on behalf of advertisers. Apache Cassandra delivered the operational speed the company needed to ensure that the platform has low latency and the required throughput.

READ MORE

Airship

Where we originally stored device data in a set of Postgres shards, our scale quickly outpaced our capacity to add new shards, so we moved to a multiple database architecture using HBase and Cassandra.

READ MORE

Apple

Apple

A year ago, Apple said that it was running over 75,000 Cassandra nodes, storing more than 10 petabytes of data. At least one cluster was over 1,000 nodes, and Apple regularly gets millions of operations per second (reads/writes) with Cassandra.

READ MORE

Backblaze

Backblaze

We needed something that would handle really high write throughput and keep scaling on the write throughput. That forced us to look at distributed stores, and Apache Cassandra was the option that fitted what we needed.

READ MORE

bazaarvoice

Bazaarvoice

EmoDB is an open source RESTful data store built on top of Cassandra that stores JSON documents and, most notably, offers a databus that allows subscribers to watch for changes to those documents in real time.

READ MORE

BEST BUY

Best Buy

Best Buy uses Apache Cassandra to manage massive spikes in holiday traffic — 7x traffic spikes and bursts > 50,000 rps — and calls it “flawless.”

READ MORE

Discord

Cassandra was the only database that fulfilled all of Discord's requirements, as they can add nodes to scale it and it can tolerate a loss of nodes without any impact on the application. Related data is stored contiguously on disk providing minimum seeks and easy distribution around the cluster.

READ MORE

Dream11

The company started its operations in 2008 and started offering single match fantasy sports in 2012. It is India's Biggest Sports Gaming platform with users playing Fantasy Cricket, Football, Kabaddi, Basketball & Hockey. Dream11 is the Official Fantasy partner of the VIVO Indian Premier League (IPL), International Council of Cricket (ICC)

READ MORE

eBay

A glimpse on our Cassandra deployment: Dozens of nodes across multiple clusters 200 TB+ storage provisioned 400M+ writes & 100M+ reads per day, and growing QA, LnP, and multiple Production clusters.

READ MORE

Equinix

Equinix uses Cassandra for its ease of operation, and always-on node architecture — and its peer-to-peer architecture guarantees no single point of failure to collect and store streaming data from infrastructure instruments.

READ MORE

Instana

"Cassandra works well; it runs really nicely and smoothly. We've never lost data, and things are easy to fix. Quite frankly, without Cassandra, we couldn't run Instana."

READ MORE

Instagram

Constant At Instagram, we have one of the world's largest deployments of the Apache Cassandra database. We began using Cassandra in 2012 to replace Redis and support product use cases like fraud detection, Feed, and the Direct inbox.

READ MORE

Intuit Mint

Mint Bills selected Apache Cassandra to store user account data. "When you are selecting between accounts on your Mint Bills app, you are actually retrieving information from Cassandra directly," Csasznik-Shaked added

READ MORE

Intuit Turbo Tax

Intuit is supporting over 42,000 Peak TPS in production in AWS, over eight clusters in production. Cassandra has to process massive amounts of data, such as entitlements, tax returns, filings, user experience, and everything needed to support TurboTax.

READ MORE

DATA STRUCTURE

Commit Log

SSTables (Sorted String Table)

Memtable

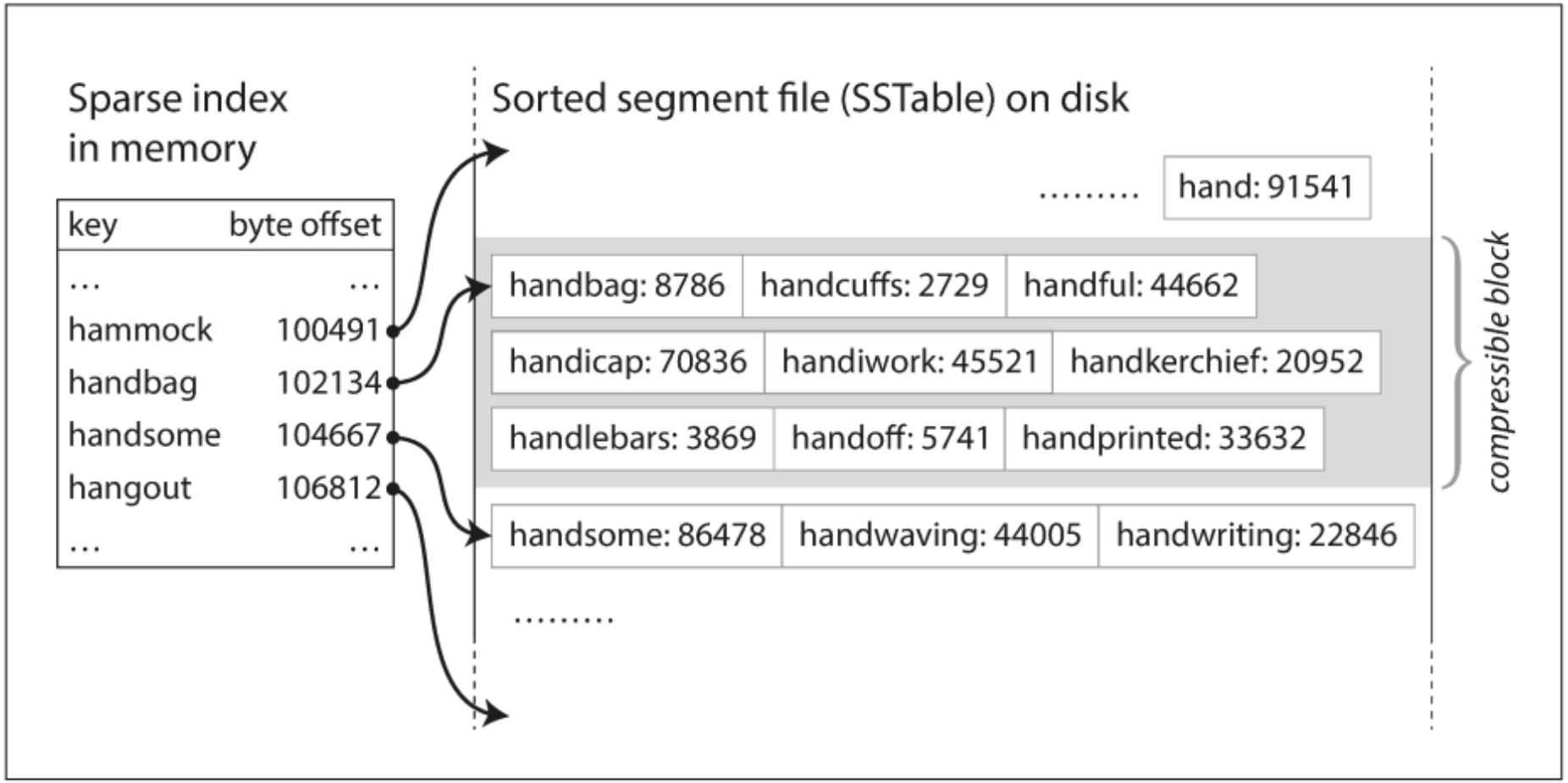
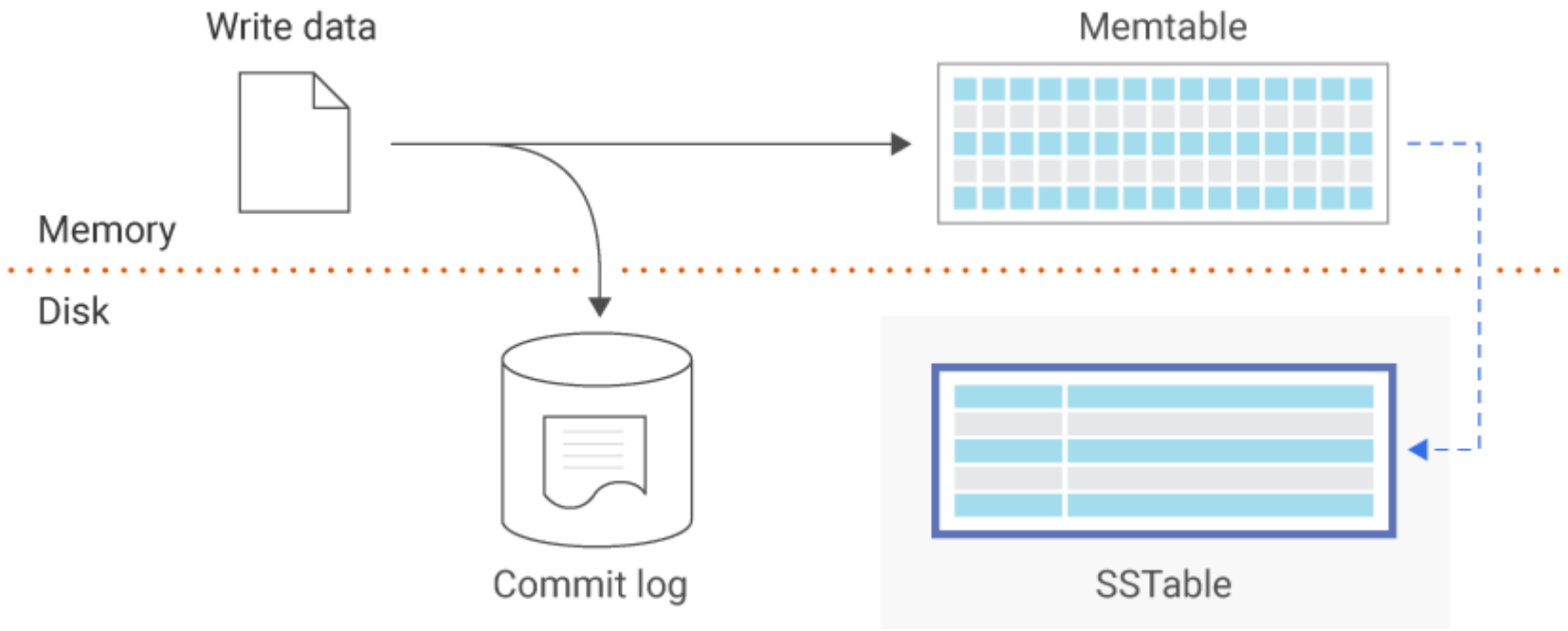


Figure 3-5. An SSTable with an in-memory index.

Memtable

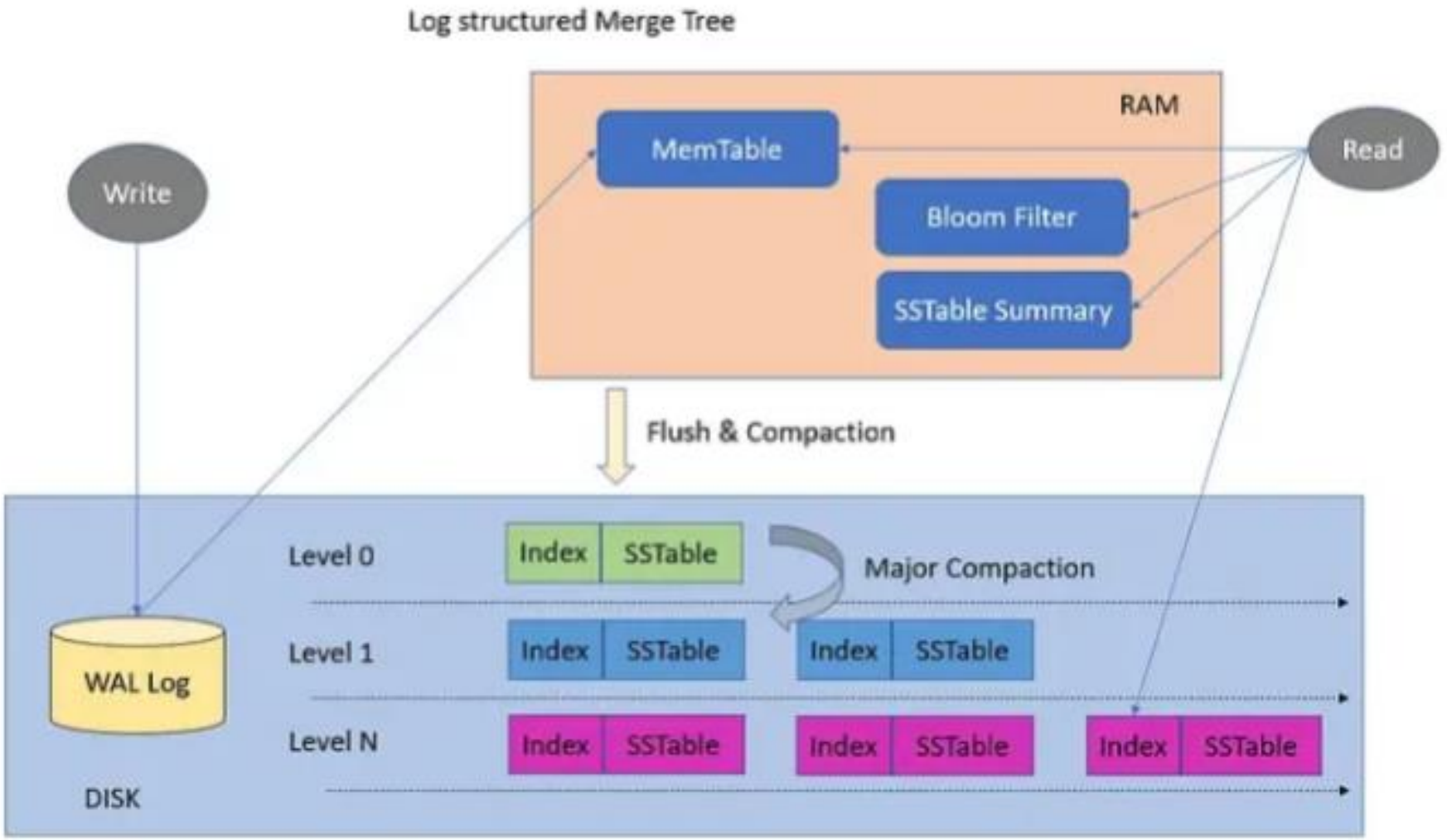


Constructing and maintaining SSTables

Merging and compacting sstables

Bloom Filter

LSM Trees


Log structured Merge Tree
RAM
MemTable
Bloom Filter
SSTable Summary
Write
Read
Flush & Compaction
Level 0
Index
SSTable
Major Compaction
Level 1
Index
SSTable
Index
SSTable
WAL Log
Level N
Index
SSTable
Index
SSTable
Index
SSTable
DISK

DATA MODEL

Data Model

- Keyspace
  - *Similar to database*
  - *Has replication factor and replication strategy*
- Table
  - *Tables are composed of rows and columns*

Data Model

- Partition
- Row

Data Model

- Column
  - *Native types (Integers, Boolean, text, timestamp, …)*
  - *Collection (list, set, map)*
  - *Counters (*
  - *User defined types*

Data Model

- Primary Key
  - *Partition Key*
  - *Clustering Key*

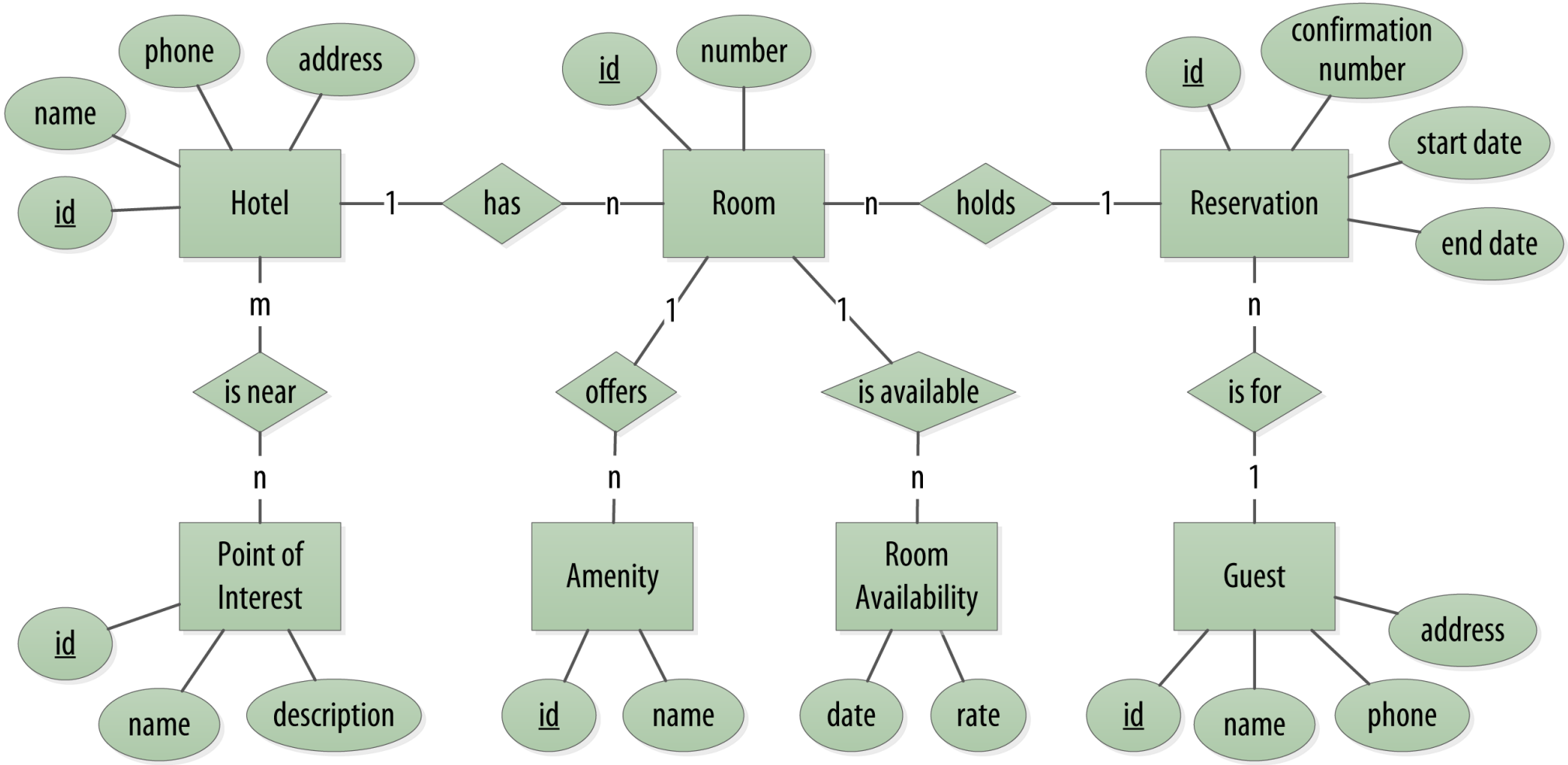
CQL

- Like SQL but avoids:
  - *Cross-partition transactions*
  - *Distributed joins*
  - *Foreign keys or referential integrity.*

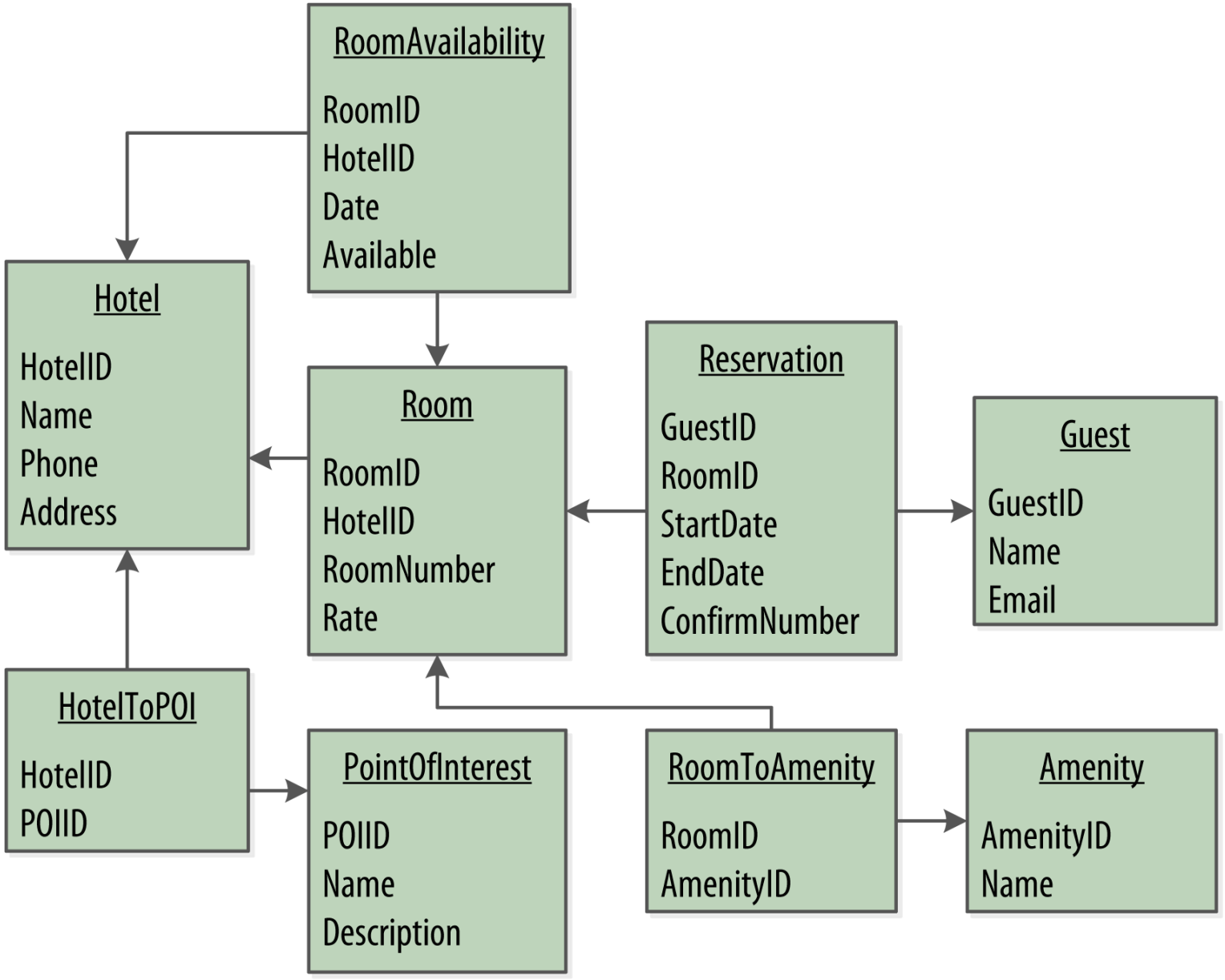
Data Modeling

- Data modeling is the process of identifying entities and their relationships
- In relational databases:
  - *data is placed in normalized tables with foreign keys used to reference related data in other tables*
- In Cassandra:
  - *data modeling is query-driven*

Conceptual data modeling



RDMS Design


RoomAvailability
RoomID
HotelID
Date
Available
Hotel
HotelID
Name
Phone
Address
Room
RoomID
HotelID
RoomNumber
Rate
Reservation
GuestID
RoomID
StartDate
EndDate
ConfirmNumber
Guest
GuestID
Name
Email
HotelToPOI
HotelID
POIID
PointOfInterest
POIID
Name
Description
RoomToAmenity
RoomID
AmenityID
Amenity
AmenityID
Name

Design Differences Between RDBMS and Cassandra

- No joins
- No referential integrity
- Denormalization
- Query-first Design
- Designing for optimal storage
- Sorting is a design decision

Data Modeling in Cassandra

- Learn more:
  - *https://cassandra.apache.org/doc/latest/cassandra/developing/data-modeling/data-modeling_logical.html*
  - *https://cassandra.apache.org/doc/latest/cassandra/developing/data-modeling/data-modeling_physical.html*
  - *https://cassandra.apache.org/doc/latest/cassandra/developing/data-modeling/data-modeling_refining.html*
  - *https://cassandra.apache.org/doc/latest/cassandra/developing/data-modeling/data-modeling_schema.html*

GUARANTEES

- Cassandra makes the following guarantees.
  - *High Scalability*
  - *High Availability*
  - *Durability*
  - *Eventual Consistency of writes to a single table*
  - *Lightweight transactions with linearizable consistency*
  - *Batched writes across multiple tables are guaranteed to succeed completely or not at all*
  - *Secondary indexes are guaranteed to be consistent with their local replicas' data*

Batched Writes

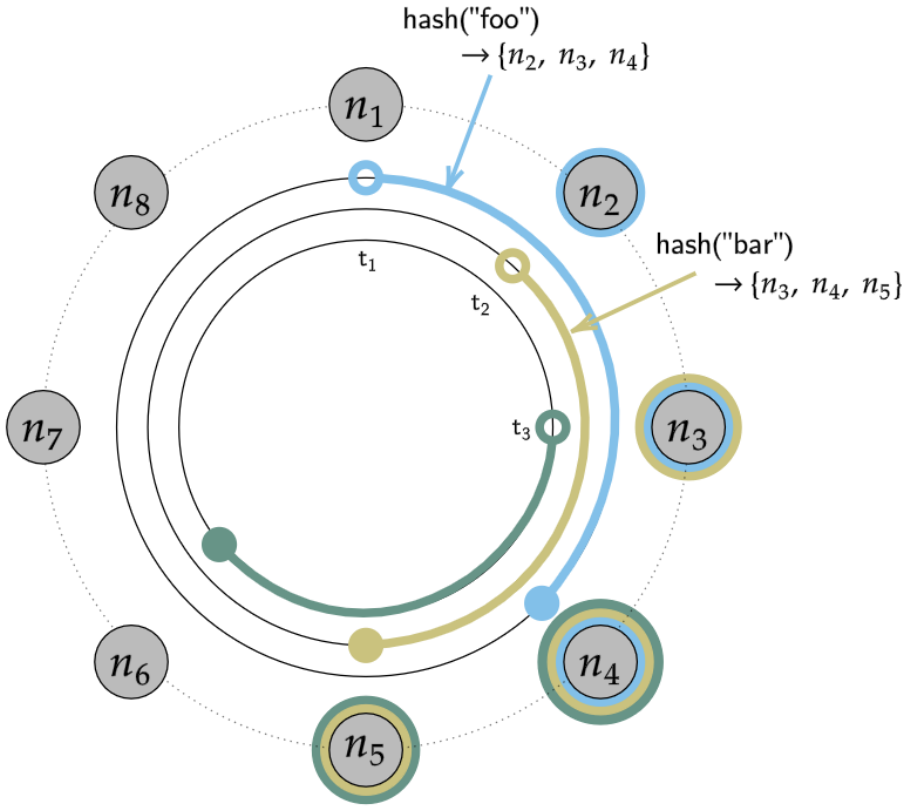
- The guarantee for batched writes across multiple tables is that they will eventually succeed, or none will. Batch data is first written to batchlog system data, and when the batch data has been successfully stored in the cluster, the batchlog data is removed. The batch is replicated to another node to ensure that the full batch completes in the event if coordinator node fails.

DISTRIBUTED ARCHITECTURE

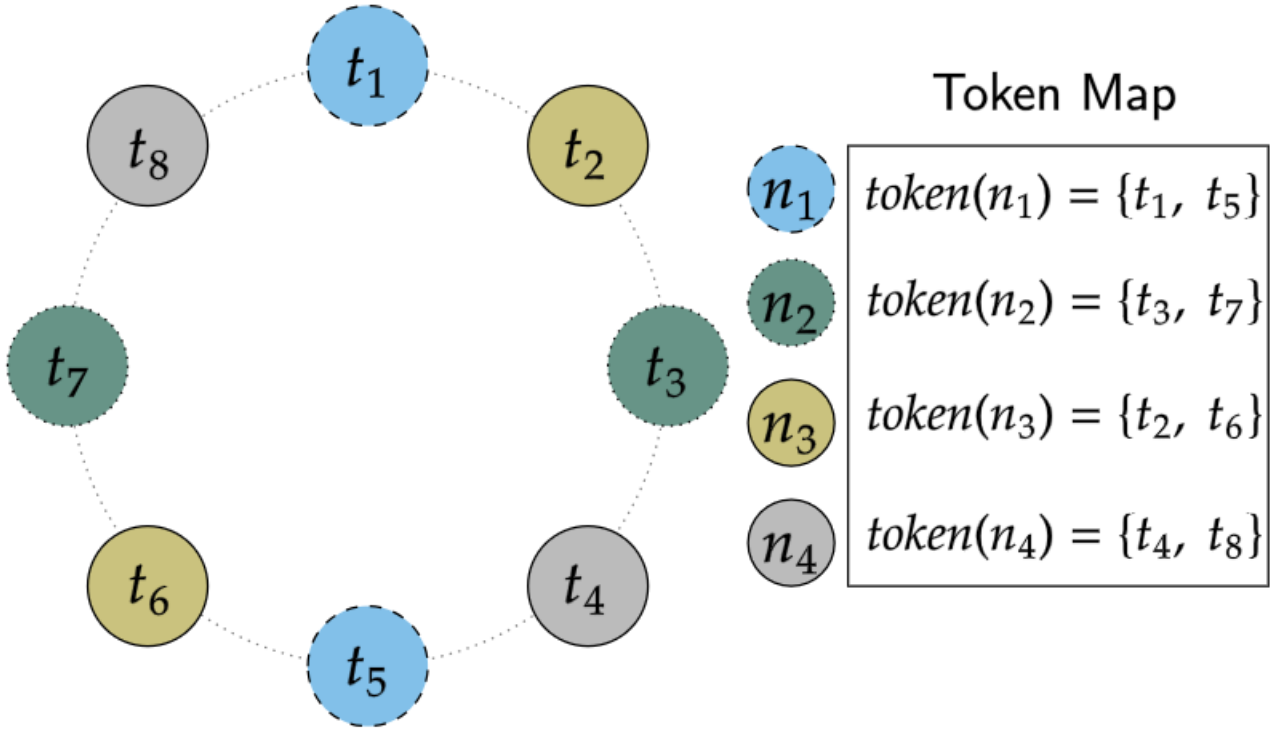
Dataset Partitioning: Consistent Hashing

token(n_1) = t_1
token(n_2) = t_2
token(n_3) = t_3
...

range(t_1, t_2] → {n_2, n_3, n_4}
range(t_2, t_3] → {n_3, n_4, n_5}
range(t_3, t_4] → {n_4, n_5, n_6}
...



Dataset Partitioning: Consistent Hashing



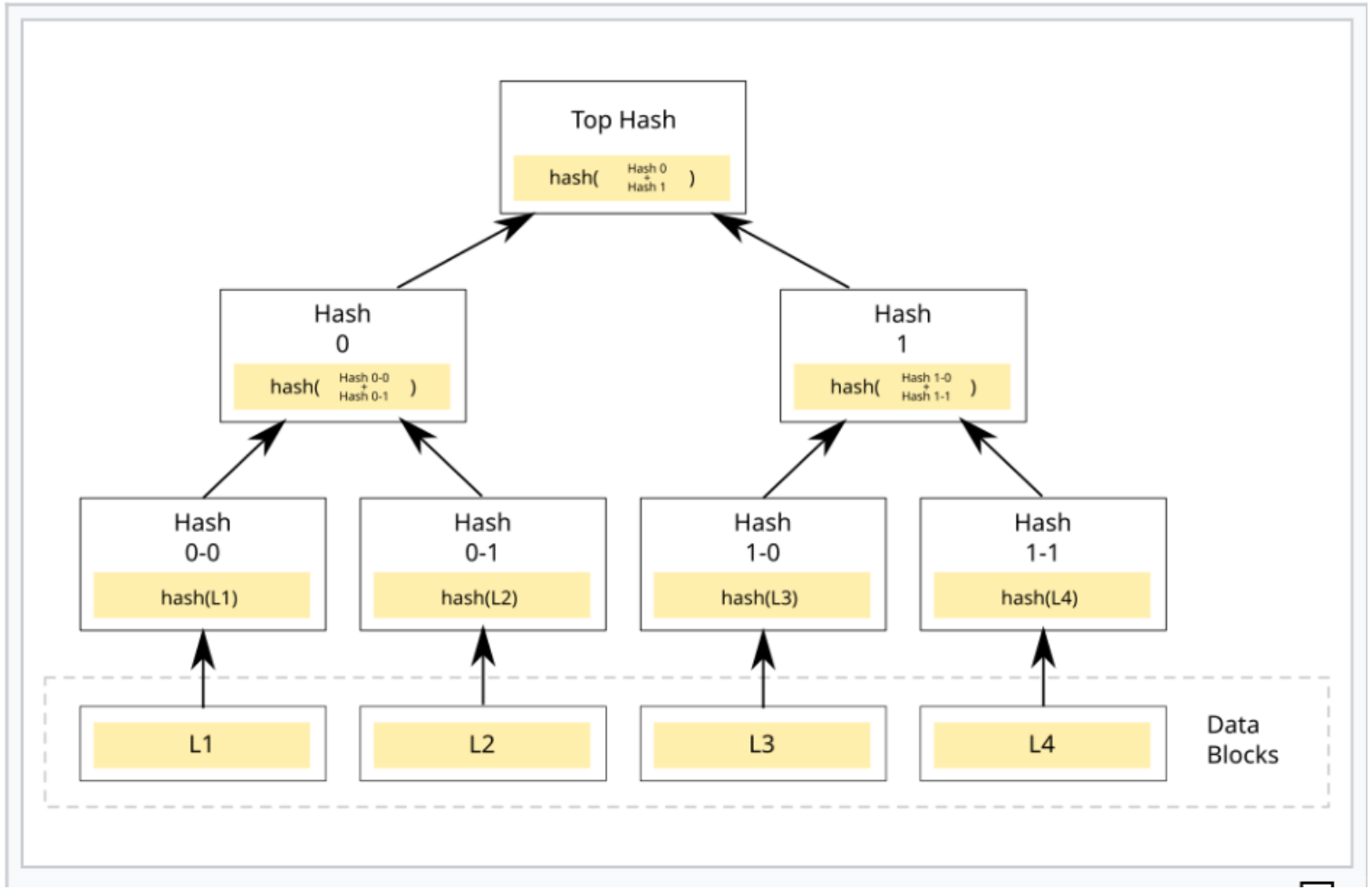
Conflict resolution: LWW

Replication Strategy

Replica Synchronization

- Replica read repair
- Hinted Handoff
- Merkle trees

Merkle Trees



Tunable Consistency

- ONE
  - *Only a single replica must respond.*
- TWO
  - *Two replicas must respond.*
- THREE
  - *Three replicas must respond.*
- QUORUM
  - *A majority (n/2 + 1) of the replicas must respond.*

Tunable Consistency

- ALL
  - *All of the replicas must respond.*
- LOCAL_QUORUM
  - *A majority of the replicas in the local datacenter (whichever datacenter the coordinator is in) must respond.*
- EACH_QUORUM
  - *A majority of the replicas in each datacenter must respond.*
- LOCAL_ONE
  - *Only a single replica must respond. In a multi-datacenter cluster, this also guarantees that read requests are not sent to replicas in a remote datacenter.*
- ANY
  - *A single replica may respond, or the coordinator may store a hint. If a hint is stored, the coordinator will later attempt to replay the hint and deliver the mutation to the replicas. This consistency level is only accepted for write operations.*

Distributed Cluster Membership and Failure Detection

- Gossip
  - *Send information using a vector clock*
  - *Every node in the Cassandra cluster runs the gossip task independently and periodically*

EXAMPLE USAGE

STORAGE ATTACHED INDEX

SHOULD I USE CASSANDRA?