Carnegie Mellon Univ.
Dept. of Computer Science
15-415/615 - DB Applications

C. Faloutsos – A. Pavlo

Lecture#28: Modern Systems

System Votes

MongoDB	32
Google Spanner/F1	22
LinkedIn Espresso	16
Apache Cassandra	16
Facebook Scuba	16
Apache Hbase	14
VoltDB	10
Redis	10
Vertica	5

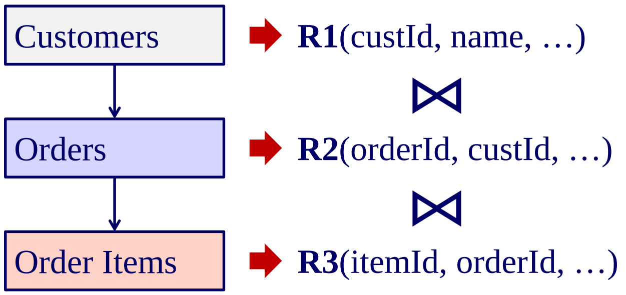
Cloudera Impala	5
DeepDB	2
SAP HANA	1
CockroachDB	1
SciDB	1
InfluxDB	1
Accumulo	1
Apache Trafodion	1

MongoDB

- Document Data Model
  - Think JSON, XML, Python dicts
  - Not Microsoft Word documents
- Different terminology:
  - Document → Tuple
  - Collection → Table/Relation

MongoDB

- A customer has orders and each order has order items.



R1(custId, name, …)

$\bowtie$

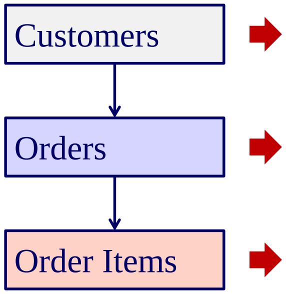
R2(orderId, custId, …)

$\bowtie$

R3(itemId, orderId, …)

MongoDB

- A customer has orders and each order has order items.



```
{ "custId" : 1234,
  "custName" : “Trump",
  "orders" : [
    { "orderId" : 10001,
      "orderItems" : [
        { "itemId" : "XXXX",
          "price"  : 19.99 },
        { "itemId" : "YYYY",
          "price"  : 29.99 } ]
    },
    { "orderId" : 10050,
      "orderItems" : [
        { "itemId" : “ZZZZ",
          "price"  : 49.99 } ]
    } ]
}
```

MongoDB

- JSON-only query API
- Single-document atomicity.
  - **OLD**: No server-side joins. Had to “pre-join” collections by embedding related documents inside of each other.
  - **NEW**: Server-side joins (only left-outer equi)
- No cost-based query planner / optimizer.

MongoDB

- Heterogeneous distributed components.
  - Centralized query router.
- Master-slave replication.
- Auto-sharding:
  - Define ‘partitioning’ attributes for each collection (hash or range).
  - When a shard gets too big, the DBMS automatically splits the shard and rebalances.

MongoDB

- Originally used **`mmap`** storage manager
  - No buffer pool.
  - Let the OS decide when to flush pages.
  - Single lock per database.

MongoDB

- Version 3 (2015) now supports pluggable storage managers.
  - **WiredTiger** from BerkeleyDB alumni. http://cmudb.io/lectures2015-wiredtiger
  - **RocksDB** from Facebook ("MongoRocks") http://cmudb.io/lectures2015-rocksdb

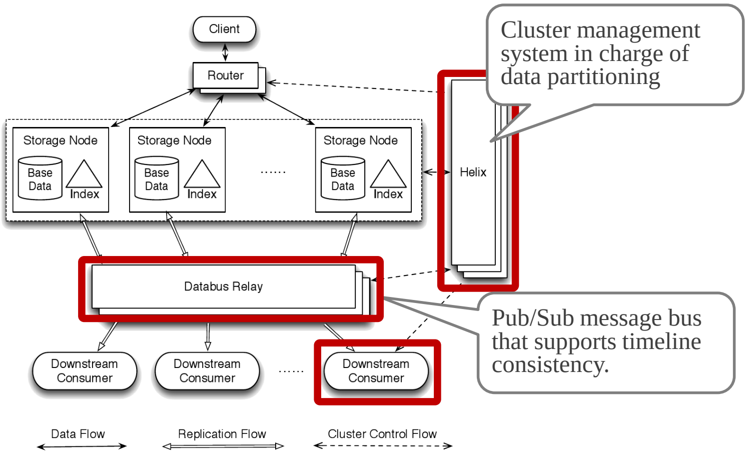
LinkedIn Espresso

- System goals:
  - Support distributed transactions across documents
  - Strong consistency to act as a single source-of-truth for user data
  - Integrates with the entire data ecosystem
- Replace legacy Oracle installations
  - Started with InMail messaging service

LinkedIn Espresso

- Distributed document DBMS deployed in production since 2012.

LinkedIn Espresso



Cluster management system in charge of data partitioning

Pub/Sub message bus that supports timeline consistency.

Source: On Brewing Fresh Espresso: LinkedIn's Distributed Data Serving Platform, SIGMOD 2013

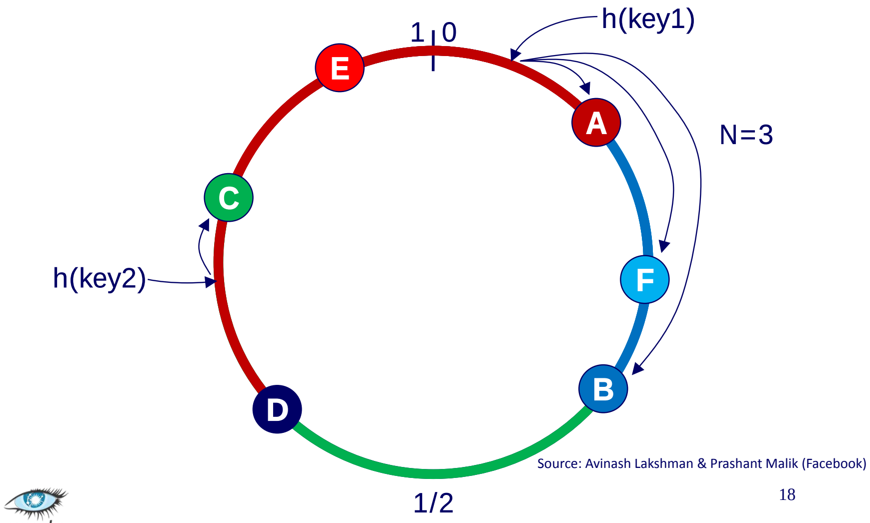
History

- Amazon publishes a paper in 2007 on the Dynamo system.
  - Eventually consistency key/value store
  - Partitions based on ***consistent hashing***
- People at Facebook start writing Cassandra as a clone of Dynamo in 2008 for their message service.
  - Ended up not using the system and releasing the source code.

Apache Cassandra

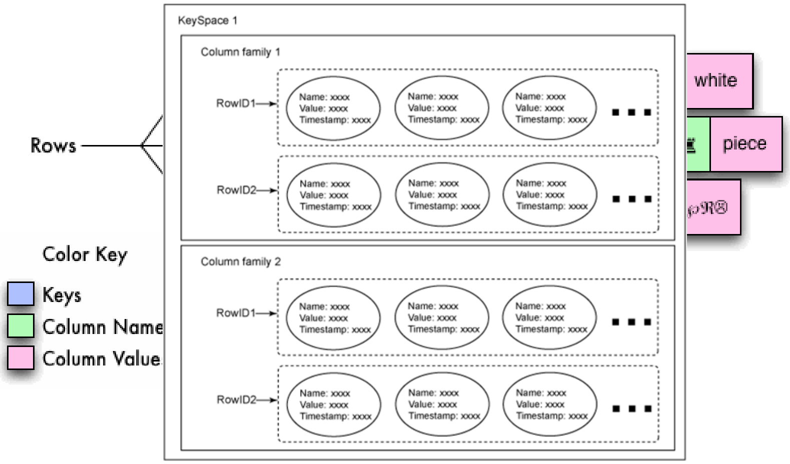
- Borrows a lot of ideas from other systems:
  - Consistent Hashing (Amazon Dynamo)
  - Column-Family Data Model (Google BigTable)
  - Log-structured Merge Trees
- Originally one of the leaders of the NoSQL movement but now pushing “CQL”

Consistent Hashing



Source: Avinash Lakshman & Prashant Malik (Facebook)

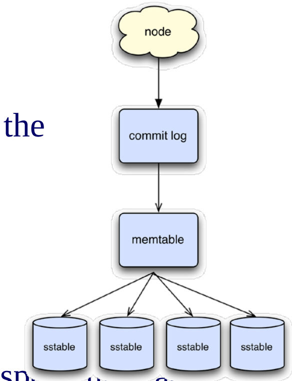
Column-Family Data Model



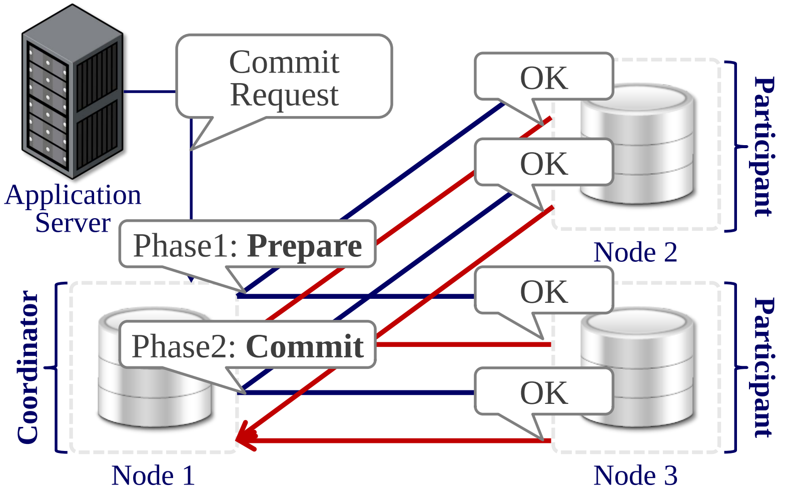
Source: Gary Dusbabek (Rackspace)

LSM Storage Model

- The log is the database.
  - Have to read log to reconstruct the record for a read.
- MemTable: In-memory cache
- SSTables:
  - Read-only portions of the log.
  - Use indexes + Bloom filters to speed up reads
- See the **RocksDB** talk from this semester: http://cmudb.io/lectures2015-rocksdb



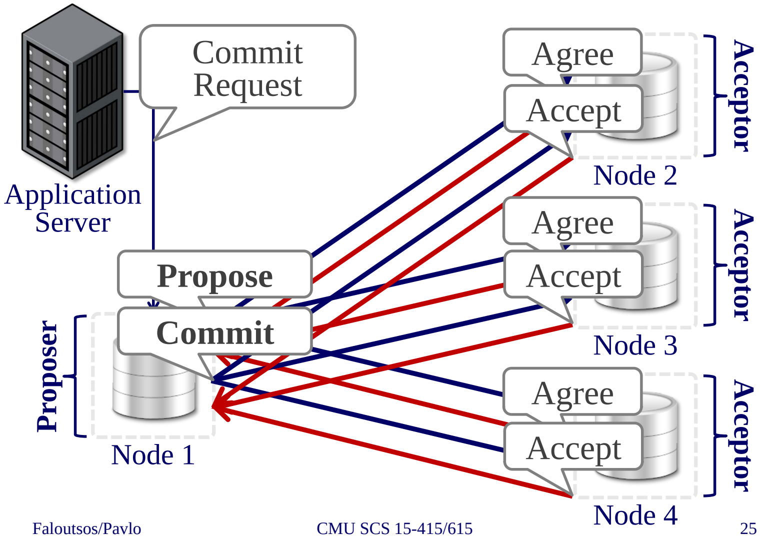
Two-Phase Commit



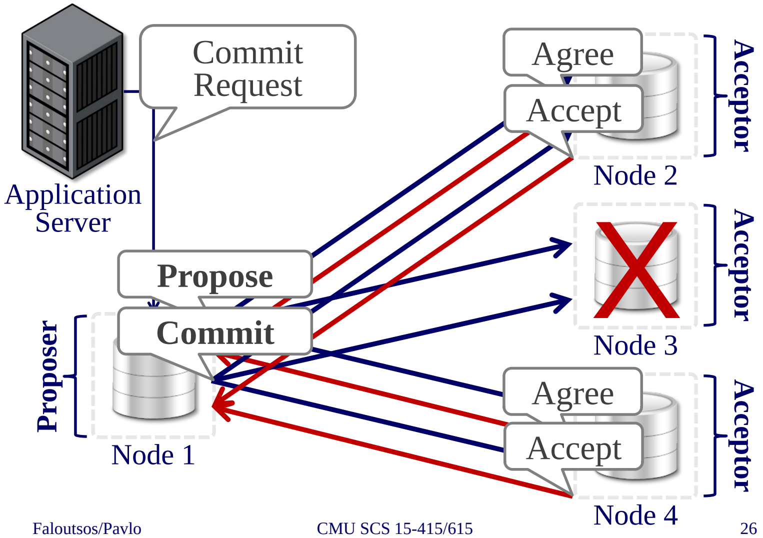
Paxos

- Consensus protocol where a coordinator proposes an outcome (e.g., commit or abort) and then the participants vote on whether that outcome should succeed.
- Does not block if a majority of participants are available and has provably minimal message delays in the best case.
  - First correct protocol that was provably resilient in the face asynchronous networks

Paxos



Paxos



Commit Request

Application Server

Propose

Commit

Proposer

Node 1

Agree

Accept

Node 2

Acceptor

Node 3

Acceptor

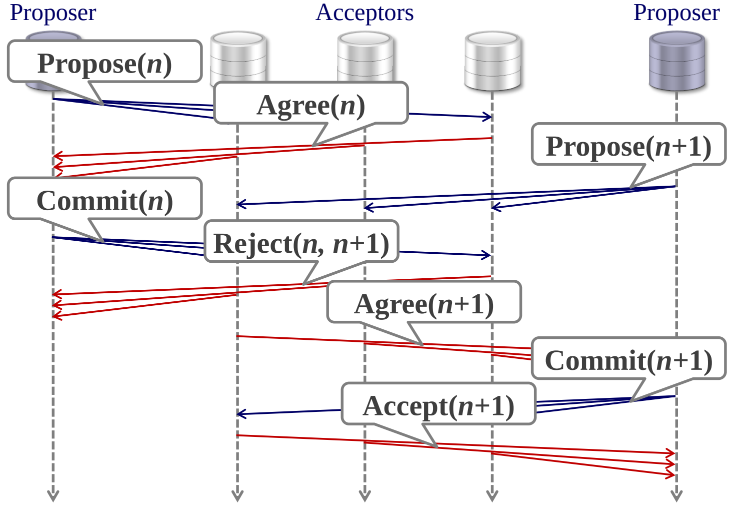
Agree

Accept

Node 4

Acceptor

Paxos



Proposer

Acceptors

Proposer

Propose(*n*)

Agree(*n*)

Propose(*n*+1)

Commit(*n*)

Reject(*n*, *n*+1)

Agree(*n*+1)

Commit(*n*+1)

Accept(*n*+1)

2PC vs. Paxos

- 2PC is a degenerate case of Paxos.
  - Single coordinator.
  - Only works if everybody is up.
- Use leases to determine who is allowed to propose new updates to avoid continuous rejection.

Google Spanner

- Google’s geo-replicated DBMS (>2011)
- Schematized, semi-relational data model.
- Concurrency Control:
  - 2PL + T/O (Pessimistic)
  - Externally consistent global write-transactions with synchronous replication.
  - Lock-free read-only transactions.

Google Spanner

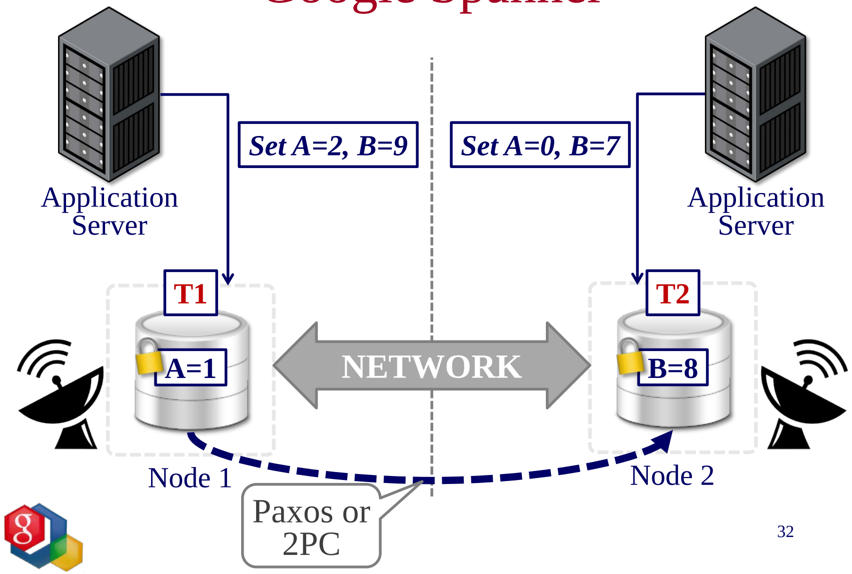
```
CREATE TABLE users {
  uid INT NOT NULL,
  email VARCHAR,
  PRIMARY KEY (uid)
};
CREATE TABLE albums {
  uid INT NOT NULL,
  aid INT NOT NULL,
  name VARCHAR,
  PRIMARY KEY (uid, aid)
} INTERLEAVE IN PARENT users
  ON DELETE CASCADE;
```

users(1001)
↳albums(1001, 9990)
↳albums(1001, 9991)
users(1002)
↳albums(1002, 6631)
↳albums(1002, 6634)

Google Spanner

- Ensures ordering through globally unique timestamps generated from atomic clocks and GPS devices.
- Database is broken up into tablets:
  - Use Paxos to elect leader in tablet group.
  - Use 2PC for txns that span tablets.
- TrueTime API

Google Spanner



Google F1

- OCC engine built on top of Spanner.
  - Read phase followed by a write phase
  - In the read phase, F1 returns the last modified timestamp with each row. No locks.
  - The timestamp for a row is stored in a hidden lock column. The client library returns these timestamps to the F1 server
  - If the timestamps differ from the current timestamps at the time of commit the transaction is aborted