Dynamo: Amazon’s Highly Available Key-Value Store

Availability is important

Tens of millions of customers at peak times

Tens of millions of shopping cart requests, 3 million checkouts per day

Hundreds of thousands of concurrently active sessions

Strict Service-Level Agreements (SLAs) translate to business value

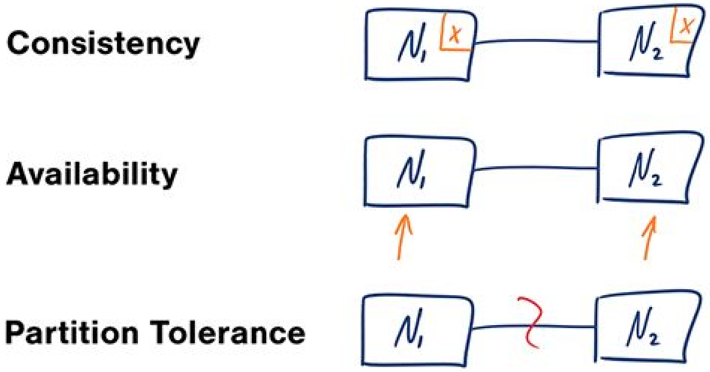
Key challenges

Failure is common

Even if each machine is available 99.999% of the time, a datacenter with 100,000 machines still encounters failures `(1-(1-p)^n)` = 63% of the time

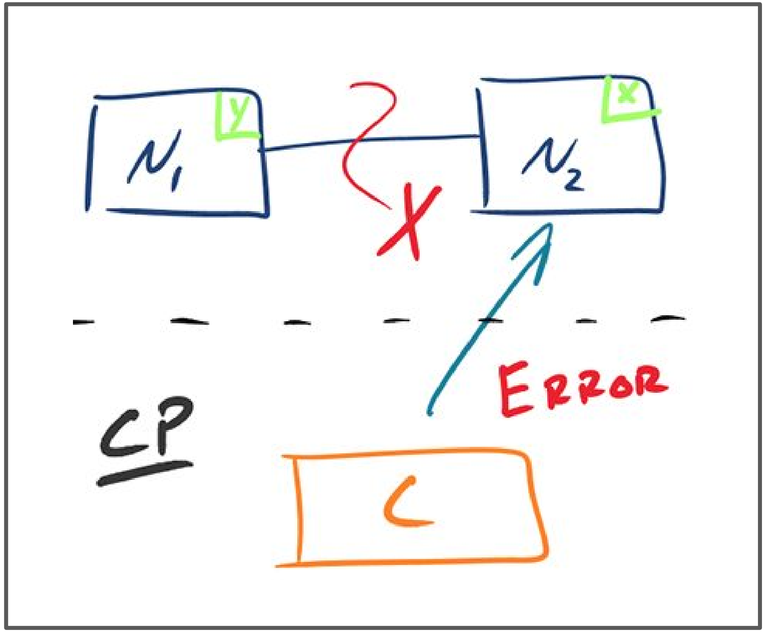
Difficult to provide availability and consistency *(linearizability)*

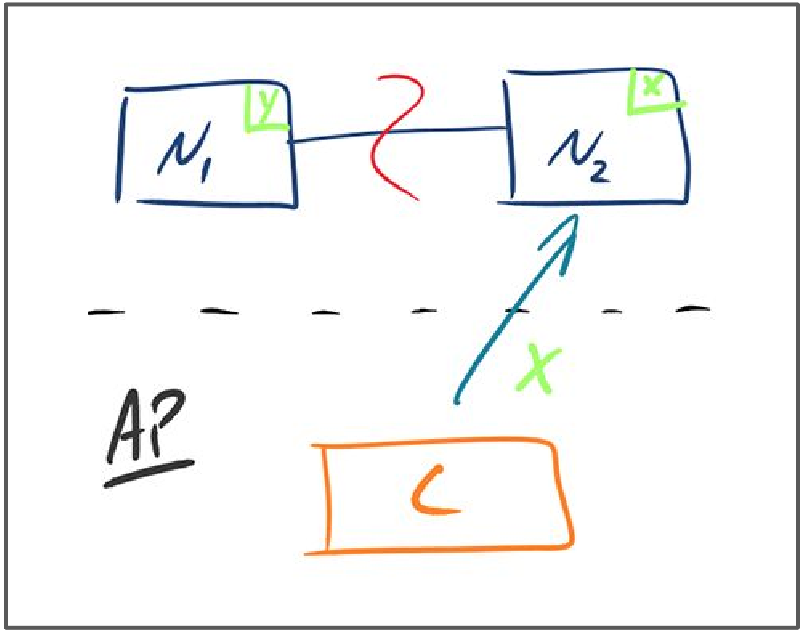
CAP Theorem



Impossible for a system to provide all three simultaneously

CAP Theorem





Dynamo

Fully decentralized, highly available key-value store

Always writeable, resolve conflicts during reads

API for clients to specify SLA requirements (99.9%)

Departure from RDBMS: simpler functionality, fewer guarantees, runs on commodity hardware

Techniques for achieving availability

Consistent hashing for partitioning key space

Vector clocks for reconciling conflicts during reads

Sloppy quorums for handling temporary failures

Anti-entropy using Merkle trees for handling permanent failures

Gossip-based protocol for membership notifications

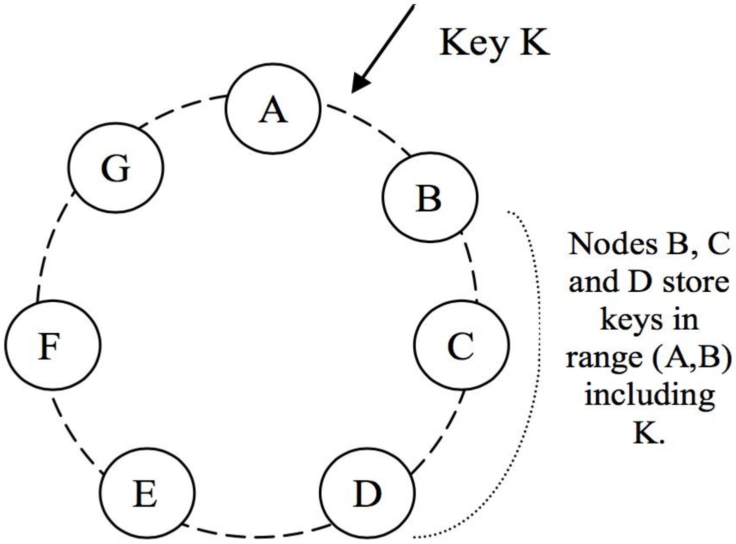
Consistent Hashing

Assign each node a random position on the ring

Node owns the preceding key range

For fault tolerance, replicate each key at N successor nodes in the ring

Virtual nodes: each physical node gets assigned multiple nodes on the ring



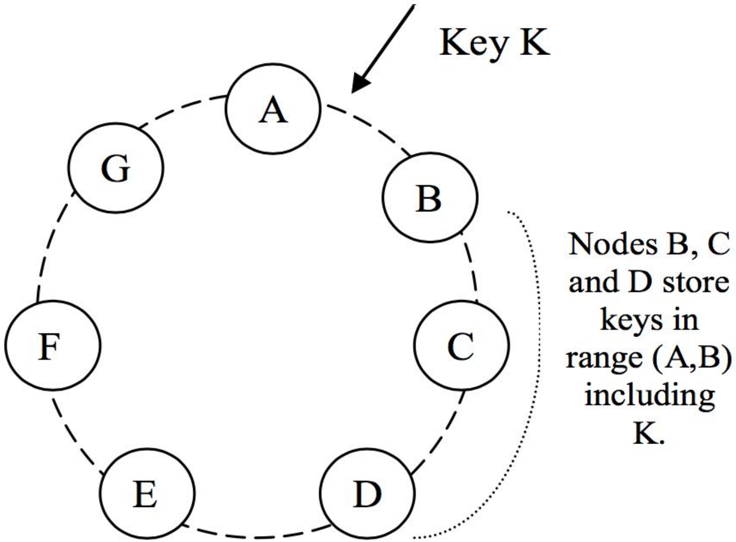
Consistent Hashing

Desirable properties?

Uniform distribution of load

Minimum object movements when nodes join or leave the ring

Number of virtual nodes can be adjusted for device heterogeneity



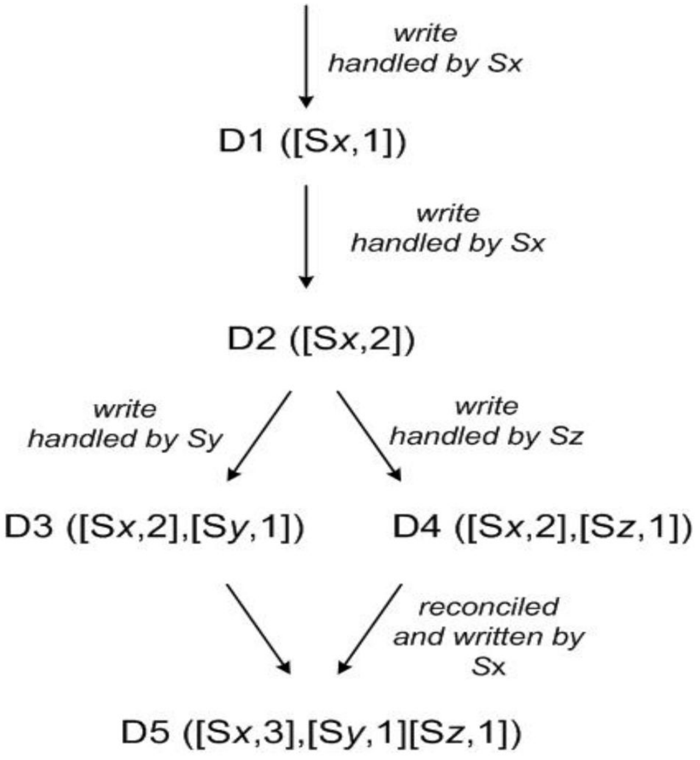
Conflict resolution

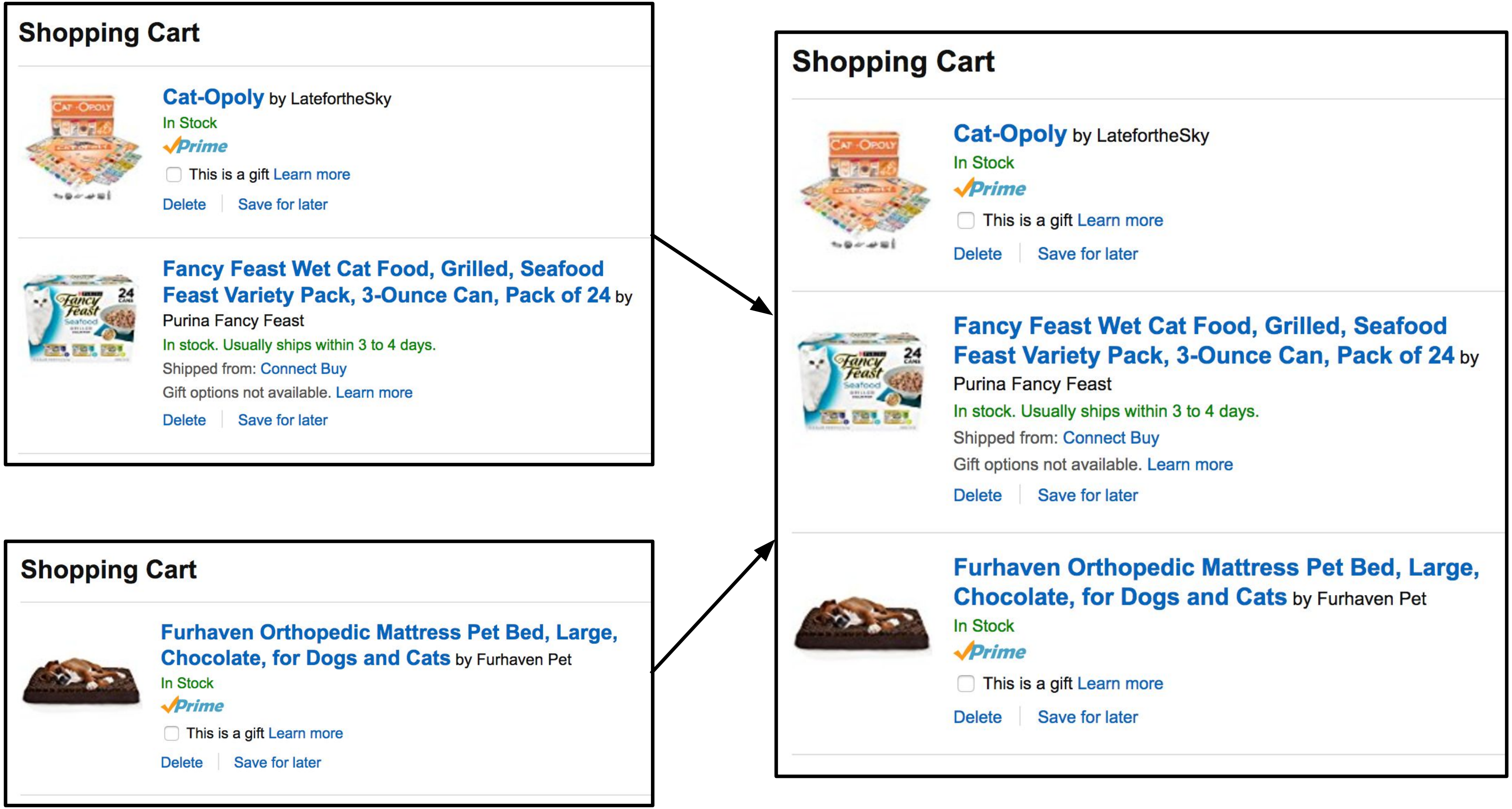
Two machines write different values to the same key

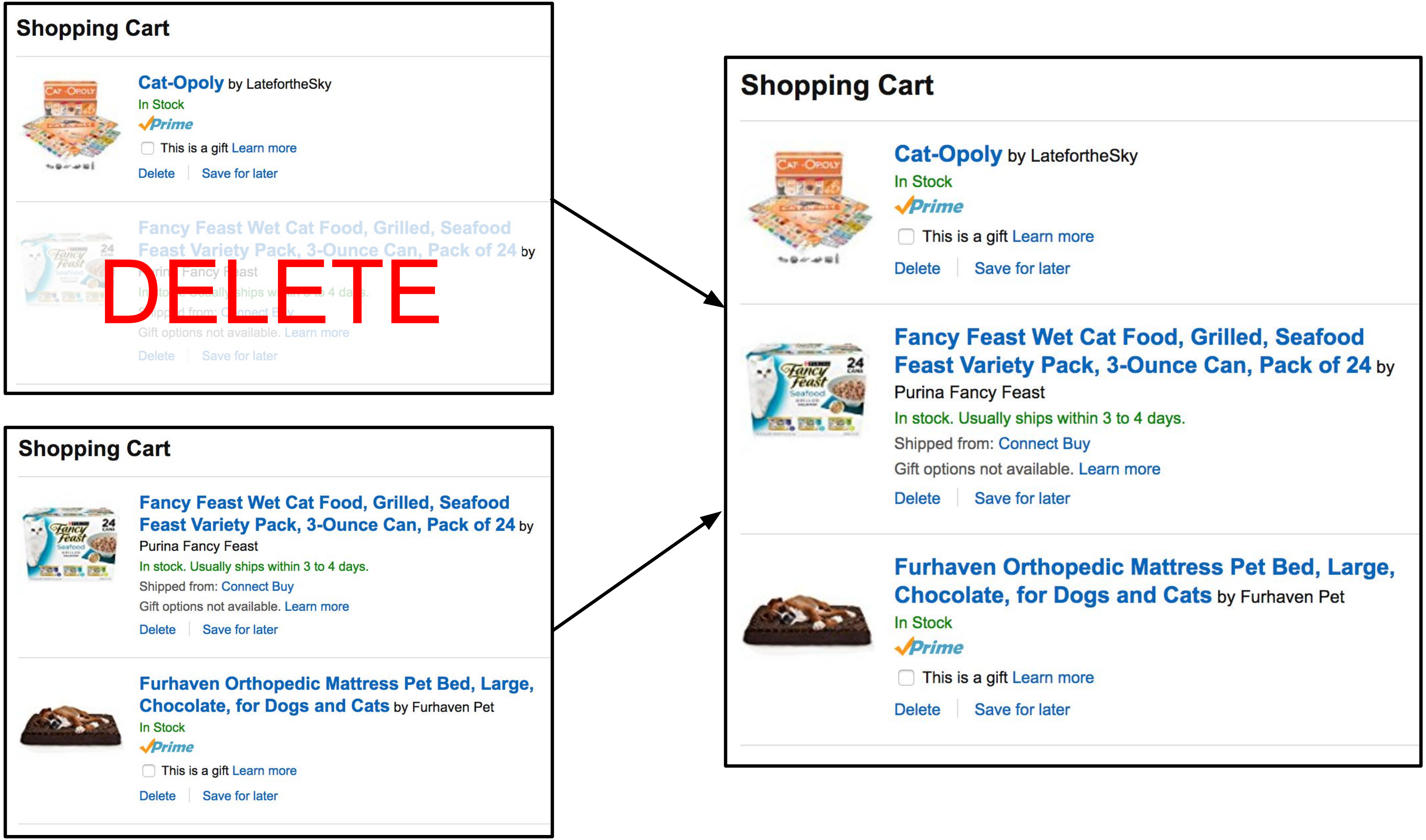
Vector clocks: list of (node, count) pairs where count is incremented on write

If one vector clock subsumes another, discard older value

Else, return all conflicting values to client




Shopping Cart
Cat-Opoly by LatefortheSky
In Stock
Prime
This is a gift Learn more
Delete
Save for later
Fancy Feast Wet Cat Food, Grilled, Seafood Feast Variety Pack, 3-Ounce Can, Pack of 24 by Purina Fancy Feast
In stock. Usually ships within 3 to 4 days.
Shipped from: Connect Buy
Gift options not available. Learn more
Delete
Save for later
Shopping Cart
Furhaven Orthopedic Mattress Pet Bed, Large, Chocolate, for Dogs and Cats by Furhaven Pet
In Stock
Prime
This is a gift Learn more
Delete
Save for later
Shopping Cart
Cat-Opoly by LatefortheSky
In Stock
Prime
This is a gift Learn more
Delete
Save for later
Fancy Feast Wet Cat Food, Grilled, Seafood Feast Variety Pack, 3-Ounce Can, Pack of 24 by Purina Fancy Feast
In stock. Usually ships within 3 to 4 days.
Shipped from: Connect Buy
Gift options not available. Learn more
Delete
Save for later
Furhaven Orthopedic Mattress Pet Bed, Large, Chocolate, for Dogs and Cats by Furhaven Pet
In Stock
Prime
This is a gift Learn more
Delete
Save for later


Shopping Cart
Cat-Opoly by LatefortheSky
In Stock
Prime
This is a gift Learn more
Delete | Save for later
Fancy Feast Wet Cat Food, Grilled, Seafood Feast Variety Pack, 3-Ounce Can, Pack of 24 by
DELETE
Gift options not available. Learn more
Delete | Save for later
Shopping Cart
Fancy Feast Wet Cat Food, Grilled, Seafood Feast Variety Pack, 3-Ounce Can, Pack of 24 by Purina Fancy Feast
In stock. Usually ships within 3 to 4 days.
Shipped from: Connect Buy
Gift options not available. Learn more
Delete | Save for later
Furhaven Orthopedic Mattress Pet Bed, Large, Chocolate, for Dogs and Cats by Furhaven Pet
In Stock
Prime
This is a gift Learn more
Delete | Save for later
Shopping Cart
Cat-Opoly by LatefortheSky
In Stock
Prime
This is a gift Learn more
Delete | Save for later
Fancy Feast Wet Cat Food, Grilled, Seafood Feast Variety Pack, 3-Ounce Can, Pack of 24 by Purina Fancy Feast
In stock. Usually ships within 3 to 4 days.
Shipped from: Connect Buy
Gift options not available. Learn more
Delete | Save for later
Furhaven Orthopedic Mattress Pet Bed, Large, Chocolate, for Dogs and Cats by Furhaven Pet
In Stock
Prime
This is a gift Learn more
Delete | Save for later

Sloppy Quorums

Write to N nodes, return success when $W < N$ nodes respond

Read from N nodes, return value(s) from $R < N$ nodes

Typically, $W+R > N$ means at least one writer and one reader overlap, so values are consistent

Sloppy here means skip nodes that have failed, such that even if $W+R > N$, the readers and writers may not overlap = not consistent!

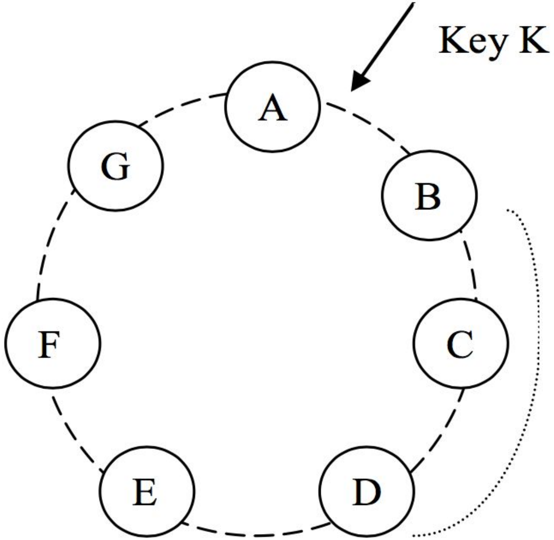
Sloppy Quorums

Example:

Typical values are N = 3, W = R = 2

Nodes C and D have failed, so key *k* is written to E and F instead

Nodes C and D recover, and now client tries to read from C and D = stale value

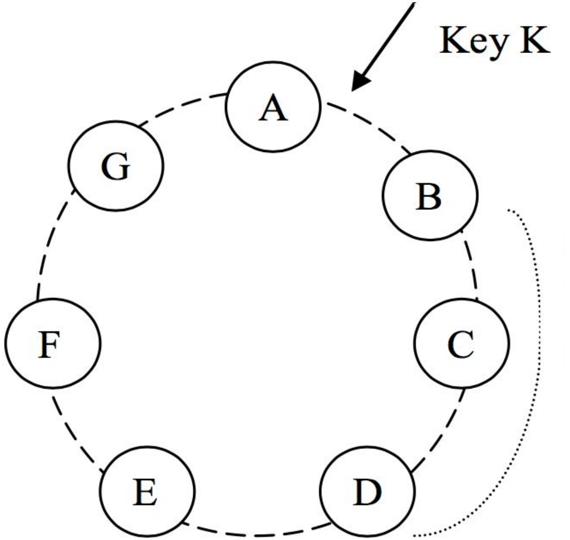


Hinted Handoff

Example:

Nodes E and F remember they are writing on behalf of C and D

As soon as C and D recovers, E and F transfer their values for *k* to C and D

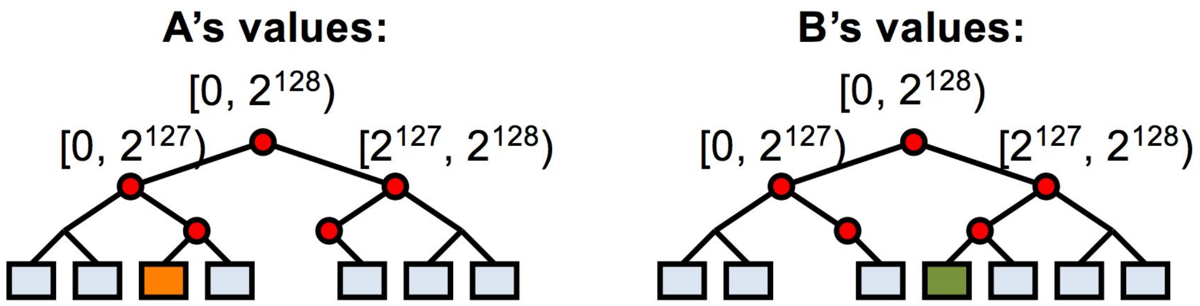


Anti-entropy using Merkle trees

Goal: minimize durability loss from above techniques

Nodes responsible for the same key spaces exchange Merkle trees

Find differences quickly while exchanging little information



Membership notification

Gossip-based protocol to propagate membership changes

Each node learns the key spaces handled by all other nodes

Result: zero-hop distributed hash table (DHT)

Clearly not infinitely scalable → finger tables?

References

http://robertgreiner.com/2014/08/cap-theorem-revisited/

http://s3.amazonaws.com/AllThingsDistributed/sosp/amazon-dynamo-sosp2007.pdf

https://pdos.csail.mit.edu/papers/chord:sigcomm01/chord_sigcomm.pdf