



Riak Intro

Us

- Shanley Kane
@shanley
shanley@basho.com
- Mark Phillips
@pharkmillups
mark@basho.com



riak

What's in store?

- At a High Level
- For Developers
- Under the Hood
- When and Why
- Some Users
- Commercial Extensions
- 1.2 and Roadmap

At a High Level

Riak

- Dynamo-inspired key/value store
 - with some extras: search, MapReduce, 2i, links, pre- and post-commit hooks, pluggable backends, HTTP and binary interfaces
- Written in Erlang with C/C++
- Open source under Apache 2 License

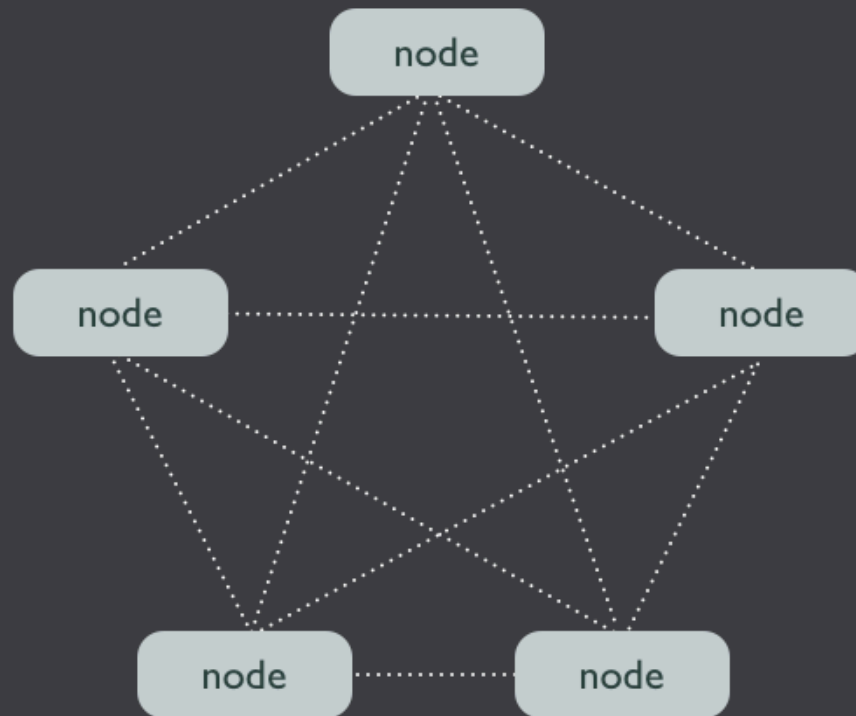
Riak's Design Goals (1)

- High-availability
- Low-latency
- Horizontal Scalability
- Fault Tolerance
- Ops Friendliness
- Predictability

Riak's Design Goals (2)

- Design Informed by Brewer's CAP Theorem and Amazon's Dynamo Paper
- Riak is tuned to offer availability above all else
- Developers can tune for consistency (more on this later)

Masterless; deployed as a cluster of nodes



For Developers

Riak is a database that stores keys against values. Keys are grouped into a higher-level namespace called buckets.

Riak doesn't care what you store.
It will accept any data type; things
are stored on disk as binaries.

key

key	value
-----	-------

bucket



bucket

key	value
key	value
key	value
key	value

Two APIs

1. HTTP (just like the web)
2. Protocol Buffers (thank you, Google)

Querying

GET/PUT/DELETE

MapReduce

Full-Text Search

Secondary Indexes (2i)

Tunable Consistency

- `n_val` - number of replica to store; bucket-level setting. Defaults to “3”.
- `w` - number of replicas required for a successful write; Defaults to “2”.
- `r` - number of replica acks required for a successful read. request-level setting. Defaults to “2”.
- Tweak consistency vs. availability

Client Libraries

Ruby, Node.js, Java, Python, Perl, OCaml, Erlang, PHP, C, Squeak, Smalltalk, Pharoah, Clojure, Scala, Haskell, Lisp, Go, .NET, Play, and more (supported by either Basho or the community).

Under the Hood

Consistent Hashing and Replicas

Virtual Nodes

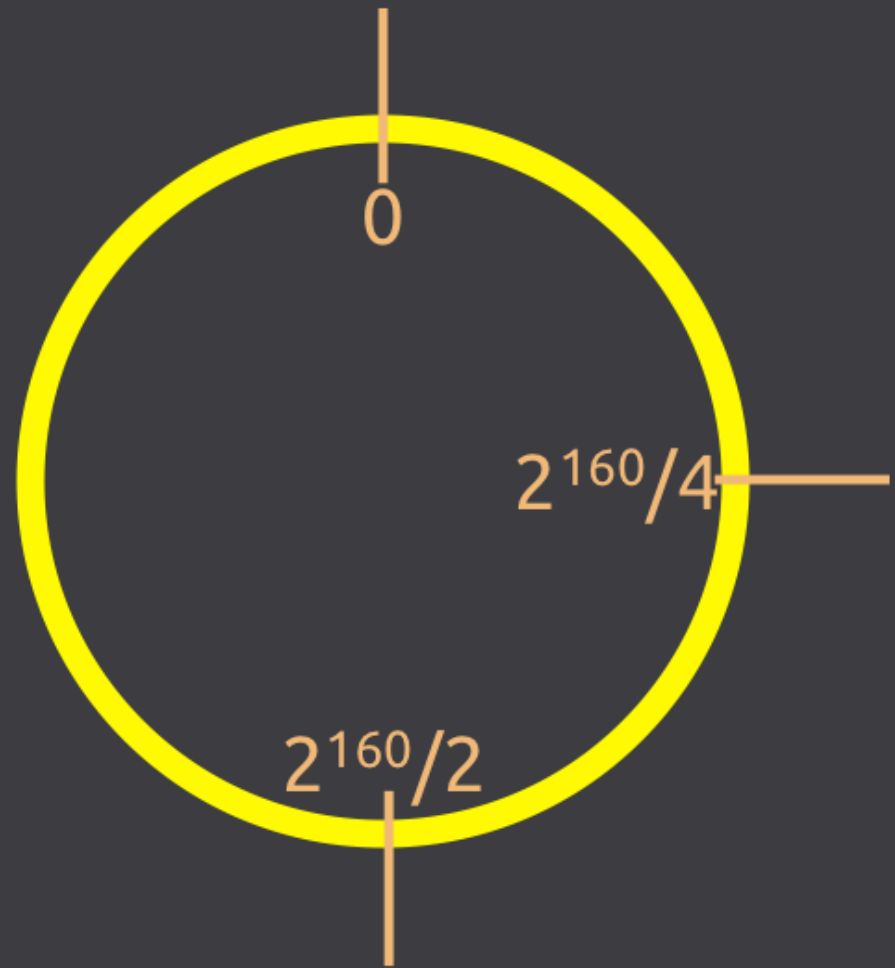
Vector Clocks

Handoff and Rebalancing

Gossiping

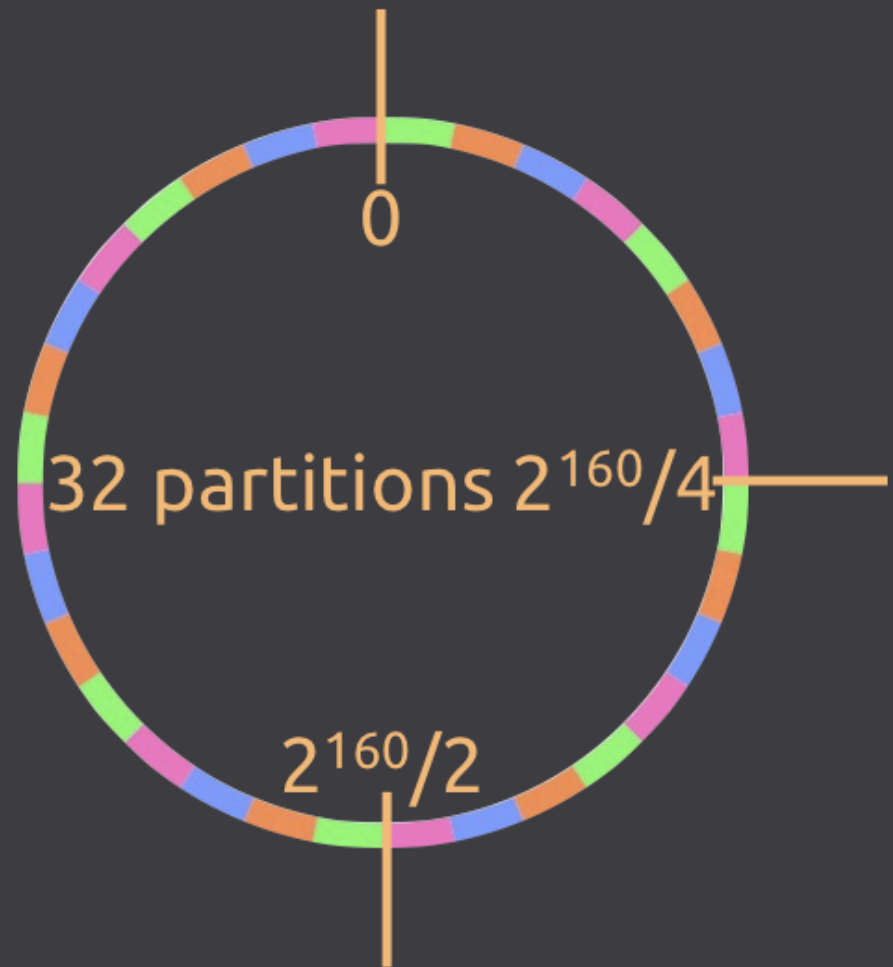
Consistent Hashing

- 160-bit integer keyspace



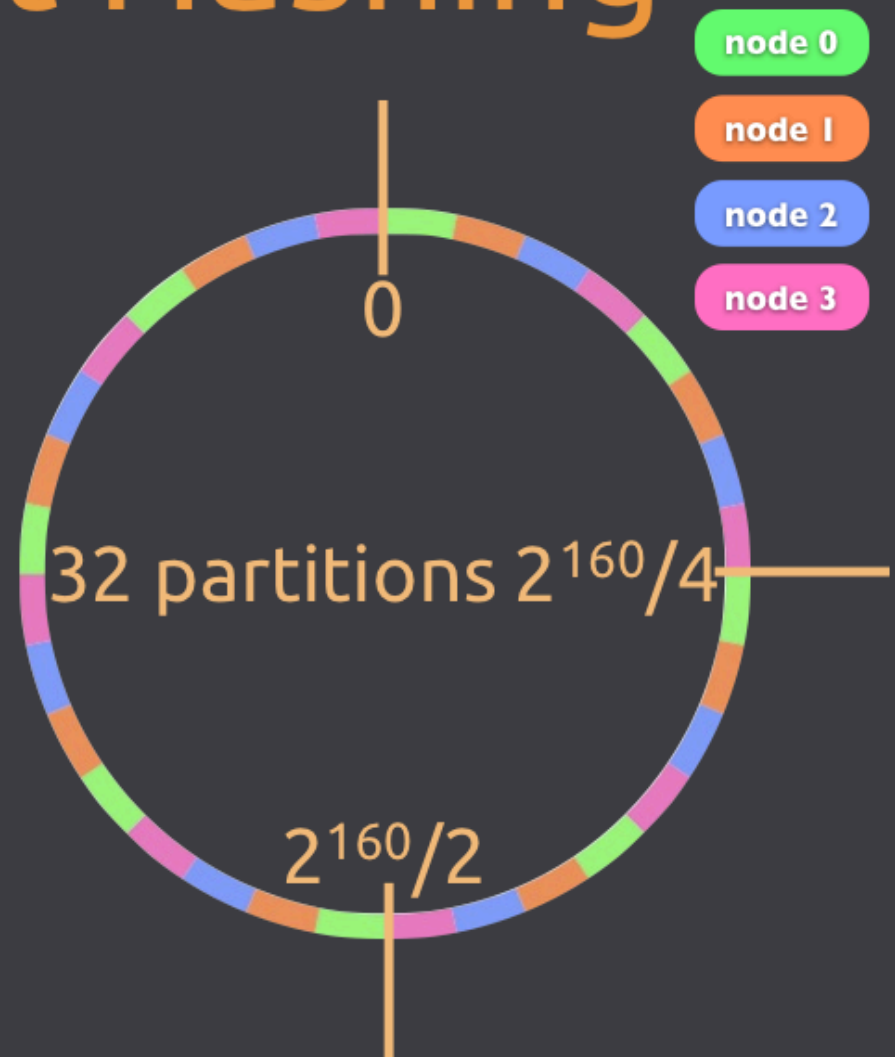
Consistent Hashing

- 160-bit integer keyspace
- divided into fixed number of evenly-sized partitions



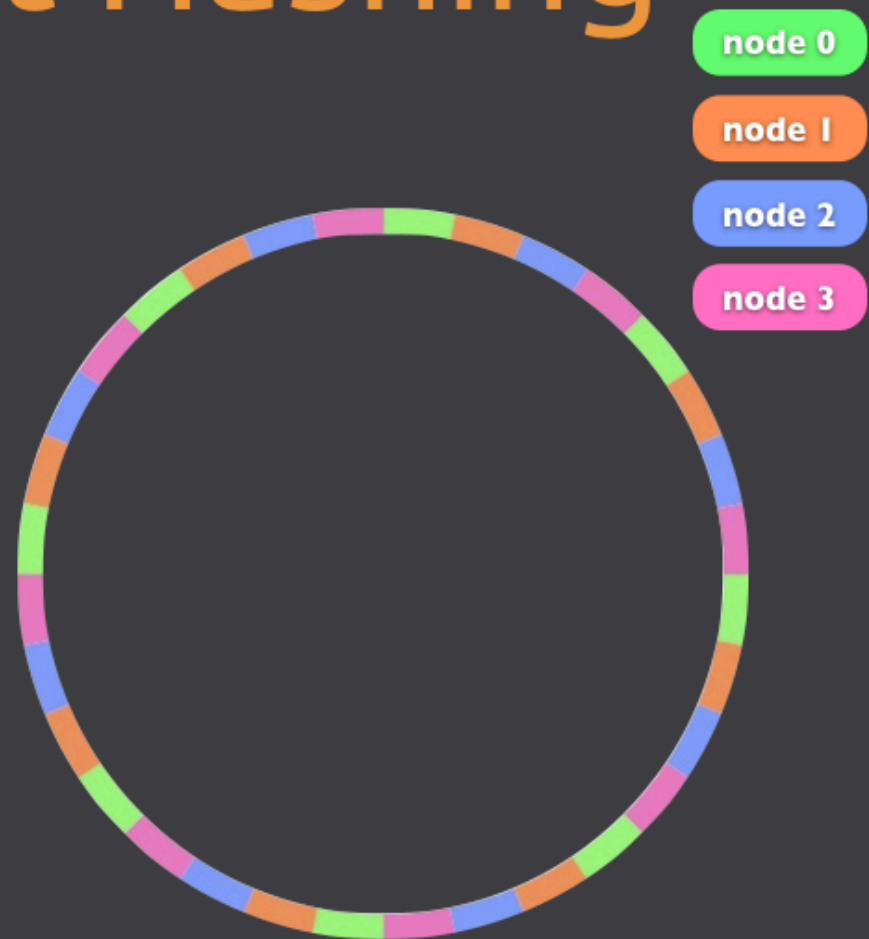
Consistent Hashing

- 160-bit integer keyspace
- divided into fixed number of evenly-sized partitions
- partitions are claimed by nodes in the cluster



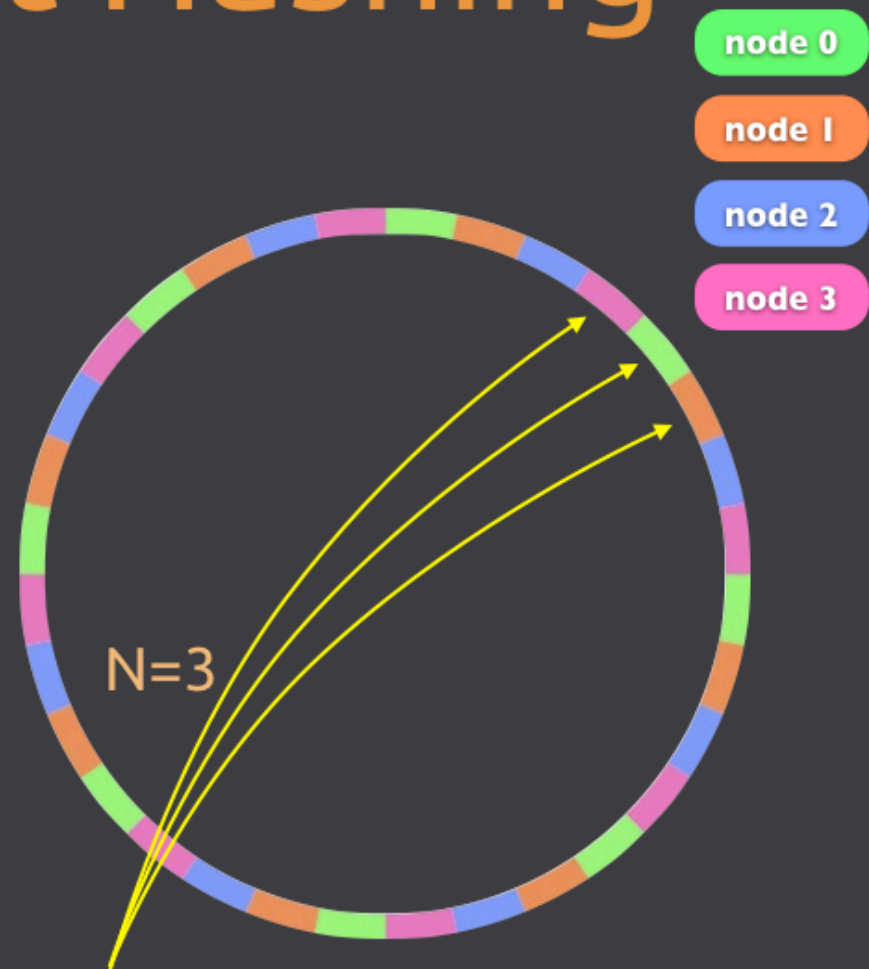
Consistent Hashing

- 160-bit integer keyspace
- divided into fixed number of evenly-sized partitions
- partitions are claimed by nodes in the cluster
- replicas go to the N partitions following the key



Consistent Hashing

- 160-bit integer keyspace
- divided into fixed number of evenly-sized partitions
- partitions are claimed by nodes in the cluster
- replicas go to the N partitions following the key



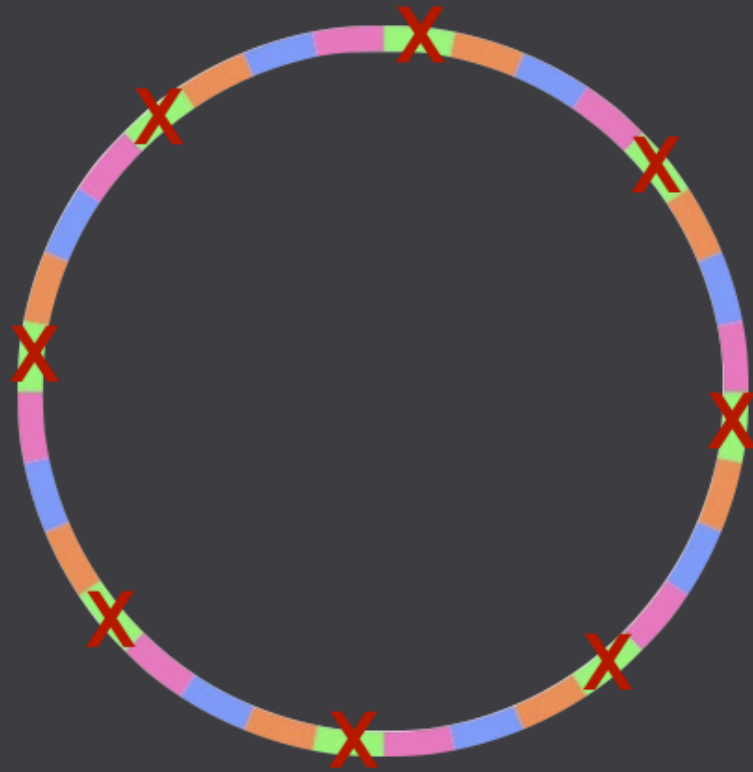
hash("meetups/nycdevops")

Disaster Scenario



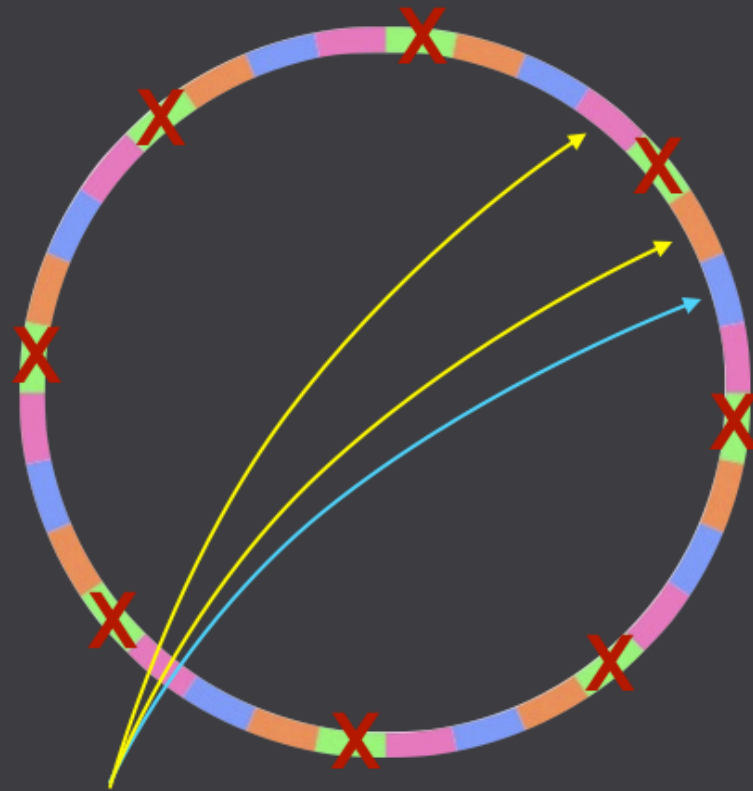
Disaster Scenario

- node fails



Disaster Scenario

- node fails
- requests go to fallback



`hash("meetups/nycdevops")`

Disaster Scenario

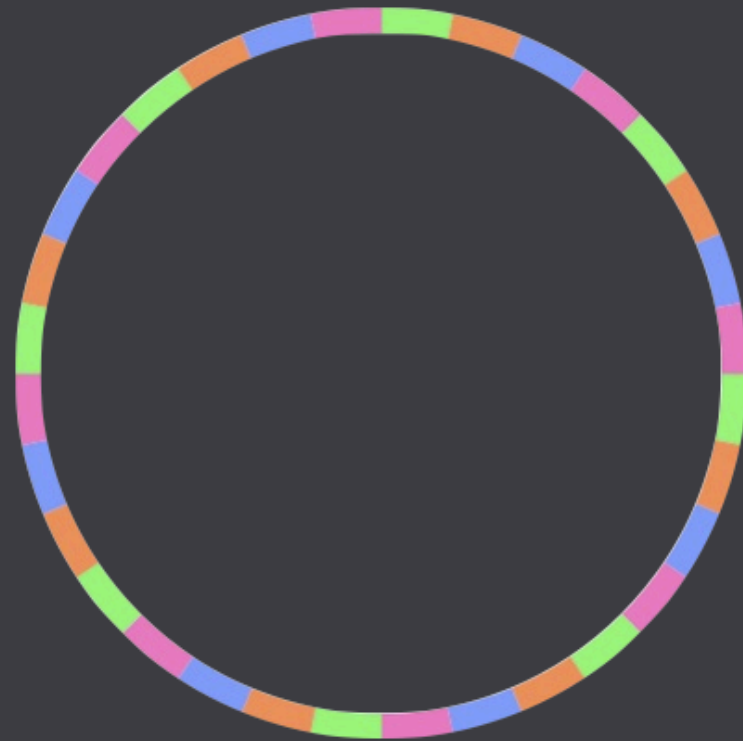
- node fails
- requests go to fallback
- node comes back



`hash("meetups/nycdevops")`

Disaster Scenario

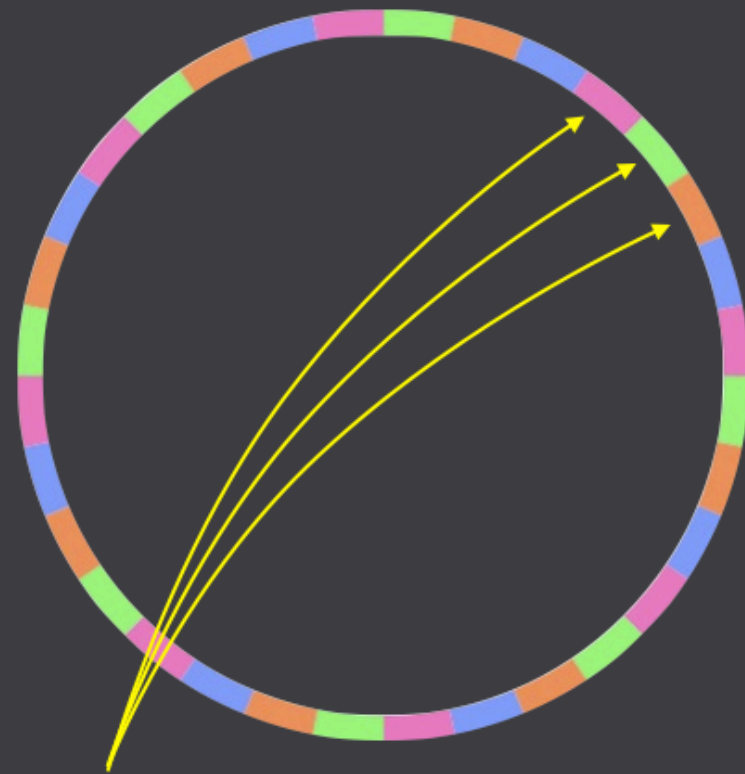
- node fails
- requests go to fallback
- node comes back
- "Handoff" - data returns to recovered node



`hash("meetups/nycdevops")`

Disaster Scenario

- node fails
- requests go to fallback
- node comes back
- "Handoff" - data returns to recovered node
- normal operations resume



`hash("meetups/nycdevops")`

Virtual Nodes

- Each physical machine runs a certain number of Vnodes
- Unit of addressing, concurrency in Riak
- Storage not tied to physical assets
- Enables dynamic rebalancing of data when cluster topology changes

Vector Clocks

- Data structure used to reason about causality at the object level
- Provides happened-before relationship between events
- Each object in Riak has a vector clock*
- Trade off space, speed, complexity for safety

Handoff and Rebalancing

- When cluster topology changes, data must be rebalanced
- Handoff and rebalancing happen in the background; no manual intervention required*
- Trade off speed of convergence vs. effects on cluster performance

Gossip Protocol

- Nodes “gossip” their view of cluster state
- Enables nodes to store minimal cluster state
- Can lead to network chatiness; in OTP, all nodes are fully-connected

Riak: when and why

When Might Riak Make Sense

When you have enough data to require >1 physical machine (preferably >5)

When availability is more important than consistency (think “critical data” on “big data”)

When your data can be modeled as keys and values; don't be afraid to denormalize

User/MetaData Store

- User profile storage for xfinityTV Mobile app
- Storage of metadata on content providers and licensing
- Strict Latency requirements



Notifications

The screenshot displays the Yammer notification module for a user named Jessica. The interface is divided into three main sections: a left-hand navigation sidebar, a central notification feed, and a right-hand community sidebar.

Left Sidebar: Features a user profile for Jessica, a 'MESSAGES' section with links to My Feed, Direct Messages, and Notifications (highlighted), and a 'COMPANY' section with links to Members, Groups, Topics, Invite, and Admin. Below that is an 'APPS' section with links to Leaderboards, Files, Images, Questions, Polls, Events, Ideas, and Org Chart.

Central Notifications: Titled 'Notifications', it lists four items:

- You were mentioned in a thread:** Sarah Schwartz (@Jessica Halper) asks when a powerpoint will be ready for a meeting on Friday (11 minutes ago).
- Phil Spitzer replied to your message:** Phil Spitzer replies to Jessica Halper, stating 'I think this is an excellent ideal' (12 minutes ago).
- Phil Spitzer likes your message:** Phil Spitzer likes Jessica Halper's message from 3 months ago about new product lines.
- Sarah Schwartz likes your message:** Sarah Schwartz likes Jessica Halper's message from 4 months ago about a marketing trip to Pepperdine University.

Right Sidebar: Includes a 'Community' section (private, created by Keith McCarty), 'Following Suggestions' (Drew Dillon, Tommy Vincent), 'Group Suggestions' (Accounting, Engineering), and 'Related Networks' (Yammer-inc.com, Geni.com, Workfeed.com, Dooms.day, Salmonellaville.com, Community.com). At the bottom, there is an 'Invite' section with an email input field and an 'Online Now' section showing several active users.

Yammer notification module powered by Riak

Session Storage

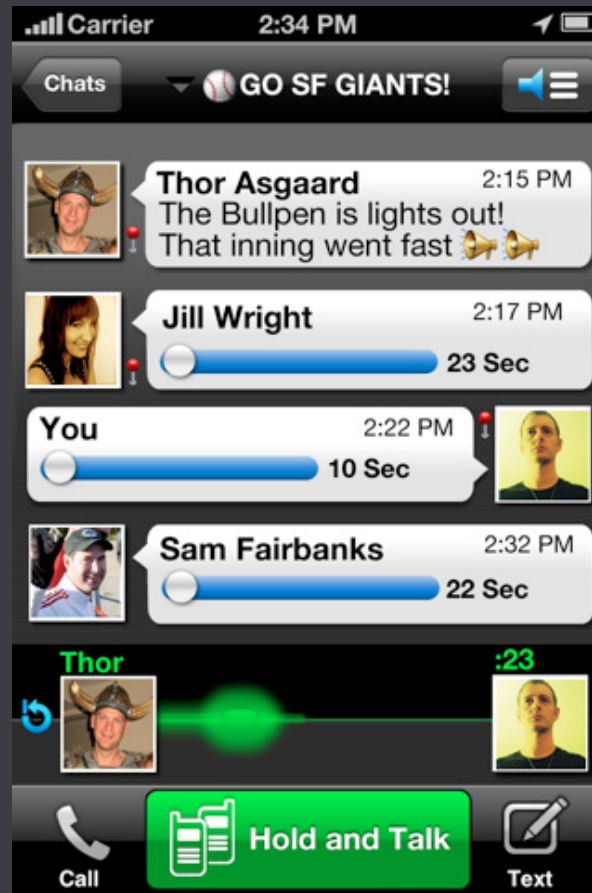
- First Basho customer in 2009
- Every hit to a Mochi web property results in at least one read, maybe write to Riak
- Unavailability or high latency = lost ad revenue



Ad Serving

- OpenX will serve ~4T ad in 2012
- Started with CouchDB and Cassandra for various parts of infrastructure
- Now consolidating on Riak and Riak Core

Riak for All Storage: Voxer



Voxer: Initial Stats

- 11 Riak nodes (switched from CouchDB)
- 100s of GBs
- ~20k Peak Concurrent Users
- ~4MM Daily Request

Walkie Talkie App Voxer Is Going Viral On iPhones And Androids, Trending On Twitter



A screenshot of a Twitter post from the user **sodmg.com** (@souljaboy). The tweet text is "Voxer. Soulja Boy." and includes a mention of "Thay SODMG". The tweet has 50+ retweets and 8 favorites. The interface shows a "Follow" button and a dropdown menu icon. Below the tweet, there are icons for Reply, Retweet, and Favorite.

sodmg.com
@souljaboy

Follow

Voxer. Soulja Boy.

Thay SODMG

50+ RETWEETS | 8 FAVORITES

5:02 AM - 6 Dec 11 via web · Embed this Tweet

Reply Retweet Favorite



Voxer: Post Growth

- ~60 Nodes total in prod
- 100s of TBs of data (>1TB daily)
- ~400k Concurrent Users
- Billions of daily Requests

Riak : Hybrid Solutions

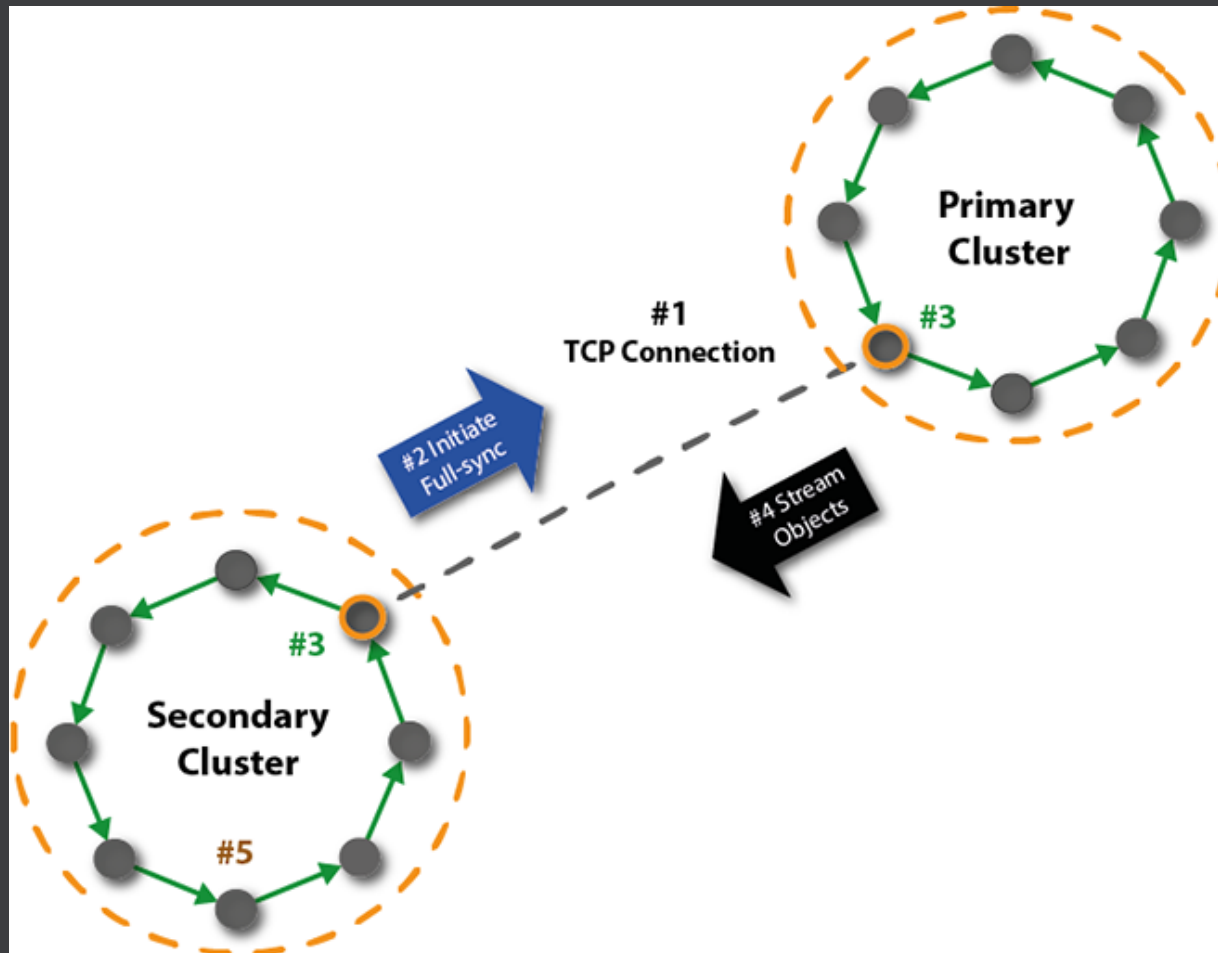
- Riak with Postgres
- Riak with Elastic Search
- Riak with Hadoop
- Secondary analytics clusters

Buy Some Software...

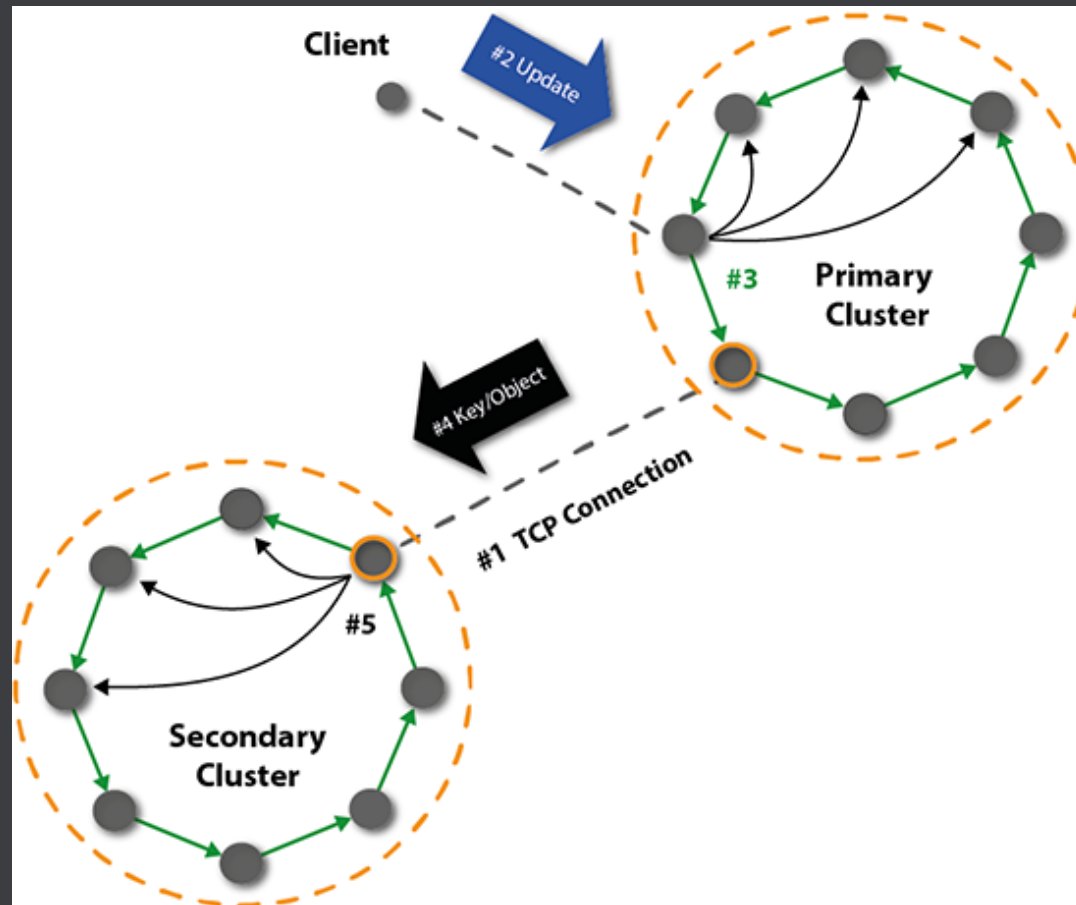
Riak Enterprise

- Multi-data center replication
- Real-time or full-time sync

Riak Enterprise: Full Sync



Riak Enterprise: Real-Time Sync



Riak Cloud Storage

- Large object support
- S3-compatible API
- Multi-tenancy
- Reporting on usage

Roadmap Stuff...

New in Riak 1.2

- LevelDB Improvements
- FreeBSD Support
- New Cluster Admin Tools
- Folsom for Stats
- KV and Search Repair work
- Much much more

Future Work

- Active Anti Entropy
- CRDTs
- Tight Solr integration
- Greater consistency
- Lots of other hotness



Riak

- docs.basho.com
- [@basho](https://twitter.com/basho)
- github.com/basho