

Riak Use Cases: Dissecting the Solutions to Hard Problems

Andy Gross <@argv0>

Chief Architect

Basho Technologies

Riak

- ✦ Dynamo-inspired key value database
 - ✦ with full text search, mapreduce, secondary indices, link traversal, commit hooks, HTTP and binary interfaces, pluggable backends
- ✦ Written in Erlang and C/C++
- ✦ Open Source, Apache 2 licensed
- ✦ Enterprise features (multi-datacenter replication) and support available from Basho

Choosing a NoSQL Database

- ✦ At small scale, everything works.
- ✦ NoSQL DBs trade off traditional features to better support new and emerging use cases
- ✦ Knowledge of the underlying system is essential
- ✦ A lot of NoSQL marketing is bullshit

Tradeoffs

- ✦ If you're evaluating Mongo vs. Riak, or CouchDB vs. Cassandra, you don't understand your problem
- ✦ By choosing Riak, you've already made tradeoffs:
 - ✦ Consistency for availability in failure scenarios
 - ✦ A rich data/query model for a simple, scalable one
 - ✦ A mature technology for a young one

Distributed Systems: Desirable Properties

- ✦ Highly Available
- ✦ Low Latency
- ✦ Scalable
- ✦ Fault Tolerant
- ✦ Ops-Friendly
- ✦ Predictable

1000s of Deployments



User/Metadata Store Comcast



User profile storage for xfinityTV mobile application

Storage of metadata on content providers, and content licensing info

Strict latency requirements

Notification Service

Yammer



FOUR LEAF CONSULTING

Welcome Jessica (edit)

MESSAGES

- My Feed
- Direct Messages
- Notifications**
- Community Feed
- More

COMPANY

- Members
- Groups
- Topics
- Invite
- Admin

APPS

- Leaderboards
- Files
- Images
- Questions
- Polls
- Events
- Ideas
- Org Chart

Notifications

You were mentioned in a thread:

Sarah Schwartz: @Jessica Halper when will the powerpoint be ready for our meeting on Friday?
11 minutes ago

[View thread »](#)

11 minutes ago

Phil Spitzer replied to your message:

Phil Spitzer in reply to Jessica Halper: I think this is an excellent idea!
12 minutes ago

[View thread »](#)

12 minutes ago

Phil Spitzer likes your message:

Jessica Halper in reply to Jesse Wilkinson: Personally, I think producing new product lines is the best strategy because it will help us expand our offering and makes us more competitive.
3 months ago
Liked by Phil Spitzer.

[View thread »](#)

12 minutes ago

Sarah Schwartz likes your message:

Jessica Halper ▶ **Marketing:** Heading down to Peppendine University tomorrow morning to film a video and attend the Social Media Garage meeting. Looking forward to the trip!
4 months ago
Liked by Sarah Schwartz.

[View thread »](#)

12 minutes ago

Community

This is a private community created by Keith McCarty.

Following Suggestions

- Drew Dillon**
Senior Sales Engineer
[Follow](#)
- Tommy Vincent**
Enterprise Business Representative
[Follow](#)

Group Suggestions

- Accounting**
[Join](#)
- Engineering**
[Join](#)

Related Networks

- Yammer-inc.com (parent)
- Geni.com
- Workfeed.com
- Dooms.day
- Salmonellaville.com
- Community.com

Invite [more](#)

Enter any email [Invite](#)

Online Now (8)

-

Yammer notification module powered by Riak

Session Store

Mochi Media



First Basho Customer (late 2009)

Every hit to a Mochi web property = 1 read,
maybe one write to Riak

Unavailability, high latency = lost ad revenue

Document Store

Github Pages / Git.io

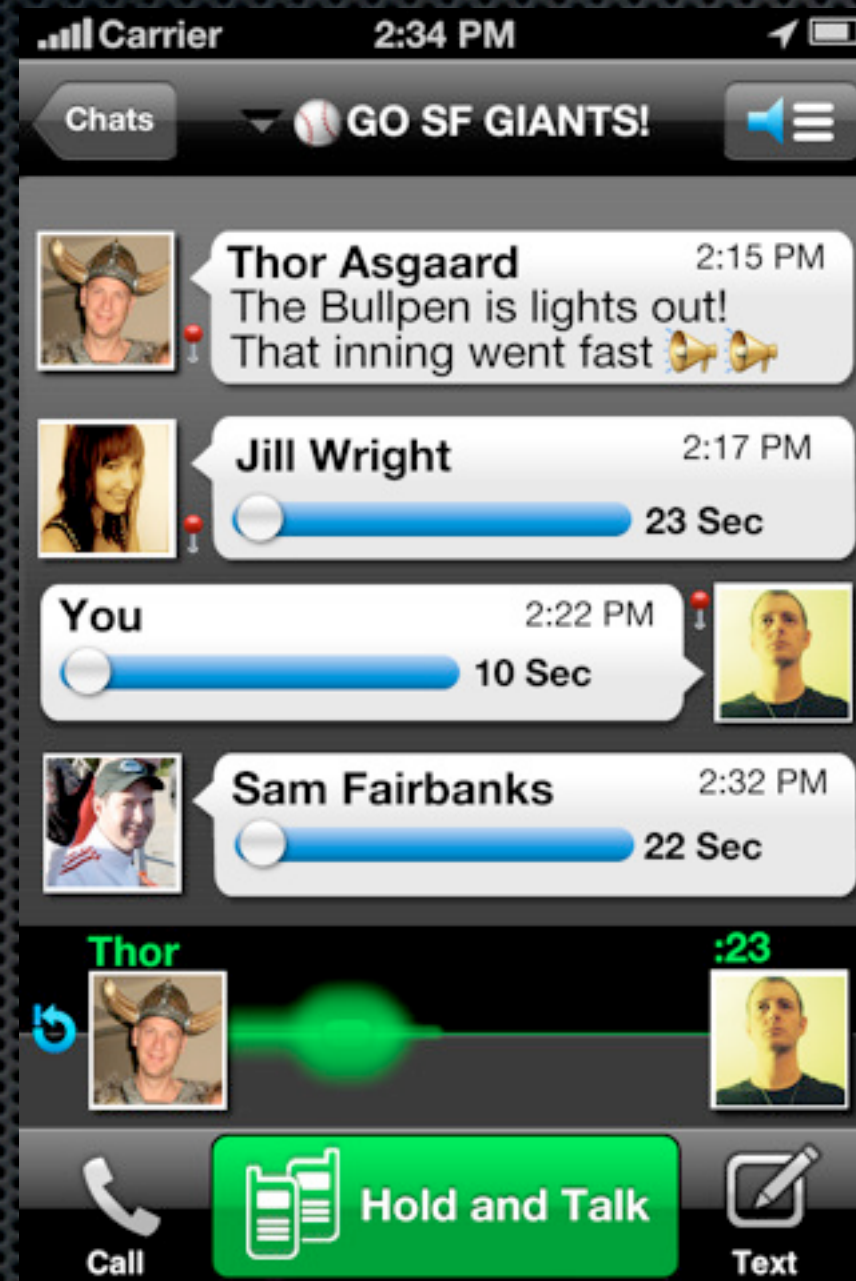


Riak as a web server for Github Pages

Webmachine is an awesome HTTP server!

Git.io URL shortener

Walkie Talkie Voxer



Voxer - Initial Stats

- ✦ 11 Riak Nodes
- ✦ ~500GB dataset
- ✦ ~20k peak concurrent users
- ✦ ~4MM daily requests

Then something happened...

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



Voxer - Current Stats

- ✦ > 100 nodes
- ✦ ~1TB data incoming / day
- ✦ > 200k concurrent users
- ✦ > 2 billion requests / day
- ✦ Grew from 11 to 80 nodes Dec - Jan

Distributed Systems: Desirable Properties

- ✦ High Availability
- ✦ Low Latency
- ✦ Horizontal Scalability
- ✦ Fault Tolerance
- ✦ Ops-Friendliness
- ✦ Predictability

High Availability

- ✦ Failure to accept a read/write results in:
 - ✦ lost revenue
 - ✦ lost users
- ✦ Availability and latency are intertwined

Low Latency

- ✦ Sometimes late answer is useless or wrong
- ✦ Users perceive slow sites as unavailable
- ✦ SLA violations
- ✦ SOA approaches magnify SLA failures

Who cares about latency?

The screenshot shows a Bloomberg news article. At the top, there's a 'MARKET SNAPSHOT' with indices for U.S., EUROPE, and ASIA. Below that, a 'Secret Fed Loans Gave Banks Undisclosed \$13B' article by Bob Ivry, Bradley Keoun, and Phil Kuntz is featured. The article text discusses the Federal Reserve's secret bailout of banks in 2008. To the right of the article is a 'More Stories' section with links to 'U.S. Stock-Index Futures Gain', 'Texas to Ask Supreme Court for Stay of Maps', 'Euro Advances After Report of IMF', and 'Asia Stocks, U.S. Futures Rally on Italy'. Below the article is an 'Advertisement' placeholder. Arrows from the 'SOA' label point to the 'U.S. Stock-Index Futures Gain' link, the 'Euro Advances After Report of IMF' link, the 'Asia Stocks, U.S. Futures Rally on Italy' link, and the 'Advertisement' placeholder.

MARKET SNAPSHOT

U.S. EUROPE ASIA

TOPIX 717.24 +10.64 1.51%

WANG 18,000.00 +340.43 +1.92%

08.70 +1.20% • EUR : USD 1.3326 0.6564% • Nasdaq 2,441.51 -0.75% • Dow 11,231.80 -0.23% • S&P 500 1,111.11 -0.11%

Bloomberg Our Company | Professional | Anywhere Visit Your Queue Sign in

QUICK NEWS VIEW MARKETS PERSONAL FINANCE SUSTAINABILITY TV RADIO Search News, Quotes, and OnlineQ

Related News: Economy • Law • Canada • U.S. • Bonds • Currencies • Finance • Insurance • Real Estate

Want to save this for later? Add it to your Queue!

Secret Fed Loans Gave Banks Undisclosed \$13B

By Bob Ivry, Bradley Keoun and Phil Kuntz - Nov 27, 2011 4:01 PM PT

Bloomberg Markets Magazine

Recommend 956
Share 74
Print
Email
Enlarge image

The **Federal Reserve** and the big banks fought for more than two years to keep details of the largest bailout in U.S. history a secret. Now, the rest of the world can see what it was missing.

The Fed didn't tell anyone which banks were in trouble so deep they required a combined \$1.2 trillion on Dec. 5, 2008, their single neediest day. Bankers didn't mention that they took tens of billions of dollars in emergency loans at the same time they were assuring investors their firms were healthy. And no one calculated until now that banks reaped an estimated \$13 billion of income by taking advantage of the Fed's below-market rates, Bloomberg Markets magazine

More Stories

- U.S. Stock-Index Futures Gain After Report on IMF Planning Loan to Italy
- Texas to Ask Supreme Court for Stay of Maps
- Euro Advances After Report of IMF
- Italy Loan Plan: Aussie, Kiwi Strengthen
- Asia Stocks, U.S. Futures Rally on Italy

Stories Rate These More News

Advertisement

SOA

Who cares about latency?



Sometimes high latency looks like an outage to the end user.

Fault Tolerance

- ✧ Everything fails
 - ✧ Especially in the cloud
- ✧ When a host/disk/network fails, what is the impact on
 - ✧ Availability
 - ✧ Latency
 - ✧ Operations staff

Predictability

“It’s a piece of plumbing; it has never been a root cause of any of our problems.”

Coda Hale, Yammer

Operational Costs

- ✦ Sound familiar?
 - ✦ “we chose a bad shard key...”
 - ✦ “the master node went down”
 - ✦ “the failover script did not run as expected...”
 - ✦ “the root cause was traced to a configuration error...”
- ✦ ***Staying up all night fighting your database does not make you a hero.***

Consistency, Availability, Latency

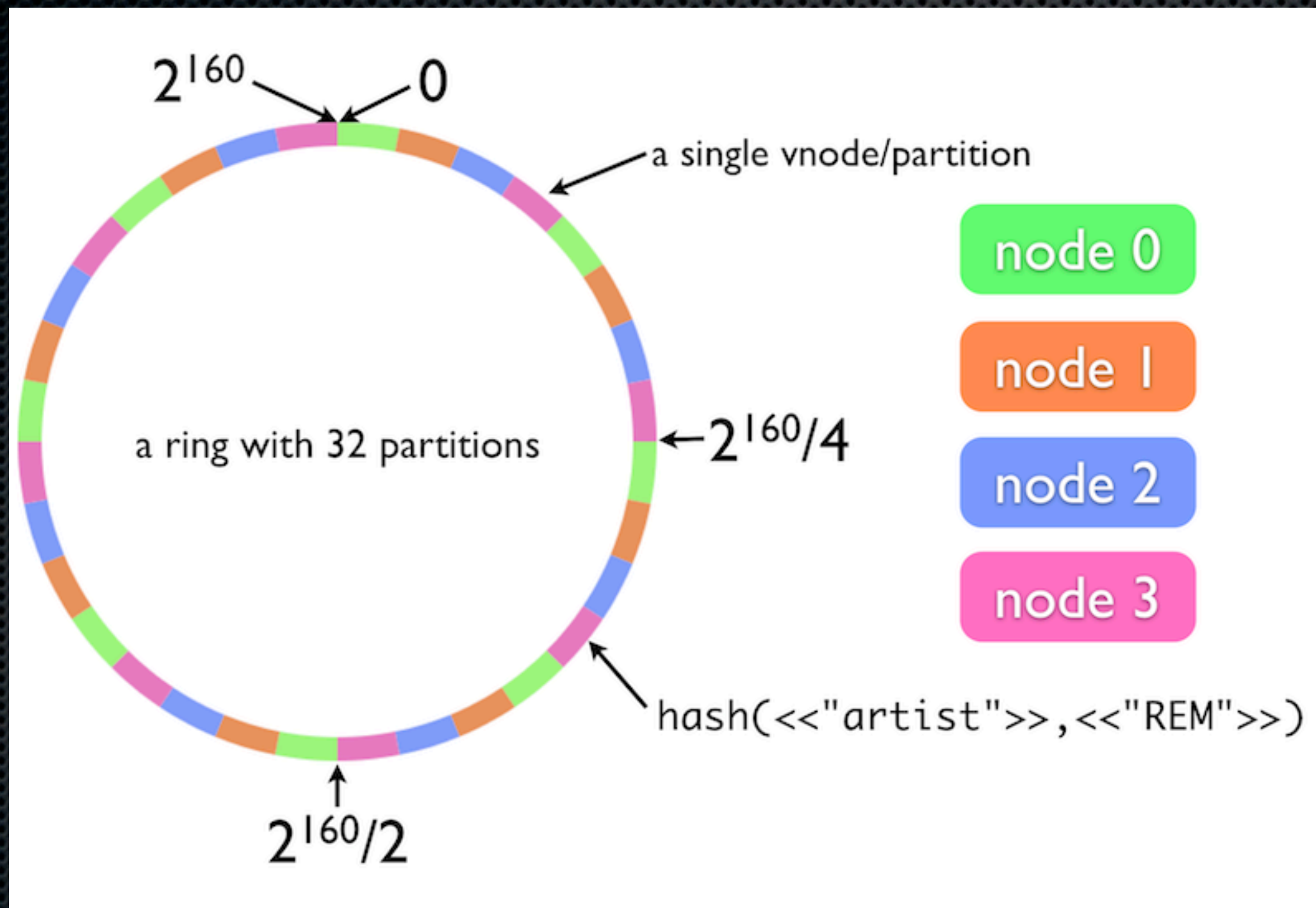
CAP

- ✦ The fundamental, most-discussed tradeoff
- ✦ When a network partition (message loss) occurs, laws of physics make you choose:
 - ✦ Consistency OR
 - ✦ Availability
- ✦ No system can “beat the CAP theorem”

Data Distribution

- ✧ Location of data is determined based on a hash of the key
- ✧ Provides even distribution of storage and query load
- ✧ Trades off advantages gained from locality
 - ✧ range queries
 - ✧ aggregates

Consistent Hashing



Virtual Nodes

- ✦ Unit of addressing, concurrency in Riak
- ✦ Each host manages many vnodes
- ✦ Riak **could** manage all host-local storage as a unit and gain efficiency, but would lose
 - ✦ simplicity in cluster resizing
 - ✦ failure isolation

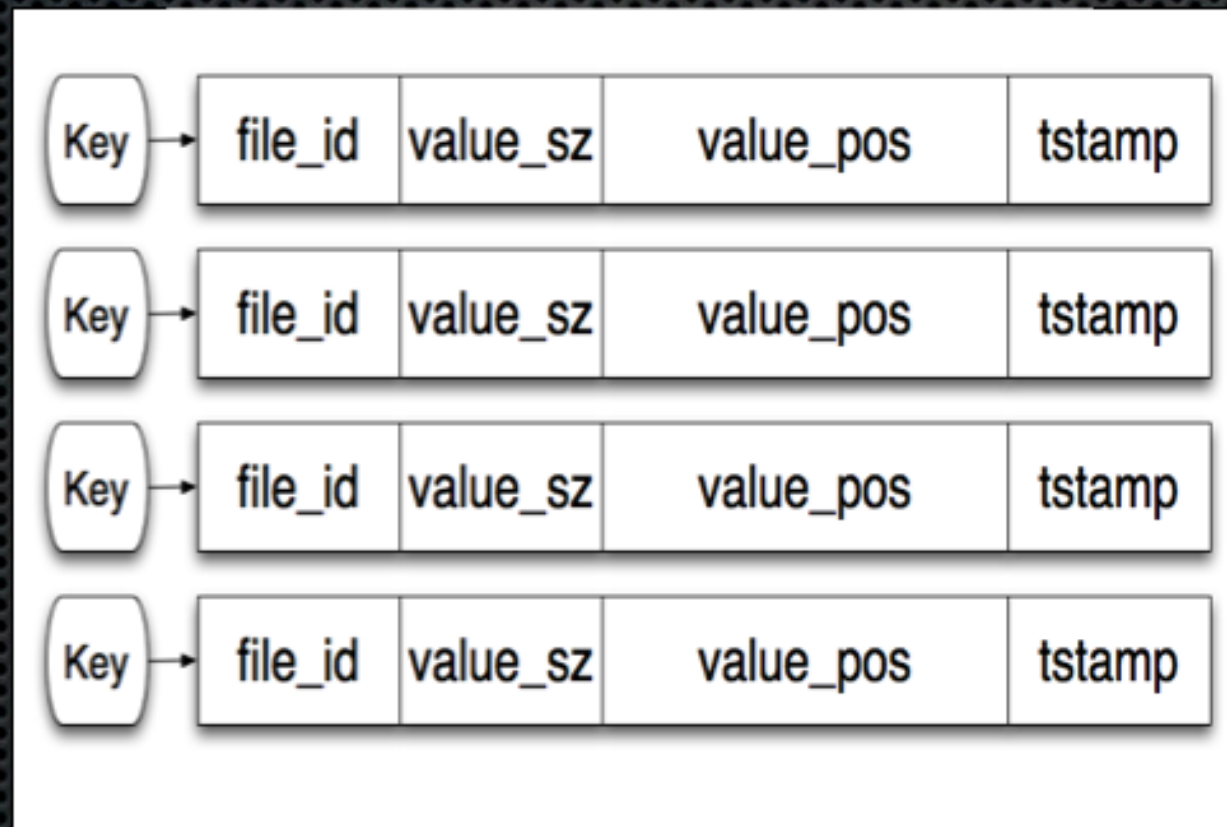
Append-Only Stores, Bitcask

Append-Only Stores

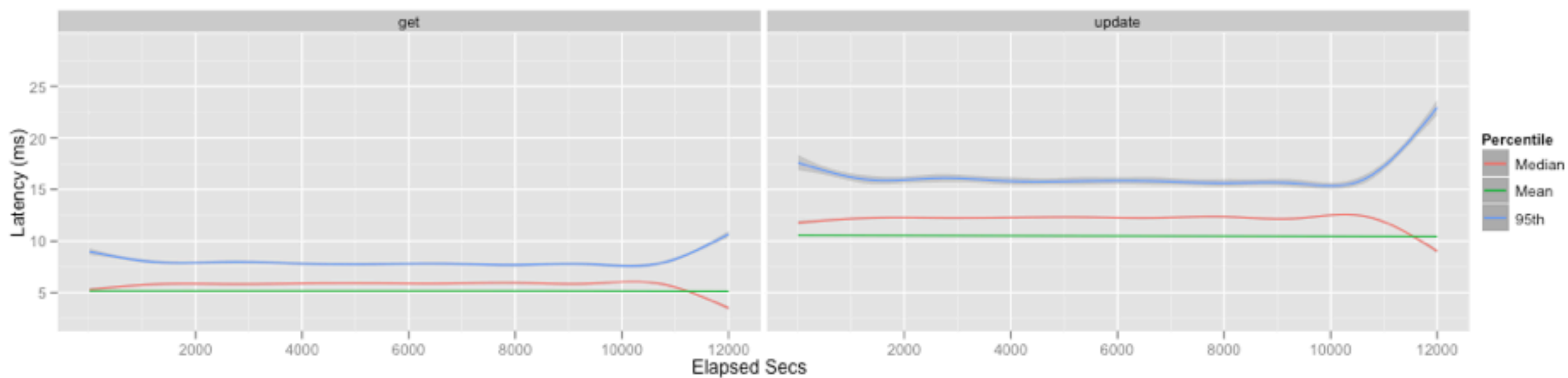
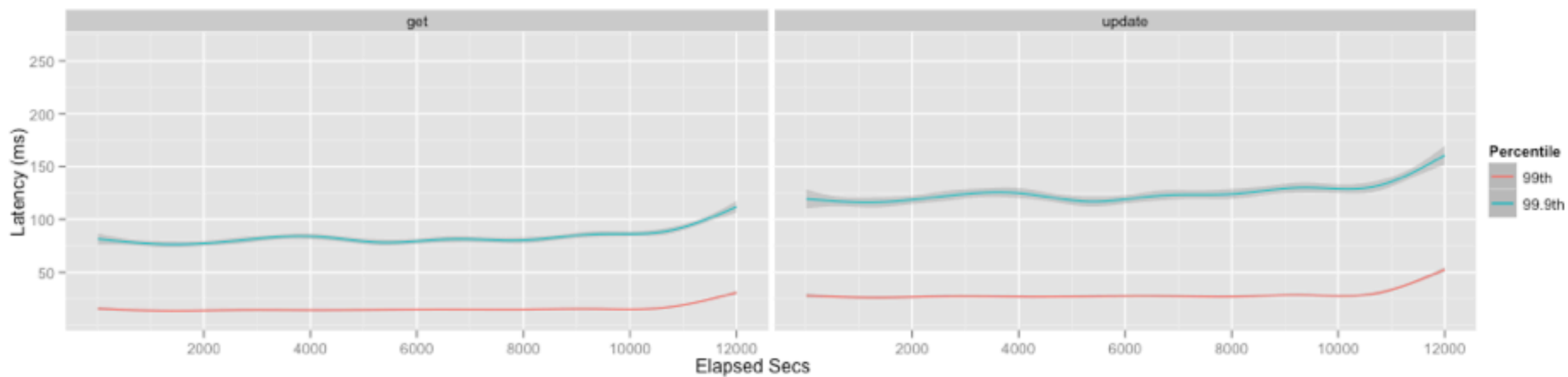
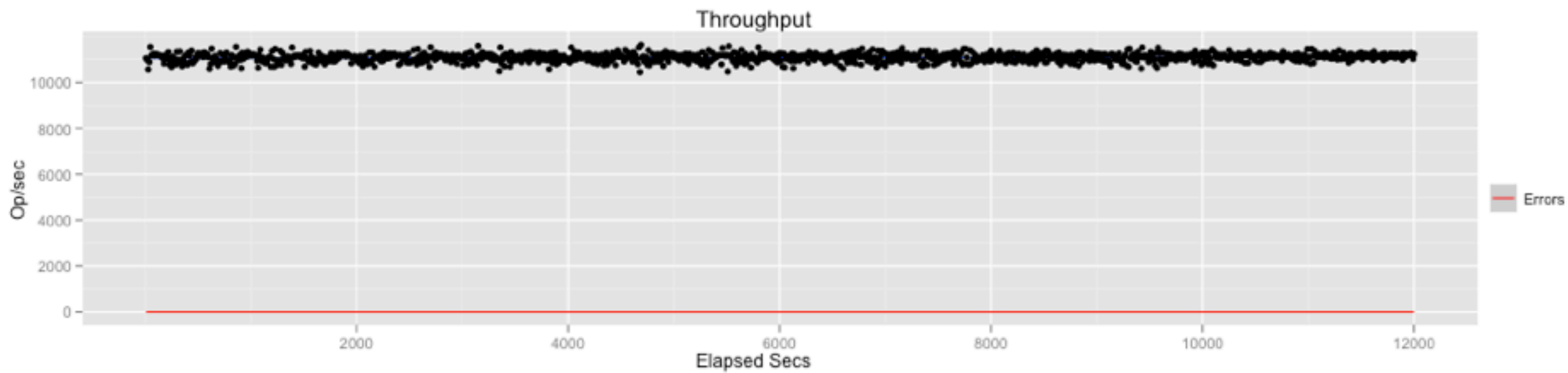
- ✦ All writes are appends to a file
- ✦ This provides crash-safety, fast writes
- ✦ Tradeoff: must periodically compact/merge files to reclaim space
 - ✦ Causes periodic pauses while compaction occurs that must be masked/mitigated

Bitcask

Low Latency: All reads = hash lookup + 1 seek
All writes = append to file



Tradeoff: Index must fit in memory



Handoff and Rebalancing

- ✦ When nodes are added to a cluster, data must be rebalanced
- ✦ Rebalancing causes disk, network load
- ✦ Tradeoff: speed of convergence vs. effects on cluster performance

Vector Clocks

- ✦ Provide happened-before relationship between events
- ✦ Riak tags each object with vector clock
- ✦ Tradeoff: space, speed, complexity for safety

Gossip Protocol

- ✦ Nodes “gossip” their view of cluster state to each other
- ✦ Tradeoffs:
 - ✦ atomic modifications of cluster state for no SPOF
 - ✦ complexity for fault tolerance

Sane Defaults

- ✦ Speed vs. Safety
- ✦ Riak ships with $N=3$, $R=W=2$
 - ✦ Bad for microbenchmarks, good for production use, durability
- ✦ Mongo ships with $W=0$
 - ✦ Good for benchmarks, horrible and insane for durability, production use.

Erlang

- ✦ Best language ever:
 - ✦ for distributed systems glue code
 - ✦ for safety, fault tolerance
- ✦ Sometimes you want:
 - ✦ Destructive operations
 - ✦ Shared memory

NIFs to the rescue?

- ✦ Use NIFs for speed, interfacing with native code, but:
 - ✦ You make the Erlang VM only as reliable as your C code
 - ✦ NIFs block the scheduler

Conclusions

- ✦ Over time, operational costs dominate
- ✦ Predictability in:
 - ✦ Latency
 - ✦ Scalability
 - ✦ Failure scenarios
- ✦ ...is essential for managing operational costs
- ✦ When choosing a database, raw throughput is often the *least* important metric.

Thanks!

- ✦ Visit us at <http://www.basho.com>
- ✦ Check out our open source code at <http://github.com/basho>
- ✦ Follow us on Twitter: @basho
- ✦ We're hiring!