A TECHNICAL OVERVIEW OF MULTI-CLUSTER REPLICATION

INTRODUCTION

Cluster replication is a critical part of modern infrastructure, providing essential business benefits for enterprise applications, platforms, and services. Riak® KV Enterprise and Riak® TS Enterprise offer Multi-cluster Replication so that Riak Clusters can be replicated for backup-clusters, analysis clusters, or for multiple datacenters. With Riak Multi-cluster Replication, data can be replicated across the datacenter or across geographic areas, providing disaster recovery, data locality, compliance with regulatory requirements, the ability to "burst" peak loads into public cloud infrastructure, and more.

This whitepaper introduces Riak Multi-cluster Replication and common use cases for this functionality. It provides an overview of features and architecture, as well as considerations for developers and operators. It also looks at some common configurations for Multi-cluster Replication, including backup/failover clusters, data geo-location, availability zones, secondary analytics clusters, and bursting from private to public cloud environments for peak loads or disaster recovery. Also, for Internet of Things (IoT) applications, edge analytics close to the device opens up new insights for many industries.

FEATURES & ARCHITECTURE

HOW IT WORKS

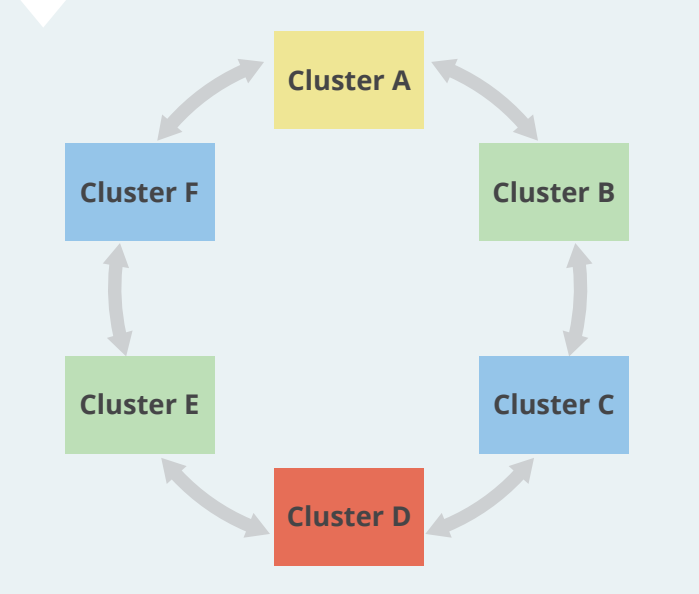
In Multi-cluster Replication, one cluster acts as a primary, or source, cluster. The primary cluster handles replication requests from one or more secondary, or sink, clusters (often located in datacenters in other regions or countries). If the datacenter with the primary cluster goes down, a secondary cluster can take over as the primary cluster. In this sense, Riak's multi-cluster capabilities are "masterless."

In Multi-cluster Replication, there are two primary modes of operation: fullsync and real-time. In fullsync mode, a complete synchronization occurs between primary and secondary cluster(s). In real-time mode, continual, incremental synchronization occurs — replication is triggered by new updates. Fullsync is performed upon initial connection of a secondary cluster, and then periodically (by default, every 360 minutes). Fullsync is also triggered if the TCP connection between primary and secondary clusters is severed and then recovered.

Multiple concurrent TCP connections (approximately one per physical node) and processes are used to maximize performance and network utilization. This replication capability supports fullsync and real-time sync between clusters, as well as their simultaneous operation. By default, this connection is unidirectional, however, bidirectional replication can be achieved by establishing two unidirectional connections between clusters (where a cluster is both source and sink). Replication can be configured for all data in the cluster or on a per-bucket basis, which allows for replication of a subset of data.

CASCADING WRITES

Riak Multi-cluster Replication includes a feature that cascades real-time writes across multiple clusters. Cascade tracking is a simple list of where an object has been written. This works well for most common configurations, however larger installations may have writes cascade to clusters that other clusters have already written to. Consider the diagram below:



In this Multi-cluster Replication configuration, a write at cluster A will begin two cascades. One goes to B, C, D, E, and end at F; and the other goes to F, E, D, C, and ends at B. Each cascade will loop around to A again, sending a replication request even if the same request has already occurred from the opposite direction. The result is three additional write requests.

Disabling cascading writes at a cluster will mitigate this issue. If cascading was disabled on cluster D, a write at A would begin two cascades. One would go through B, C, and D; and the other through F, E, and D. This reduces the number of extraneous write requests to one.

Riak Enterprise Multi-cluster Replication allows for flexible topologies that can be further utilized to address the issue of write duplication. A sample topology can be found in the Riak KV Enterprise Documentation under the section entitled Cascading Real-time Writes.

ARCHITECTURAL STRATEGIES

This section reviews the common use cases and architectural strategies for Riak Multi-cluster Replication.

PRIMARY CLUSTER WITH FAILOVER

One of the most common architectural patterns for Multi-cluster Replication is maintaining a primary cluster that serves traffic and a backup cluster for emergency failover. Maintaining a backup cluster can also be an important component of regulatory compliance and/or ensuring business continuity during an adverse event.

In this configuration, a primary cluster serves as the production cluster from which all read and write operations are served. A backup cluster is maintained in another datacenter. In the event of a datacenter outage or critical failures at the primary site, requests can be directed to the backup cluster either by changing the DNS configuration or allowing your load balancer to route traffic to the healthy cluster.

Depending on business requirements, operators may use either fullsync or real-time sync for replicating data from the primary to the backup cluster. For some use cases, keeping data up to date within a 24-hour period is sufficient, however, sometimes more frequent syncs may be required. If hot failover is required, real-time sync can be used. Maintaining continuous real-time sync both keeps the data more up to date on the secondary, and reduces the load/time required for a fullsync operation.

The failover strategy should also be fully defined upfront:

1. Specify the conditions in which a failover mode will be invoked
2. Decide how traffic will be directed to the backup cluster
3. Document and test the failover strategy to ensure success.

It is also recommended that the failover strategy be tested periodically so any potential issues can be resolved in advance of a crisis.

ACTIVE-ACTIVE CLUSTER CONFIGURATION

For some use cases, it may be desirable to maintain two (or more) active, synced clusters that are both responsible for serving data to clients. This approach is generally used to achieve data locality in cases where clients are served at low latency by whichever datacenter is nearest to them. An added benefit of this approach is business continuity in the event of a datacenter failure — all traffic can be served from one cluster if the other cluster goes down.

Bidirectional replication is often desirable in this configuration to ensure that both clusters have all data and updates. For data locality, requests can be load balanced across geographies with client requests directed to the closest datacenter. For situations where not all data needs to be shared across all datacenters (or if certain data, such as user data, must only be stored in a specific geographic region to meet privacy regulations), Multi-cluster Replication can be configured on a per-bucket basis so only shared assets and popular assets are replicated.

AVAILABILITY ZONES

Availability zones are common for enterprises providing cloud services, either as a public platform or for internal consumption. Availability zones provide efficient Multi-cluster Replication and data redundancy within a geographic region (such as a coast or a country). In this configuration, data is replicated within an availability zone's series of datacenters. In the event that one of datacenters experiences an outage or serious failure, data can still be served from other datacenters within the same region.

There are multiple approaches to setting up availability zones using Riak Enterprise. One approach is to have a primary site in a region to which all reads and writes for specific users, applications, or data sets are directed. This primary cluster can then be replicated to one or more proximal secondary clusters, either using real-time sync or fullsync, depending on the business case.

In other approaches, data can be replicated in real-time from one cluster to both another datacenter and other cold backups maintained for emergency conditions. The right approach is dependent on your business requirements and constraints, including availability, data criticality, and bandwidth and infrastructure costs. If this structure is relevant to your business use case, please contact us to further discuss the best way to meet your needs.

SECONDARY ANALYTICS CLUSTERS

As mentioned earlier, many Riak Enterprise users need to serve heavy production traffic and perform other computationally intensive tasks such as MapReduce. Since the request patterns of writing and reading data differ significantly from distributed search, analytics, and aggregation tasks, performing both types of computation on the same cluster isn't ideal. Analytics workloads can cause degraded performance and periods of higher latency for regular GET, PUT, and DELETE operations.

An alternative to running these workloads on the same cluster is to replicate data from the primary cluster, which is responsible for serving all production requests, to a secondary cluster on which analytic and other computations can be performed. Replication can be configured to occur on a given interval depending on the nature and temporal requirements of the analytics tasks. Additionally, all of the data or only some of the data (via per-bucket replication settings) can be replicated depending on your needs.

To decide if a secondary analytics cluster is right for your use case, it is necessary to determine the profile of production traffic on your system as well as the profile of the analytic/aggregation tasks you must perform. This will determine if both workloads can be handled on the same cluster with the desired performance and latency results. Please get in touch to learn more and gain access.

PUBLIC CLOUD USE CASES

Public clouds are becoming a critical component of enterprise infrastructure. Riak is designed to be easy to use and operate on public clouds, and is partnered with many of the leading cloud providers.

There are several use cases for Riak's Multi-cluster Replication in the public cloud. Many enterprises want to maintain a cold or hot backup of their cluster in a public cloud for business continuity in the event of a datacenter outage in their private infrastructure. For this use case, please see the earlier section on Primary Cluster with Failover.

For other customers, the public cloud can provide a more cost-effective way of meeting peak loads, rather than building out private infrastructure to accommodate them. For example, many retailers and media providers need to offer increased capacity over the holiday season. Riak Enterprise is used to scale out capacity on public clouds over these periods, either with fullsync or real-time sync depending on the business needs.

Finally, some enterprises run certain applications or services entirely on public clouds. For these users, redundancy and data locality across public cloud availability zones is necessary for optimal performance and resiliency. Riak Enterprise's Multi-cluster Replication allows clusters to be easily replicated across availability zones.

NEXT STEPS

Full documentation for Riak KV Enterprise is available at http://docs.basho.com.

If you are evaluating Riak Enterprise and are interested in Multi-cluster Replication, please contact us. We would be happy to arrange a tech talk with your team, or answer any questions about our product and how customers are using it in production to meet their business goals. If you want to try Riak Enterprise, we can provide a free developer trial that you can set up on your own hardware and evaluate on your own time. Finally, our Professional Services Team can assist you in planning, setting up, and optimizing your multi-datacenter strategy.

ABOUT BASHO TECHNOLOGIES

Basho, the creator of the world's most resilient databases, is dedicated to developing disruptive technology that simplifies enterprises' most critical distributed systems data management challenges. Basho has attracted one of the most talented groups of engineers and technical experts ever assembled devoted exclusively to solving some of the most complex issues presented by Big Data and IoT. Basho's distributed database, Riak® KV, the industry leading distributed NoSQL database, is used by fast growing Web businesses and by one-third of the Fortune 50 to power their critical Web, mobile and social applications. Built on the same foundation, Basho introduced Riak TS, which is the first enterprise-ready NoSQL database specifically optimized to store, query and analyze time series data. The Basho Data Platform helps enterprises reduce the complexity of supporting Big Data applications by integrating Riak with Apache Spark, Redis, and Apache Solr.