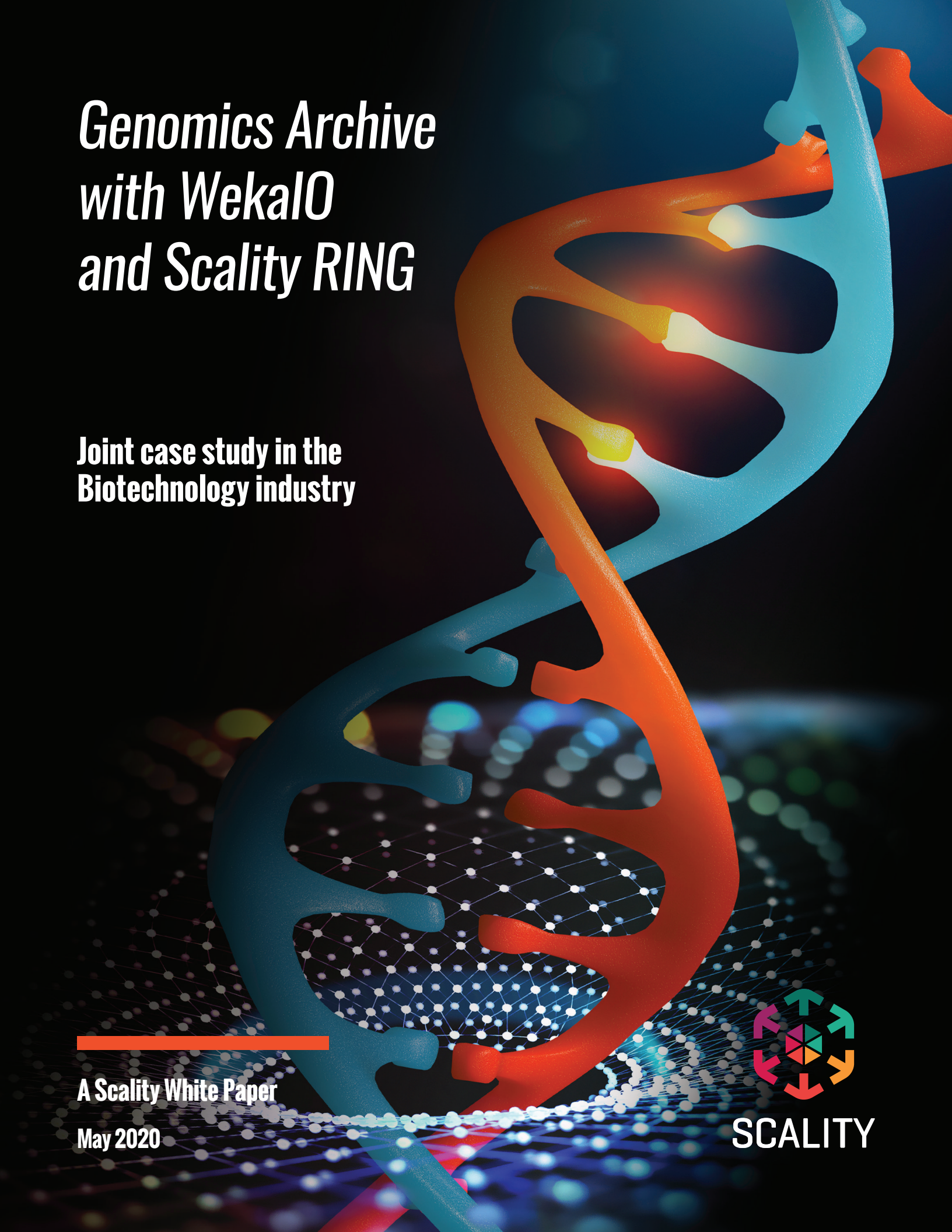
Genomics Archive with WekaIO and Scality RING

Joint case study in the Biotechnology industry



A Scality White Paper

May 2020

SCALITY

Introduction **3**

Hybrid Cloud Use Cases **4**

Scality concepts and definitions: 4

Case Study **5**

Business problem and previous architecture 5

New hybrid cloud architecture 5

Additional architectural details 9

Business benefits of Hybrid Cloud architecture 9

Hybrid Cloud DR Considerations **10**

Conclusion **11**

RING8

Introduction

Having a business continuity plan is crucial for virtually all organizations, but maintaining a secondary datacenter does not always make sense—if it is even an option at all. The costs of building (or leasing), staffing and operating a datacenter dedicated to disaster recovery (DR) alone probably do not make sense for most businesses. However, the combination of on-premises infrastructure and the public cloud enables hybrid cloud use cases that are changing the way enterprise IT leaders think about data management and data protection.

Most enterprises deploy mission-critical applications that depend on continuous access to data in order to maintain business processes. The critically important nature of data and the need to provide continuous access forces many enterprises to consider storing data in multiple, geographically-distributed datacenter locations. These multiple copies provide assurance that application data remains available even in the event of a primary datacenter outage or loss. For most businesses, this has historically meant maintaining two physical datacenter locations with synchronization of data between the two sites.

Replicating on-premises data to the public cloud, coupled with the ability to quickly spin up application components in the cloud in the event of an on-premises failure, provides a variety of DR options that range from immediate failover to longer recovery time objectives for less critical workloads. Hybrid cloud DR provides the option to have multiple sites that are geographically remote from the primary site. Additionally, the OPEX costs of a hybrid cloud DR solution may be much more affordable than the CAPEX costs of a more traditional on-premises DR only architecture depending on how often the cloud resources need to be spun up.

This paper provides an overview of a real-world Scality RING hybrid cloud DR solution of genomics data in the biopharmaceutical industry and some considerations enterprises should take into account when considering a hybrid cloud DR architecture.

Hybrid Cloud Use Cases

As Scality's customers incorporate more and more public cloud services into their IT operations, they seem to implement the following three hybrid cloud data management use cases most frequently:

- **Disaster Recovery:** maintaining a copy of on-premises data in the cloud for business continuity.
- **Archiving:** transitioning on-premises data to cheap cloud storage for long-term archive and data retention.
- **Service Bursting:** replicating a subset of on-premises data to the cloud to leverage the flexibility and elasticity of cloud services for peak demand of compute intensive workloads such as artificial intelligence, machine learning, media rendering/transcoding.

While the remainder of this paper will focus on the DR use case, both archiving and service bursting are supported architectures as well.

Scality concepts and definitions:

Scality RING: software-defined, scale-out file and object storage. RING deploys on any standard x86 servers and acts as a single, distributed system that scales across thousands of servers, multiple sites, and an unlimited number of objects. Recognized as a leader in file and object storage by both IDC and Gartner, the RING provides high performance across a variety of workloads at up to 90% lower TCO than legacy storage.

Extended Data Management (XDM): a set of RING features that enable multi-site replication of data across RING and public cloud storage services. XDM supports Amazon Web Services, Microsoft Azure, Google Cloud Platform, Wasabi and other public clouds. The XDM component can be installed separately and even in a different location from the core RING services; a key enabling feature for some hybrid cloud DR architectures.

In-band and Out-of-band: architectures where the RING XDM component is the target of a given application and in the data path (In-band) or where a storage location (RING, public cloud, etc.) is the target of a given application and XDM is outside the data path (Out-of-band). In the latter architecture, XDM is able to detect updates to existing data from a storage location and replicate that data to other locations. As illustrated later in this paper, hybrid cloud DR with RING can be implemented with XDM either in-band or out-of-band.

RING8

Case Study

Biopharmaceutical enterprise uses RING for Big Data Hybrid Cloud DR

This case study is about a global biotechnology company that invests in scientific innovation to create transformative medicines for people with serious and life-threatening diseases.

Business problem and previous architecture

A genome is a person's complete set of DNA. Almost every cell in our bodies contains a full copy of the approximately three billion DNA base pairs of which the human genome is made. A key component of this biopharmaceutical company's business is analyzing and researching genomic data - data which can grow exponentially based on the research project, number of data sets, number of participants, etc. Time to market with new research and medical solutions is critical, yet, like any business, the biopharmaceutical company is always looking to optimize spending and reduce costs where possible.

The biopharmaceutical company's previous architecture consisted of two on-premises datacenters. EMC Isilon was used for file-based storage. Traditional backup infrastructure and a high-performance file system were also in place. The Isilon storage was at 100% capacity which prompted the company to consider a new architecture.

As the amount of data grew beyond single digit petabytes, the company was running into issues scaling and managing the file-based Isilon storage. As storage capacity was nearing 100%, backups were taking too long, and too much effort and operational costs were going into managing and flattening the ever-increasing file system. The fact that capacity was at 100% was also hindering the amount of genomic data that could be added to the system, hence slowing down or limiting the number of research studies that could be worked on in parallel.

What architecture would allow the biopharmaceutical company to complete research projects faster and at a lower cost?

New hybrid cloud architecture

The company decided to switch to a new, modern architecture for the following reasons:

- They wanted the ability to leverage the cloud for disaster recovery, compute bursting, cheap archive storage, and other use cases.

RING8

- They wanted to move away from CAPEX of owning and operating a secondary datacenter to OPEX of "renting" cloud services.
- They wanted a more scalable, object-based storage solution that could grow easily beyond multiple PB, without the limitations of traditional file storage.

In response to those requirements, this biopharmaceutical company replaced its previous architecture with Scality RING object storage and WekaIO's WekaFS high-performance HPC file system.

The use case is big data collection for genomics data that is ingested via WekaIO. The new hybrid cloud architecture is a combination of on-premises infrastructure in a single datacenter and Amazon Web Services resources. The deployment is a large WekaIO cluster (22 nodes). RING serves as the archive for WekaIO with multiple PB of data in production and capacity projected to grow to 10 PB.

For the biopharmaceutical company, hybrid cloud DR is accomplished by replicating RING data to Amazon S3 and maintaining "warm" VMs with WekaIO in Amazon EC2. The data replication is out-of-band, so data is written to RING initially and then read and replicated by the XDM component which is also running in EC2.

As illustrated below, normal operating conditions have WekaIO writing data to RING. Network traffic and failure detection is managed by Global Server Load Balancing (GSLB).

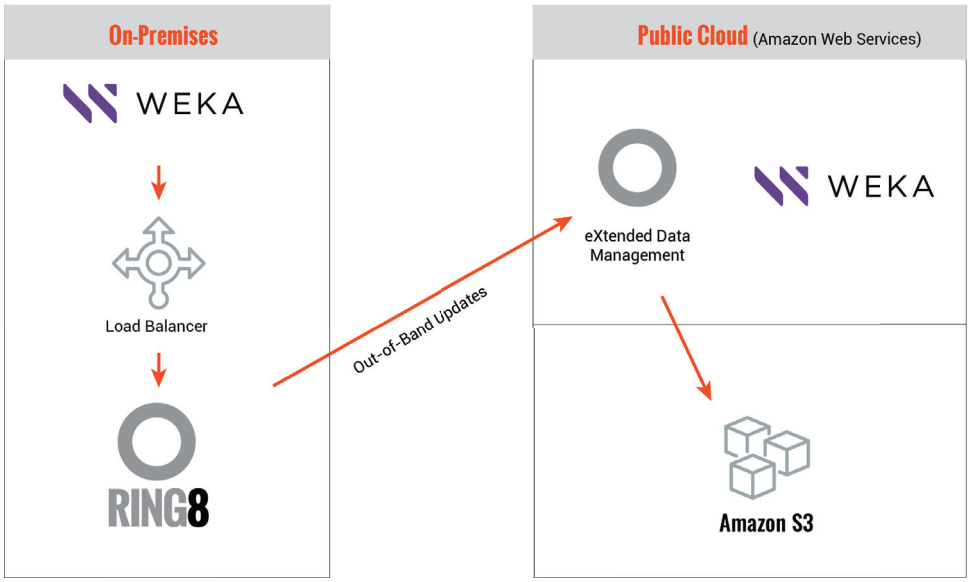


Figure 1. Out-of-band Normal Conditions

In the event of a connectivity loss to RING for whatever reason (network, hardware, etc.), the load balancer redirects traffic to the XDM component in AWS.

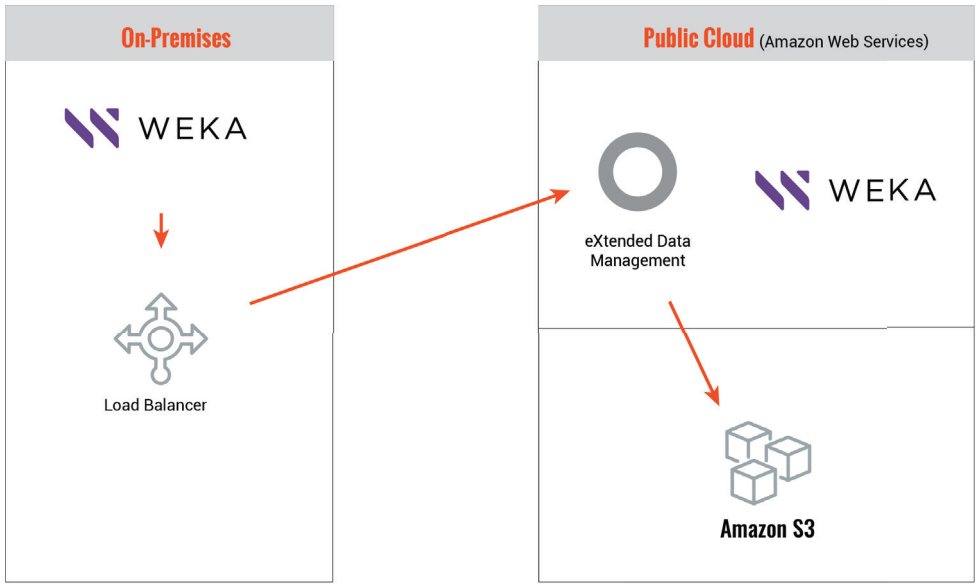


Figure 2: RING Issue

In the event of a complete datacenter loss, the EC2 VMs running Weka become the primary application and continue to write to the XDM component.

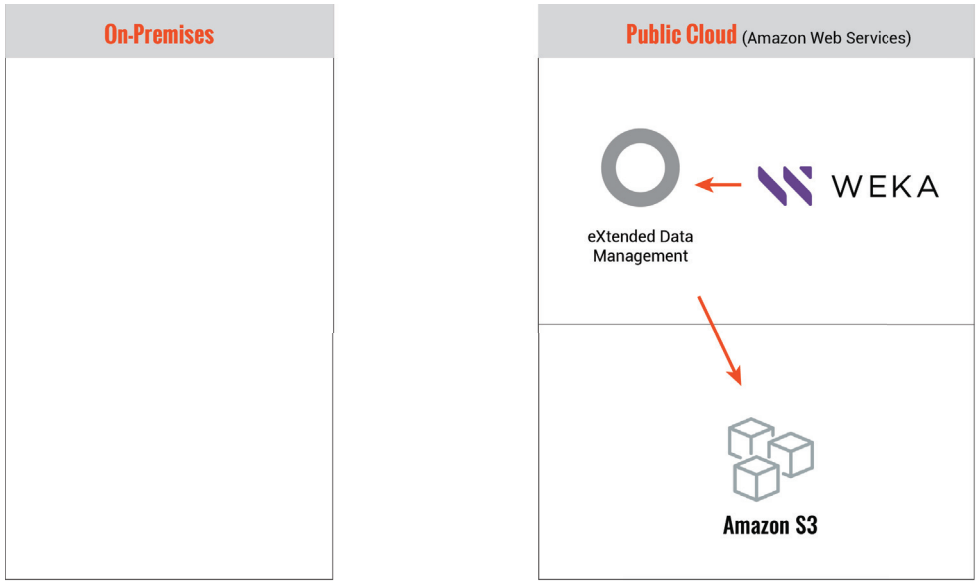


Figure 3: Datacenter Loss

RING8

In all of the above scenarios, XDM maintains a consistent metadata namespace, so when it is time to failback to normal operating conditions, XDM can write any data that was written directly to Amazon S3 during an on-premises outage back to RING.

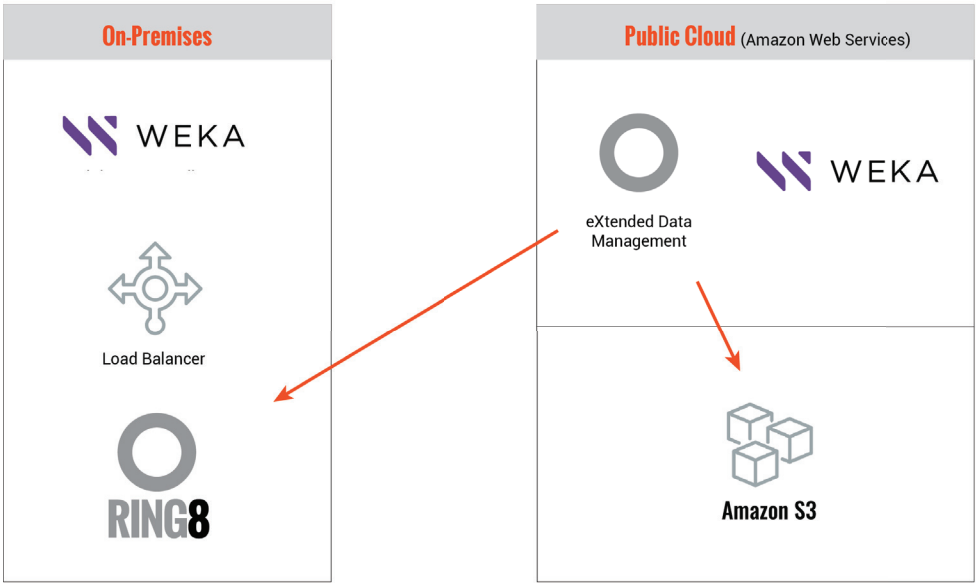


Figure 4: Failback / Return to Normal Conditions

An alternative architecture would be implementing the XDM component on-premises and in-band.

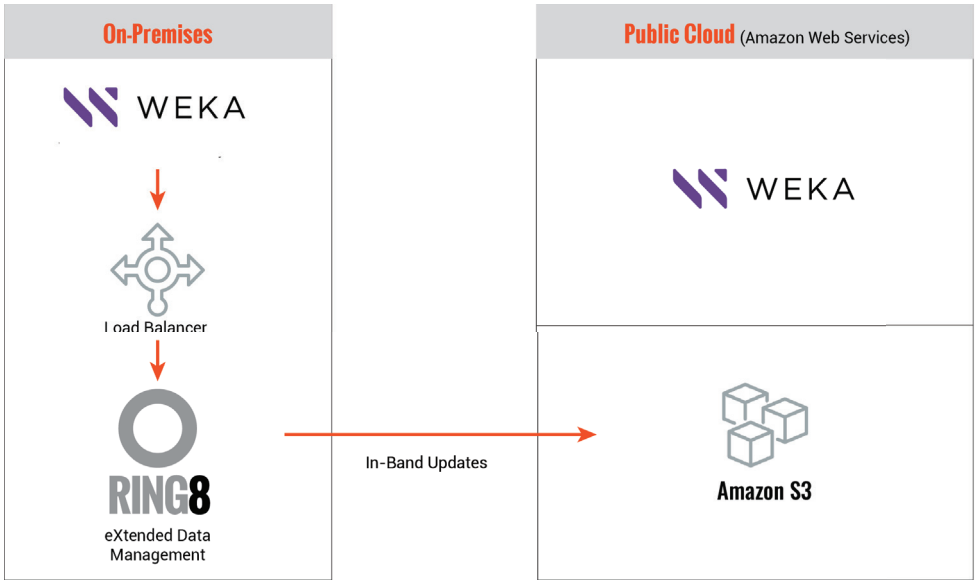


Figure 5: In-band Normal Operating Conditions

In the above in-band scenario, a secondary XDM system can be deployed in the cloud in the event of an on-premises failure. Metadata from the on-premises XDM can be backed up to the cloud such that this secondary XDM instance can be spun up along with the secondary WekaIO instance to enable a fully operational environment in the cloud.

Additional architectural details

In the case of this biopharmaceutical company's implementation, some additional architectural details include:

- Account and access keys are mirrored between on-premises and public cloud so that the WekaIO application and XDM can seamlessly read/write data either to RING during normal operations or Amazon S3 in the event of a failover.
- Failover is intentionally not automated. Every organization and every application is different. Some applications are business-critical and must have an immediate failover capability, while other applications can be down for some amount of time before the impact becomes problematic. The biopharmaceutical company's big data application can be unavailable for short time periods without serious disruption to the business. Not failing over to the cloud every single time there is a short outage or network hiccup saves the biopharmaceutical company unnecessary data egress charges when restoring data from the public cloud which can be one of the more expensive items of a hybrid cloud DR solution. In other words, if the connectivity between WekaIO and RING is interrupted, it is less expensive to wait until that issue is resolved than to automatically failover to the cloud, write a potentially significant amount of data, and then have to pull that data out of the cloud a short-time later.

Business benefits of Hybrid Cloud architecture

With this new hybrid cloud architecture, the biopharmaceutical company achieved the following business benefits:

- Able to complete research projects up to 12 times faster given the ability to more easily scale the storage system and support more concurrent research projects. As an example, a project that previously took four days (96 hours) can now be completed in eight hours resulting in dramatically faster time to market.
- Removed the on-going costs of maintaining and operating a second datacenter. AWS resources are now “rented” only when needed at a lower overall cost.
- Removed the traditional backup infrastructure overhead and costs.

RING8

Scality RING provides on-premises data protection across multiple servers and XDM provides a secondary copy in AWS.

- In addition to cloud DR, the replication and tiering features of XDM support the cloud archive and compute bursting use cases outlined above.

Hybrid Cloud DR Considerations

There are many factors that impact performance, recovery times, and costs to consider when architecting a hybrid cloud DR solution.

- **Recovery Objectives:** recovery point objectives (RPOs) and recovery time objectives (RTOs) are influenced by the sizing of the XDM component and the network bandwidth. More critical workloads with shorter RPO and RTO windows should leverage more resources for the XDM components and a higher bandwidth, perhaps dedicated network between on-premises and public cloud such as AWS Direct Connect.
- **Automated Failover:** as discussed above, the time to failover and level of automation could impact the costs of failback. If a very short outage results in an automated failover to the cloud during which significant data is written to the cloud, it is likely that restoring that data upon failback will result in significant cloud egress costs.
- **Cloud Storage Classes and Recovery Objectives:** cloud storage comes in many classes (tiers) ranging from "hot" to "cool" to "archive." The former classes typically offer better performance at a higher cost while the latter classes reduce performance and data retrieval times at a lower cost. For applications that require shorter recovery objectives, it is likely that a "hot" tier such as Amazon S3 would be best fit for cloud storage. For applications that can have longer recovery objectives, it is possible that an "archive" tier such as Amazon S3 Glacier would be an option as well as reduce the overall costs of the solution. However, once again, retrieval and egress costs should be taken into account when modeling the architecture.

RING8

Conclusion

For organizations that do not have access to or would prefer not to continue to invest in on-premises infrastructure for DR, hybrid cloud architectures present compelling options.The flexibility to replicate data to the public cloud, leverage different storage classes to manage costs based on application availability requirements, and quickly spin up application components in the cloud in the event of an on-premises failure can enable hybrid cloud DR scenarios that are more efficient and more cost effective than multiple on-premises datacenters.

Scality RING with Extended Data Management provides Scality's customers the freedom to select their cloud provider of choice as well as the flexibility to implement both in-band and out-of-band hybrid cloud DR architectures. The combination of industry-leading and globally proven RING for on-premises storage with the power and flexibility of the public cloud is helping to define the next generation of data protection.