2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks

Experiences with Building Disaster Recovery for Enterprise-Class Clouds

Long Wang[†], Harigovind V Ramasamy[†], Richard E Harper[†], Mahesh Viswanathan[‡], and Edmond Plattier*

[†]IBM T. J. Watson Research Center
Yorktown Heights, NY, USA
{wanglo,hvramasa,reharper}@us.ibm.com

[‡]IBM Cloud
Yorktown Heights, NY, USA
maheshv@us.ibm.com

*IBM Global Technology Services
La Gaude, France
edmond_plattier@fr.ibm.com

***Abstract*—The ability to recover from disasters is an important requirement for many enterprises. With enterprise-class workloads increasingly hosted on the cloud, cloud customers have come to expect disaster recovery (DR) as a necessary feature from cloud platforms. This paper identifies key challenges in providing DR as a service on enterprise cloud platforms, and portrays DR solutions for a managed cloud platform. In particular, we present the reference architecture for DR solutions, and describe our practical experiences in providing a portfolio of DR solutions for the cloud platform. The solutions cover diverse target recovery sites, such as an equivalent cloud site, a dedicated recovery site, and a customer-owned site. From the experiences, we provide insights into and lessons on implementing DR for enterprise-class clouds.**

Keywords-Cloud, Disaster Recovery, Automation

I. INTRODUCTION

Cloud platforms offer the convenience of rapid deployment, pay-as-you-go, and easy scalability. Cloud providers also offer a variety of sophisticated management capabilities as services, e.g., virus scanning, patching, monitoring, load balancing, security, and on-demand software. That has led enterprise-class clients (such as banks, financial institutions, hospitals, governments, and utility companies) to steadily migrate their large yet critical workloads to cloud environments.

Enterprise clients' resiliency requirements have always been very stringent, because the business impact of any service interruptions can be very large. Many cloud platforms are designed to be multi-tenant, hosting multiple clients. Consequently, failures on a given cloud site may have particularly dire consequences, affecting multiple or even all enterprise workloads spanning multiple clients.

Of particular concern are long-lived outages due to disasters, be they natural disasters such as earthquakes or floods, or man-made disasters caused by security attacks, terrorism, or operator errors. In many countries, including the United States, many businesses (such as financial institutions and health care providers) are required by law to have processes in place to recover from disasters within specified time periods. Disaster recovery on the cloud refers to the ability to recover the cloud infrastructure and the workloads hosted atop the infrastructure after a disaster. To ensure business continuity for enterprise-class cloud clients, the cloud infrastructure provider must ensure that systems, processes, policies, and procedures for recovering workloads are in place well before a disaster strikes. That requires careful and comprehensive planning and preparation before any occurrence of a disaster, and tightly-controlled execution of plans after the declaration of a disaster.

In this paper, we present our experiences with providing disaster recovery (DR) as a service on Cloud Managed Services (CMS), IBM's flagship *managed* cloud platform. CMS is distributed globally across multiple sites and hosts numerous enterprise clients. By *managed* cloud, we mean that the cloud manages IT services such as monitoring, patching, security, load balancing, and even certain applications (e.g., Oracle and SAP) on behalf of those CMS clients.

Given the diversity of client requirements and workloads as well as geographical constraints, a one-size-fits-all DR approach is impractical. Three DR architectures have been built at different CMS cloud sites: (i) cloud to cloud, (ii) cloud to non-cloud special-purpose site, and (iii) cloud to non-cloud client-managed site. In this paper, we present the first two DR architectures (due to page limitation, the third one is not discussed) and detail specific requirements that necessitated these distinct architectures, challenges in meeting the requirements, solutions devised, and lessons learned.

DR itself is an established field with many academic and commercial solutions [13][2][5][8][14][16][20]. However, DR for enterprise cloud platforms is relatively new, and has some unique challenges and additional constraints when compared to traditional DR. For example, in managed cloud environments like CMS, steady-state processes such as patching and monitoring are the cloud providers' responsibility, and these processes are expected to be managed by the cloud provider post-recovery at the secondary site (DR site). As another example, in traditional DR, examination and testing of the DR procedure may be performed once or few times annually. However, in a cloud environment DR test may be performed on-demand whenever the client desires. While commercial cloud DR solutions are available, many [14][16][20] focus on DR of native cloud applications only, i.e. application which are developed on top of resilient cloud services and APIs. Moreover, details about commercial DR solutions are typically limited to product brochures, user manuals, white papers, or web pages [14][17][21][22], which rarely delve into technical details of how the entire DR solutions are actually architected and implemented.

We make the following contributions in this paper. We provide comprehensive technical description of two DR architectures that have been implemented in a commercial managed cloud platform. Based on these real-world experiences, we propose a reference architecture for DR of enterprise-class clouds that can serve as a useful template for DR practitioners and researchers alike. We expose insights and lessons learned in addressing challenges unique to enterprise-class clouds.

978-1-4799-8629-3/15 $31.00 © 2015 IEEE
DOI 10.1109/DSN.2015.53

II. Challenges in Enterprise Cloud DR

CMS is a geographically distributed managed cloud platform hosting multiple enterprise clients running diversified enterprise workloads. Providing DR on such a platform raises a number of challenges.

Aggressive SLAs. Two metrics – Recovery Time Objective (RTO) and Recovery Point Objective (RPO) – are used to specify service level agreements (SLAs) for a Disaster Recovery service. RTO refers to the amount of time for operations to be restored after the occurrence of a disaster. RPO refers to the point in time up to which data restoration will be completed after a disaster; it is a measure of the amount of data lost during the recovery procedure. When Line-of-Business applications are migrated to or hosted on cloud, required RTO and RPO values are quite low (hours or minutes). Achieving such RTO/RPOs, especially at scale, absolutely requires automating critical DR functions.

Scale & Diversity. The aggregation of multiple large enterprise workloads into a single multi-tenant cloud site poses an overwhelming challenge for DR. For example, daily update of ten 10TB databases could easily consume over 1Gbps in replication bandwidth over Wide Area Network (WAN) distances, and CMS supports many such databases. Moreover, client requirements and system platforms are diverse in the enterprise cloud. Automation and flexibility are required for the DR solutions.

Inter-dependencies and Coordination of Server DR and Application DR. Besides restarting VMs, enterprise-cloud DR consists of recovering applications within the mandated RTO. This has usually been neglected in cloud DR solutions, relegated to manual processes, or left to the client's responsibility. Cloud-scale application recovery requires deep understanding of the inter-dependencies between servers and applications, and careful, automated orchestration.

DR of Management Capabilities. Because enterprise workloads are rigorously managed, it is necessary to continue to meet the clients' manageability, backup/restore, monitoring, security, patching, and auditing requirements while in residence at the DR site. A key technical challenge is to quickly reintegrate recovered workloads into the new management systems at the DR site within specified RTOs.

Location Requirements for the DR Site. Many enterprise clients do not allow their data to leave their respective countries. For them, the DR site cannot be located outside of the country where the primary cloud site is. Moreover, the primary cloud site and the DR site should be reasonably close to minimize the high-bandwidth replication network cost. Building a cloud site only to host recovered workloads is expensive. That may necessitate more flexible solutions that reuse data centers in the same country of the clients.

On-demand, Non-Intrusive DR Tests. The ability to recover from disasters must be tested regularly (even on-demand, for some clients) and audited, typically without disrupting the clients' DR readiness or production workload. To this end, automation of DR test is a must.

III. Overview of Enterprise Cloud DR

When a disaster takes out a technology infrastructure, a series of procedures need to be performed to recover workloads hosted on that infrastructure (primary site) at a secondary technology platform (secondary or DR site). In this paper, the primary site is always an enterprise-class cloud site (e.g., a CMS site), whereas the DR site may or may not be a cloud site.

Though there are vastly different requirements of DR from diverse enterprise clients and DR sites vary a lot in their characteristics, there are common concepts, use cases, and components in the different scenarios of DR for enterprise clouds. In this section, we discuss these common DR concepts, phases, components, and present a uniform reference architecture for various DR solutions.

A. Disaster Recovery

Figure 1 illustrates the typical DR life cycle consisting of a number of phases, which map to basic DR use cases.

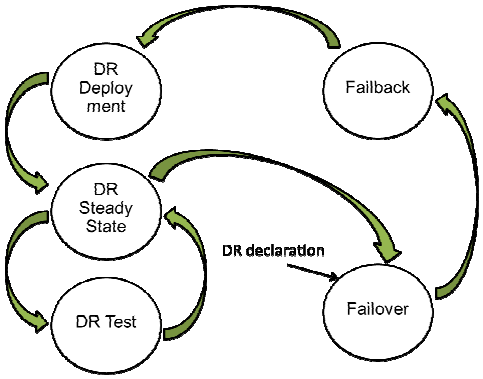

Figure 1: DR Life Cycle and Basic DR Use Cases

DR Deployment is the start phase of the DR life cycle. In this phase, the primary cloud site and the secondary site(s) are set up for supporting DR, and clients and their workloads are enrolled for DR protection. If a cloud system has distributed sites and built-in DR capabilities (which may be exposed through APIs), development of client workloads using such capabilities would also fall in this phase.

DR Steady State is the normal mode of operations, characterized by continuous data replication from the primary cloud site to the secondary site. DR-enabled machines may be modified or reconfigured in this phase, e.g. by adding new disks or installing new applications. Continual monitoring of replication and DR readiness is required in DR Steady State.

DR Test is the phase in which a client conducts tests of the recovery procedures by emulating a disaster occurrence and verifying that the procedures indeed recover workloads per SLAs. Cloud clients may demand DR testing annually, multiple times per year, or even on-demand. DR Testing is typically designed to be non-disruptive to production workloads and steady-state operations at the primary site, including data replication. Even if an actual disaster occurs during the DR test, recovery must happen correctly.

Failover refers to recovery of the primary cloud site at the DR site after a disaster has been *declared*. *DR declaration* is usually an executive-level decision, and may happen upon actual occurrence of a disaster, or even pre-emptively (e.g., a hurricane forecast). Upon DR declaration, failover procedures start to execute as a complete automatic process. This phase is marked by workload downtime and is bounded by RTO guarantees. After failover, clients' workloads run on the DR site.

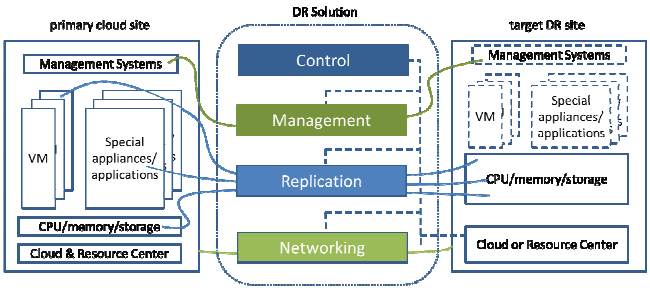

Figure 2: Reference Architecture for Cloud DR

Failback refers to the movement of clients' workloads from the DR site back to the primary site upon repair and restoration of the primary site to full functionality. After failback, the DR life cycle goes back to the DR deployment phase. If the DR site provides the same performance, management capabilities, administration and sufficient resource capacity, there is less incentive for failback, and both the cloud provider and clients may not want to do the failback but instead leave the production workloads running at the DR site. However, often DR site(s) are considered temporary place(s) for hosting client workloads, and consequently, may have performance, management, resource planning, and/or administration capabilities not comparable with those in the primary site. Repair and restoration of the primary site may not be possible in certain disaster situations. But the more common type of disasters (for example, those caused by power outages or operator errors) allow for the eventual restoration of the primary site. The client workloads can also be "failed back" to another cloud site if the original primary site cannot be repaired or the cloud provider/clients want to move the workloads to another cloud site.

B. Reference Architecture for Cloud DR

Though there are many DR solutions addressing a variety of requirements, conceptually their essential DR functionalities are similar. Figure 2 captures those functionalities in the form of a reference architecture. A DR solution consists of four types of components: replication, networking, management, and control.

1) Replication

Replication components perform the functions required for replicating workloads from the primary cloud site to the target DR site (or multiple DR sites). The primary site and the DR site are usually tens or hundreds of kilometers apart.

Different replication mechanisms have different RPO capabilities. For example, backup-based DR techniques have typical RPO values of 1 day while asynchronous electronic replication-based DR has RPO values of minutes or less.

Many replication modes exist, but the two key ones are active-active and active-passive. Active-active replication means that for each VM at the primary site, there is an active VM replica at the secondary site and the VM replicates to the active replica (the replica can be in the suspended mode). Active-passive replication means that for each VM at the primary cloud site, its state is replicated into the storage at the secondary site, but the VM replica is not provisioned until the disaster occurrence. Active-passive replication has lower costs than active-active replication as there is no need to host VM replicas at the secondary site in the DR steady state phase. The major benefit of active-active replication is the reduced RTO, since VM provisioning at the secondary site is done prior to and not during failover.

Replication can be performed at different levels: storage-level, virtual or physical host-level, and application-level. The three levels are indicated by the three curves through the *Replication* box in Figure 2. Sometimes, a combination of these replication levels may to be applied.

Storage-level replication works by copying any updates to VMs' state at the primary site's storage to the secondary site's storage. This type of replication relies on mirroring features of the underlying storage subsystem (available in certain enterprise-class storage systems, e.g., SAN Volume Controller [23]). An advantage of storage-level replication is that it can cover all VMs and all clients; no per-client or per-VM agent installation or configuration is needed. VM-level replication does not require special features of storage system, but requires installation of a replication agent in each VM. The per-VM agent installation may be cumbersome and different versions of the agent on different operating systems may need to be developed. Application-level replication is needed for certain services, applications, or middleware because generic storage-level or VM-level replication may not satisfy their special requirements. For example, if SAP HANA [26] appliances are replicated using a non-SAP-certified mechanism, SAP support and warranty become void. So we have to use HANA System Replication, the SAP-certified replication mechanism for HANA appliances.

2) Networking

Network design is critical for the DR of enterprise client's workloads. Several issues must be addressed in the design.

Replicated data traverses the following path from the primary cloud site to the secondary site: LAN in the primary cloud, WAN between the two sites, and finally, LAN in the secondary site. The physical network link for the WAN must have adequate bandwidth to handle the replication traffic in a large-scale enterprise cloud. This may necessitate acquiring or renting a dedicated link.

As multiple clients' workloads are replicated over the same data path simultaneously, the network design should support multiple replication streams and secure segregation of data streams (e.g., using VPNs, VLANs, and/or MPLS). Moreover, network design should also account for the possibility of multi-level replication (e.g., storage-level and application-level).

Network design should also provide secure access channels (through adequate authentication and authorization mechanisms) for administrators and clients. Administrators of the primary cloud site need to access certain resources in the secondary site for purposes of DR, management, reconfiguration, etc. Clients need to access their recovered machines in both DR test and failover phases.

Large enterprise clients have multiple organizations and complicated internal organizational structures, with each organization having its own cloud account(s). The cloud network design should allow for adequate segregation (e.g., using different VLANs) between accounts within the same enterprise client. The client's internal organizations and corresponding VLANs should be available in the DR site after failover. Depending on the RTO values, the VLANs can be pre-created before failover or recreated during failover.

Enterprise clouds provide clients with sophisticated management capabilities. Networks at the primary cloud site and the secondary site should be designed to support these capabilities. For example, load balancing is a key feature of enterprise clouds and may be implemented in the switching or routing facility of the cloud. Such switching/routing configurations for load balancing should be restored at the DR site after failover.

3) Management

Enterprise workloads are rigorously managed. There is an abundant set of management services for enterprise workload such as virus scanning, patching, high availability, directory services (e.g., Microsoft's Active Directory), load balancing, VLAN/firewall, monitoring, backup/restore, and compliance.

The DR site may be a cloud or just a platform with basic hypervisor capability, e.g. VMware vCenter, libvirt, etc. The management capabilities may or may not exist in the DR site before failover; so *Management Systems* in the DR site is illustrated in Figure 2 as dash-line box.

Usually, recovered workloads run on a DR site temporarily after failover and will be migrated to the original primary cloud site via failback soon after the original primary cloud site is repaired, or will be failed back to another cloud site (a newly created site or an existing site). So it may be acceptable to provide a limited set of management functions at the DR site after failover, if the full set is not available. Certain management functions are usually considered indispensable for the operation of enterprise workloads and must be provided at the DR site, e.g. virus scanning, security patch, Active Directory, backup, and basic monitoring.

4) Control

Control components orchestrate and coordinate the replication, networking, and management components discussed above, together with the computation resources at the DR site to implement all use cases of the DR life cycle. Control components fall into two main categories: (i) orchestration/automation components and (ii) administration components. Orchestration/automation components drive the automatic steps and coordinate manual steps for each use case in the DR life cycle. Automation is achieved by invoking the right DR components and the resource center to perform the right operations at the right time. Administration components usually take the form of self-service portal(s) for cloud administrators, clients, and the secondary site's administrators to launch a number of DR operations. Clients may use the portal for specifying which VMs should be replicated, initiating DR test or failover, viewing replication/recovery status, defining user roles, and specifying access permissions. Administrators may use the portal for enrolling clients into DR protection and other administrative tasks.

Orchestration/automation components play a critical part in the failover phase, where strict RTO requirements need to be met. Figure 3 show a high-level outline of an example failover procedure with three stages: environment recovery, workload recovery, and management recovery. During environment recovery, networking and management components are recovered to be functional enough to support workload recovery. Certain management components (e.g., authentication and authorization components like Active Directory) need to be recovered prior to recovering workloads. During workload recovery, the resource center at the DR site provisions VMs in order of priority (in this example, we consider active-passive replication) and loads them with replicated data. That is followed by application recovery on the started VMs. After validation of the recovered VMs and workloads, management tools are recovered. During management recovery, the rest of the management capabilities are recovered and applied to the recovered workloads.

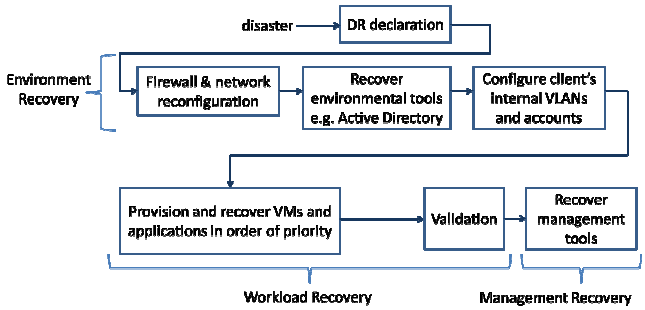

Figure 3: Outline of a Sample Failover Procedure

IV. Building DR for an Enterprise Cloud

The reference architecture captures the most important DR components and functions. Various embodiments of the reference architecture must adapt to different practical constraints, requirements and/or topologies.

Table 1: DR Solutions for an Enterprise Cloud Platform

	DR1	DR2	DR3
Description	Failover to similar cloud site	Failover to dedicated DR site	Failover to custom site
RPO/RTO Specifications	15min/4 hours	15min/4 hours	15min/4 hours
Regional Availability	Another cloud site in same region	No cloud site but a purpose-built DR site in same region	No cloud site but a customer-owned site in same region
VM Provisioning at DR site	cloud's VM provisioning	Dedicated DR site's VM recovery mechanism	VMWare vCenter
Replication	Storage-based and application-based; Active-active	Host-based and application-based; Active-passive and Active-active	Storage-based; Active-passive
Post-failover Management	Full management	Limited management	Limited management
Networking	VLANs over dedicated Fiber link	VPN, MPLS or Point-to-Point with Layer 2 or Layer 3 routing	VLANs over dedicated link
Control	Custom DR Orchestrator	Dedicated DR site's DR automation	Custom DR Orchestrator

Table 1 summarizes three DR solutions we have implemented for the CMS platform, denoted as DR1, DR2, and DR3. Location requirements had to be satisfied for each solution, i.e. the primary cloud site and the DR site must be in the same region or country. DR1 is for a cloud-to-cloud scenario. In DR2, the secondary site is a purpose-built DR site. For DR3, neither a cloud site nor a special DR site was available in the region; therefore, a custom site with client-provided resources was re-purposed as a DR site. The RPO/RTO values listed in Table 1 are specifications based on

Service Level Agreements (SLAs) or Service Level Objectives (SLOs) specified in contracts. In DR1 and DR3, the actual RPO is near-zero; the actual RTO is less than 2 hours. In DR2, the actual RPO and RTO are less than 1 min and 4 hours, respectively.

We provide details of the DR1 and DR2 solutions below. Due to page limitations, the details of DR3 are not given in this paper. Basically, DR3 is a light-weight version of DR1 with fewer management capabilities and certain custom features (e.g., same IP address for recovered workloads as the primary ones) implemented at a DR site that was built by re-purposing the customer's data center.

A. *DR1: Cloud-to-Cloud Solution*

When there is a cloud site in the same region of the primary cloud site with equivalent management capabilities and networking architecture, the DR1 solution is the preferred method for protecting enterprise workloads, as it greatly alleviates the problems of recreating management capabilities as well as enterprise clients' VLANs and internal organization structures at the DR site. DR1 solution is applied to most sites of the CMS cloud platform.

Overview. Figure 4 illustrates the architecture of the DR1 solution. Two cloud sites with identical management services function as a DR pair in an active-active configuration, with data protection enabled via storage-level replication. The identical storage configuration of the two cloud sites allows for use of storage-level replication. The relationship between the two sites are pairwise symmetric, i.e., each cloud site in a DR pair is a DR site for the other. Before such a DR pairing can be established, the two sites are linked by a duplex replication network (typically, at least 2x10Gbps).

Storage devices at both sites are paired with each other via Global Mirroring [24], which is based on the *SAN Volume Controller* [23] component in Figure 4. Global Mirroring is an asynchronous data replication technology for providing aggressive RPOs with little performance impact on applications at the primary site. Host-level replication (such as Oracle Data Guard [25] and rsync) is also supported.

A typical CMS client may have around 40 VMs with an average disk size of 200GB for each VM. The average storage utilization is around 80%, i.e. 80% of disk capacity is occupied with data; so the initial replication traffic for such a client is about 40 VMs x 200 GB/VM x 0.8 = 6400 GB. The maximum speed of initial replication is estimated as *Available Bandwidth* x *Transmission Efficiency (80% in our case)*, i.e., 10Gbps x 0.8 = 8 Gbps. Full transmission efficiency of 10Gbps cannot be achieved in practice due to protocol overhead and other reasons. So, it takes about 6400GB x 8/8Gbps = 6400s = 1.78 hours to finish the initial replication assuming no throttling of initial replication. After that, our observation shows that the average daily rate of data change is around 10%-15%. If we use the higher number in that range, there is about 6400 GB x 0.15=960GB of data change every day, and the occupied bandwidth is around 960GB x 8 / (24x3600s)=0.089Gbps, which is 1.1% of the 8Gbps effective bandwidth. While many clients have tens of VMs, there are also big clients who have thousands of VMs. For a client with 1200 VMs, it takes more than 2 days for initial replication, which will occupy 33.3% of available bandwidth in steady state. Practical experience shows that for proper replication performance, average steady state bandwidth utilization must be well under 70%. Networking devices and software at both sites are pre-configured to enable data replication and to enforce segregation of data streams for different clients. These devices include firewalls and routers, and are indicated by the *Network Termination* component in Figure 4.

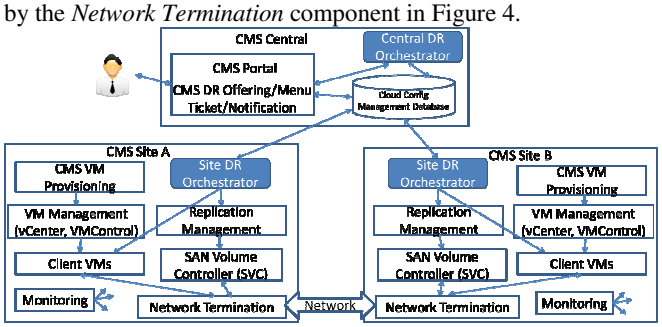

Figure 4: Architecture of the cloud-to-cloud DR solution

As active-active replication is used in DR1, DR tests show low RTO values, i.e. a few tens of VMs were recovered in a few minutes and 1200 VMs were recovered within 2 hours.

DR Life Cycle. Consider a cloud customer whose production workloads are hosted on one of the cloud sites in the DR site pair. The customer can sign up for DR-as-a-Service at the central *CMS Portal* (Figure 4) through a DR entitlement process, during which the customer's virtual environment (including VLANs and security zones) is pre-deployed on the other cloud site via the *CMS VM Provisioning* and *Network Termination* components. Some applications may require application-level replication, in which case the active instances at the two sites need to communicate with each other.

Post entitlement, a customer may request DR enablement for any of his/her VMs. Upon this request, a VM is created by the *CMS VM Provisioning* component at the DR site with identical memory, CPU, number of data disks, and sizes of disks. The VM has an IP address different from the primary VM. All managed services that were enabled on the primary VM are automatically activated on the newly created VM. Data replication is then enabled for the VM's disks via Global Mirroring. Consistency among multiple VMs' states at the DR site is achieved by using the Consistency Group (CG) feature of the Global Mirroring technology, which ensures write-order consistency among storage of all hosts in a CG. In the steady-state phase, the CGs are set to read-only mode at the DR site because the states of the secondary VMs are mirrored from the primary site.

The failover operation begins upon DR declaration (see Figure 1), and is executed in an automated fashion by the Site DR Orchestrator (Figure 4), or DRO. Each cloud site has a DRO. Failover is similar to the process shown in Figure 3, but with some differences. Environment recovery is already completed before DR declaration during DR entitlement as described above, and management recovery is not needed as management functions are already provided by the DR cloud site (certain minor reconfiguration may be needed). So failover mainly consists of workload recovery. The DRO at the DR site retrieves the list of VMs to be recovered from the *Cloud config management database* at *CMS Central* site, and prioritizes their recovery according to client specifications. If

the VMs contain Global Mirrored data, the DRO invokes a script to (i) stop Global Mirroring replication for the VMs' CGs, and (ii) change access of the CGs from read-only to read/write. Finally, the DRO administers a *shoulder tap*, i.e., a notification to the application layer of the active VMs at the DR site that application-level recovery can begin. The DRO does so by logging into the VMs and invoking shoulder-tap scripts (installed by the client at a predefined location).

Since we have an active-active configuration, failback is equivalent to performing the failover actions at the other site using the automation provided by the DRO.

B. *DR2: Cloud-to-Dedicated DR Site*

IBM has a tradition of providing disaster recovery services for enterprise customers long before the cloud era, and dedicated DR sites have been set up all over the world for this purpose. A single DR site may service tens to hundreds of enterprise customers whose workloads ran on traditional non-cloud platforms. For a CMS site in a region where there is a dedicated DR site but no other CMS site, we apply this DR topology to satisfy clients' location requirement.

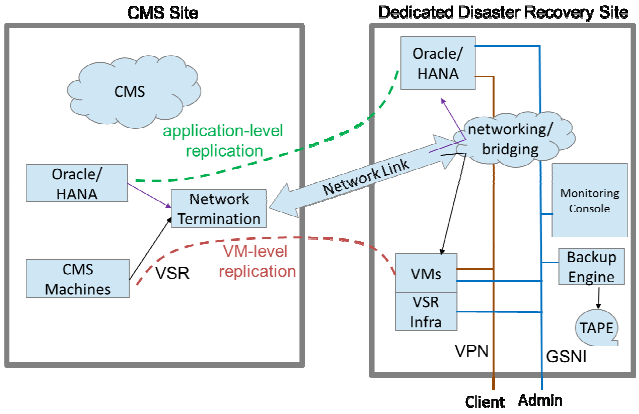

Figure 5: Architecture of DR to Dedicated Site (DR2)

Overview. DR2 solution (Figure 5) is based mainly (but not exclusively) on active-passive replication between the CMS cloud site and the DR site. DR2 uses active-passive because DR2 is a lower-priced offering than DR1, and active-passive typically has lower cost, since there is no need for resources to support active VM instances during normal operation. Active-passive replication is performed using Virtual Server Recovery technology [18] (*VSR Infra* component in Figure 5). VSR is used for automation of the DR process (similar to the DRO in DR1), including the dynamic provisioning and restart of machines at the DR site upon DR declaration. VSR replication is at VM-level (or host-level for physical machines) and supports consistency groups. VM-level replication is used because the dedicated DR site was built to support both non-cloud and cloud-hosted workloads. Application-level replication is also supported in DR2 for certain middleware in active-active mode. In particular, when Oracle database and SAP HANA workloads are enabled for DR, active instances of the machines hosting these middleware are pre-created in the DR site. Vendor-specific application-level replication [25][26] is performed between the primary and secondary instances.

Various types of network links with different bandwidths can be used between a CMS cloud site and the dedicated DR site. Their bandwidths range from 100Mbps (IPSEC VPN) to 2.5Gbps (Optical Carrier 48). Selection of network type depends on the site scale and required replication bandwidth.

DR Life Cycle. A CMS client contacts the manager of the DR site to enroll into the DR Service. Then the cloud administrators install VSR agents in the client's machines to be protected. The agents replicate data of the machines to VSR infrastructure at the DR site. With active-passive replication, the replicas at the DR site can have the same host names and IP addresses as the primary ones.

If application-level replication is required, administrators of the DR site create machines in the DR site and set up the active-active replication. In addition, the administrators of the DR site configure the networking properly to reflect the client's network settings in the CMS cloud site (see Section III.B on what network settings should be configured).

Figure 3 actually depicts the failover procedure for DR2. DR tests for real clients show that it typically takes less than 10 minutes for workload recovery on tens of VMs. Even with environment and management recovery, DR test or failover for hundreds of VMs is well within the 4-hour RTO.

When the primary cloud infrastructure is repaired, the client can request its workloads be failed back to the cloud site. The VSR engine provides host-level replication for failback. First, the client's machines are provisioned at the primary cloud if they are not already there. Then these machines are booted by a VSR-provided ISO image via a remote SCSI mount. This ISO image directs the booted machine to replicate data from VSR storage of the DR site.

Management Recovery. Recovering management systems is a major task in DR2, whereas in DR1 we can leverage cloud management capabilities already present in the DR site. Figure 5 shows a subset of management capabilities at the DR site (monitoring, backup, and networking). *GSNI* refers to a virtual network created for cloud administrators to access and configure resources of the DR site.

Redundant instances of certain management tools (e.g., backup servers and HANA Studio [26]) are provisioned at the DR site in the steady-state phase, and will be ready for use at failover time. A number of other infrastructure, security, management and access functions for the CMS cloud site need to be restored at the DR site after failover. These functions include backup, DNS for resolving IP addresses, Active Directory (AD) for authenticating/allowing users' access into systems, Lightweight Directory Access Protocol (LDAP) for allowing delivery staff access to managed resources, Delivery Access Server for remote management, VLANs for segregating data streams of clients and management systems, security rules implemented on firewalls and routers, load-balancing implemented on switches, and monitoring capabilities. A separate virtual network is created for cloud administrators to access resources of the DR site. DNS, AD, LDAP, and Delivery Access servers, as well as load balancers are hosted on VMs that are replicated and recovered using VSR, just like client workload VMs. Similarly, SAP Manager [26] machines are replicated and recovered just like regular VMs.

V. Lessons Learned from Our DR Experiences

Enterprise cloud DR should cover not only workloads but also the systems management services and the manage-from environment. Cloud infrastructures typically consist of a manage-to environment (which hosts the customer workloads) and a manage-from environment (which controls the rest of the cloud and manages the manage-to environment). Upon DR declaration, systems and services in the manage-from environment at the DR site have to be functional before the start of the recovery of the managed-to systems. Enterprise workloads require rich systems management services (e.g., monitoring, virus scanning, patch, and backup) throughout the DR life cycle, including the period during which workloads may be in residence at the DR site upon failover. Systems management software (e.g., agents in the customer workload VMs) that underlie these services may have critical dependencies on the client workload and environment (e.g., resource UUIDs, VLAN IDs, network identities, and configurations in the primary site). State transitions during the DR life cycle may have subtle effects on these dependencies, e.g., IP address of a patch management server may be different in the primary site and DR site. Such changes need to be carefully tracked and accounted for, to ensure continued and uninterrupted manageability of the DR-protected VMs. For this purpose, the systems management software may need to be recovered, adapted, (re-)configured, and/or (re-)activated one or more times (e.g., during failover and failback) during the DR life cycle. Failure to do so may render the recovered VMs without systems management services, a situation unacceptable for many enterprise-class workloads.

Standardization vs. customizability. Standardization is the one of the pillars of cloud computing. However, enterprise clients, particularly larger ones, expect a good degree of customizability to fit their existing policies and environments. To allow for broader adoption of their platforms, cloud providers need to be sensitive to those expectations. For example, clients have different preferences on whether to retain or change the network IDs (IP addresses and hostnames) of their VMs after failover. Enterprise DR designers may have to accommodate both cases.

Data management is central to DR design. At the core of any DR solution is the data that needs to be protected and recovered. Data replication and management drives much of the design effort and cost. It is important to estimate the amount of data that needs to be moved per unit time between the sites as accurately as possible, as this estimate is critical to many design factors, such as network bandwidth and network type. The estimate should also account for projected growth in data movement as more VMs obtain DR-protection.

If the goal of the recovery effort upon failover is to recover applications, then data replication is most efficiently done at the application level. Application-based data replication is superior in maintaining state and control than infrastructure-based replication. The latter is more resource-hungry and resource-expensive, and maintains no application state. However, not all applications have in-built replication capability. For those applications, replication has to be done at the infrastructure level and additional mechanisms (e.g., log replay) may be needed to correctly recover application state.

Regulations and regional requirements may trump technology in DR. Depending on the country and geographical region, enterprises may be subject to strict data protection regulations that may thwart lower cost DR solutions. For example, some countries have strict laws that prohibit movement of certain types of data outside national borders. Cloud DR designers need to keep such requirements in mind and provide flexible choices for DR across country borders and within country borders.

The location of the DR site(s) should also satisfy geographic diversity requirements. The distance between the primary and DR sites should be far enough to account for risk factors such as the prevalence of natural disasters and the potential radius of their impacts. What constitutes acceptable distance between the primary and DR sites may vary from continent to continent. For example, the acceptable distance may be measured in tens of kilometers for Europe and in hundreds of kilometers for North America.

Reducing DR cost. DR protection is expensive. A network capable of handing the replication traffic between the primary and DR sites in a timely fashion supportive of RPO requirements may be costly to provision. The cost of protecting a VM in active-active mode is potentially more than twice the cost of the VM. That is due to the cost of the second VM plus associated storage, replication, orchestration, and management costs. Cost reduction may be obtained by utilizing the reserved resources in the DR site for non-critical purposes (e.g., running dev/test workloads) until DR declaration. However, such a setup may significantly complicate DR orchestration. An active-passive configuration may further help reduce the cost. However, it may not be suitable for workloads with very stringent RTOs (e.g., minutes), due to the risk of not being able to provision VMs in a timely fashion upon DR declaration.

Automation is a must to achieve low RTOs. Failover and failback processes place high resource demands on the DR and primary environments respectively, because of the need to resume or restart 100s, if not 1000s, of VMs closely together. Those processes may rack up heavy loads on storage devices, network services, and hypervisors in the environment, which may lead to additional failures. To meet the low RTOs demanded by enterprise workloads, careful, complex, and automated orchestration is required. Manual processes have little or no place in enterprise-cloud DR.

DR Testing should be flexible and non-disruptive. Enterprise clients expect the ability to test DR readiness for their cloud-hosted workloads. While annual DR tests are common, clients increasingly want the flexibility to perform DR tests on-demand at times and frequencies of their choosing. Cloud DR architects need to take these requirements into consideration, while simultaneously ensuring that the DR tests are non-disruptive to the ongoing protection of workloads actively running in the primary site.

VI. Related Work

Data replication and data backup are key to DR and have been covered extensively in the literature (e.g., [1][2][6]). Han et al. [1] describe a scheme for basic data replication, failure detection and switchover. SecondSite [6] describes a LAN-based replication mechanism that is extended to a WAN. PipeCloud [2] provides synchronous replication

consistency across WAN for DR. Besides replication, other problems relevant to DR have also been studied in the literature. Keeton et al. [3] discuss how to schedule recovery order of different data during DR. Bianco et al. [4] study optimization of VM placement and storage assignment during recovery for ensuring high network performance of recovered workloads. Zhang et al. [5] propose a design of Storage Area Network for storage-level cloud DR. The approaches described in the above works address specific problems in enabling DR; however, they do not describe a complete DR solution for enterprise clouds.

Prior work on cloud DR includes survey of cloud DR [13], creation of optimum DR plans [7][8], discussion of DR requirements [11][12], and modeling and analysis of DR approaches or DR-protected clouds [9][10]. These studies focus on theoretic analysis, optimization, or high-level discussion of cloud DR. Robbins [19] reports on manually migrating workloads from one Amazon datacenter to another as part of a DR exercise. Our work differs from prior work in that we describe practical challenges and real-world experiences in designing and implementing automated DR services for enterprise cloud customers from a cloud provider's perspective.

A number of cloud providers and third-party corporations deliver cloud DR services to clients, such as Amazon [14], VMWare [16], Rackspace [20], Microsoft [14], EVault [17], Rightscale [21] and CSC [22]. There are product brochures, user manuals, or web pages that give high-level descriptions of their DR services or in-depth presentation on a few individual DR techniques. However, they do not provide internal details or comprehensive technical description of their entire DR architectures. We have provided such details in this paper, with the aim of addressing challenges unique to enterprise-class clouds covering both born-on-the-cloud applications and traditional enterprise applications that were not built on top of cloud services or APIs.

VII. CONCLUSIONS

DR is required for many critical business applications of enterprise clients. Providing DR for enterprise clouds raises some unique challenges not encountered in traditional DR. In this paper, we presented a reference architecture for enterprise cloud DR, and described real-world experiences in providing a diverse array of cloud DR solutions. The insights and lessons drawn out of our DR experiences can be useful for researchers and practitioners alike.

REFERENCES

[1] H. Han, L. Li, D. Zhu, "Research and implementation on remote disaster recovery system," International Conference on Computer Science & Service System (CSSS), 2012.

[2] T. Wood, H. Andrés Lagar-Cavilla, K. K. Ramakrishnan, P. Shenoy, and J. Van der Merwe. "PipeCloud: using causality to overcome speed-of-light delays in cloud-based disaster recovery," Proceedings of the 2nd ACM Symposium on Cloud Computing 2011 (SOCC '11).

[3] K. Keeton, D. Beyer, E. Brau, A. Merchant, C. Santos, and A. Zhang. "On the road to recovery: restoring data after disasters," Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems 2006 (EuroSys '06).

[4] A. Bianco, J. Finochietto, L. Giraudo, M. Modesti, F. Neri, "Network planning for disaster recovery," 16th IEEE Workshop on Local and Metropolitan Area Networks, 2008. LANMAN 2008.

[5] J. Zhang, N. Zhang, "Cloud Computing-based Data Storage and Disaster Recovery," International Conference on Future Computer Science and Education (ICFCSE), 2011.

[6] S. Rajagopalan, B. Cully, R. O'Connor, and A. Warfield. "SecondSite: disaster tolerance as a service," SIGPLAN Note. 47, 7, 2012.

[7] P. Fallara, "Disaster recovery planning," Potentials, IEEE, vol. 22, (5) pp. 42-44, Dec. 2003-Jan. 2004.

[8] O.H. Alhazmi, Y.K. Malaiya, "Evaluating disaster recovery plans using the cloud," Proc. Reliability and Maintainability Symposium (RAMS), 2013.

[9] M. Pokharel, S. Lee, J. S. Park, "Disaster Recovery for System Architecture Using Cloud Computing," 10th IEEE/IPSJ International Symposium on Applications and the Internet (SAINT), 2010.

[10] B. Silva, P. Maciel, E. Tavares, A. Zimmermann, "Dependability models for designing disaster tolerant cloud computing systems," 43rd IEEE/IFIP Intl. Conf. on Dependable Systems and Networks (DSN), 2013 .

[11] L. L. Hoong, G. Marthandan, "Factors influencing the success of the disaster recovery planning process: A conceptual paper," Intl. Conf. on Research and Innovation in Information Systems (ICRIIS), 2011.

[12] L. Turnbull, H. Ochieng, C. Kadlec, and J. Shropshire. "Improving service continuity: IT disaster prevention and mitigation for data centers," In Proceedings of the 2nd annual conference on Research in information technology (RIIT '13), 2013.

[13] M. Khoshkholghi, et al. "Disaster recovery in cloud computing: A survey," Computer & Information Science vol. 7, (4) pp. 39-54, 2014.

[14] Microsoft Azure, "Site Recovery – Cloud Disaster Recovery," [online] 2014, http://azure.microsoft.com/en-us/services/site-recovery/ (Accessed: April 3, 2015).

[15] Amazon Web Servics, "Using Amazon Web Services for Disaster Recovery," [online] 2014, https://media.amazonwebservices.com/AWS_Disaster_Recovery.pdf (Accessed: April 3, 2015).

[16] VMware, "VMware vCloud Air Disaster Recovery," [online] 2015, http://www.vmware.com/files/pdf/vchs/VMware-vCloud-Hybrid-Service-Disaster-Recovery-DS.pdf (Accessed: April 3, 2015).

[17] Evault, "Cloud Disaster Recovery Services," [online] 2015, http://www.evault.com/products/disaster-recovery/ (Accessed: April 3, 2015).

[18] IBM Global Technology Services, "Virtualizing disaster recovery using cloud computing," [online] 2015, https://www-935.ibm.com/services/uk/en/it-services/vsr_whitepaper.pdf (Accessed: April 3, 2015).

[19] Bill Robbins, "Disaster recovery in a cloud environment," [online] 2011, http://www.ibm.com/developerworks/cloud/library/cl-disasterrecovery/cl-disasterrecovery-pdf.pdf (Accessed: April 3, 2015).

[20] Rackspace, "IT High Availability Disaster Recovery Plans with Rackspace," [online] 2015, http://www.rackspace.com/disaster-recovery-planning (Accessed: April 3, 2015).

[21] RightScale, "Disaster Recovery," [online] 2015, http://www.rightscale.com/solutions/cloud-computing-uses/disaster-recovery (Accessed: April 3, 2015).

[22] CSC, "Business Continuity Management and Disaster Recovery," [online] 2015, http://www.csc.com/cybersecurity/offerings/45397/86285-business_continuity_disaster_recovery (Accessed: April 3, 2015).

[23] IBM, "Spectrum Virtualize - SAN Volume Controller," [online] 2015, http://www-03.ibm.com/systems/storage/software/virtualization/svc/ (Accessed: April 3, 2015).

[24] IBM, "Global Mirror Technical Whitepaper," [online] 2008, http://www-01.ibm.com/support/docview.wss?uid=tss1wp100642 (Accessed: April 3, 2015).

[25] Oracle, "Data Guard Concepts and Administration," [online] 2008, http://docs.oracle.com/cd/B19306_01/server.102/b14239/toc.htm (Accessed: April 3, 2015).

[26] SAP, "SAP HANA," [online] 2014, http://hana.sap.com/ (Accessed: April 3, 2015).