WHITE PAPER–**JUNE 2023**

Continuous Recovery

Protecting workloads in a PVDC spanning between two geo data centers

Table of contents

Executive Summary

Many cloud providers participating in VCPP have successfully adopted VMware Cloud Director Availability (VCDA). The product improvements in the latest releases related to the simplified implementation and operation processes, useful features, and deep integration with vSphere and VMware Cloud Director (VCD), in addition to the attractive pricing model, made VCDA the preferred product for providing DR and migration services for their tenants' workloads.

However, there are still cases when the distributed nature of data used by replication engine creates challenges. This makes the answer to the question "How to deploy and configure VCDA appliances in a way to ensure DR service agility?" not so obvious and the goal of this document is to explore one such case.

Introduction

VCDA relies on two replication engines – Classic and VMC. The Classic data engine utilizes the Host Based Replication (HBR) technology. HBR enables replication of Virtual Machines without any modification at the ESXi kernel level or inside Guest VM operating system. While HBR operates at layer of the hypervisor VCDA leverages HBR capabilities for providing DR and migration services at the cloud level. This document is applicable for cases when VCDA is configured to use the Classic data engine and discusses how to design a resilient DR service.

The design of any other technologies mentioned in this document is out of scope and will not be explored in detail.

VCDA and HBR maintain a set of data required to operate the replications between source and destination. Part of this data is centralized in VCDA manager in the recovery site, another part is distributed inside VCDA replicators, and third part is located outside of VCDA appliances on the datastores. At this moment when the latest VCDA release is 4.5, data maintained by a single replicator in its own database cannot be reconstructed if the appliance fails for any reason, for example a failure of the datastore where the appliance resides.

Use Case Requirements

SOLUTION BUSINESS REQUIREMENTS	TYPE	DETAILS	DESIGN QUALITY
SBR001	Disaster Recovery	# Simplify the recovery process, where recovery of the workloads with no configuration changes at the application level and most importantly its security posture.	• Manageability • Availability • Recoverability
SBR002	Disaster Recovery	# Achieve the lowest possible RTO (Recovery Time Objective) while conducting full or granular workloads failover.	• Manageability • Recoverability • Performance
SBR003	Disaster Recovery	# Tenants would be able to control the failover and failback process and own it end-to-end with the least intervention from the service provider.	• Manageability • Security

Document Dictionary (Terms and Acronyms)

This section will provide the design specific terms and acronyms used across the document.

TERM	DESCRIPTION
Regions (R)	- Regions are the geographic locations of your data centers. - Different regions might offer different service qualities in terms of latency, solutions portfolios, and services.

Availability Zone (AZ)	- Is an isolated data center within a co-located data centers within a single region, where no single data center is to be shared between multiple availability zones as to preserve the level of segregation, availability, and dependency.
Continuous Recovery	- Is a regional function which enables tenants to perform recovery operations (failover/failback) without any change at the application level from multiple vectors (compute, storage, network, and security).
RxAZx	- Stretched components across AZxA and AZxB.
RxAZxA	- Availability Zone X, Data Center A.
RxAZxB	- Availability Zone X, Data Center B.
RxAZxC	- Availability Zone X, Data Center C (usually witness).
RxAZx	- Availability Zone X.

High Level Logical Design and Assumptions

Technologies of Choices

Together, the combination of the hereunder VMware technologies enable service providers to achieve the use case requirements.

CLOUD COMPONENT	DESCRIPTION	CLOUD CAPABILITY	SERVES WHAT?
VMware Cloud Director ®	Layer of software that abstracts virtual resources and exposes cloud artifacts to consumers.	● Cloud Management Portal / Cloud Broker. ● Cloud Consumption Portal / Cloud Broker.	● Resources
VMware Cloud Director Availability ®	Tenant workloads replication technology providing Cloud to Cloud and DR to Cloud capabilities.	● Cloud Management Portal Integration. ● Cloud Consumption Portal / Cloud Broker Integration. ● On-Premises DR to Cloud Capability. ● Cloud to Cloud DR Capability. ● Migration to Cloud Capability.	● Resources
VMware vSphere ®	Virtualization platform providing abstraction of physical infrastructure layer for Cloud Director.	● Compute Resources Service Layer. ● Cloud Management Portal Integration.	● Management ● Resources
VMware vSAN ®	For Software defined storage integration and capabilities.	● Storage Resources Service Layer.	● Management ● Resources
VMware NSX-T ®	For Software defined networking integration and capabilities.	● Software Defined Networking Service Layer. ● Software Defined Security Service Layer. ● Cloud Management Portal Integration.	● Resources

		● Cloud Resources Portal Integration.	

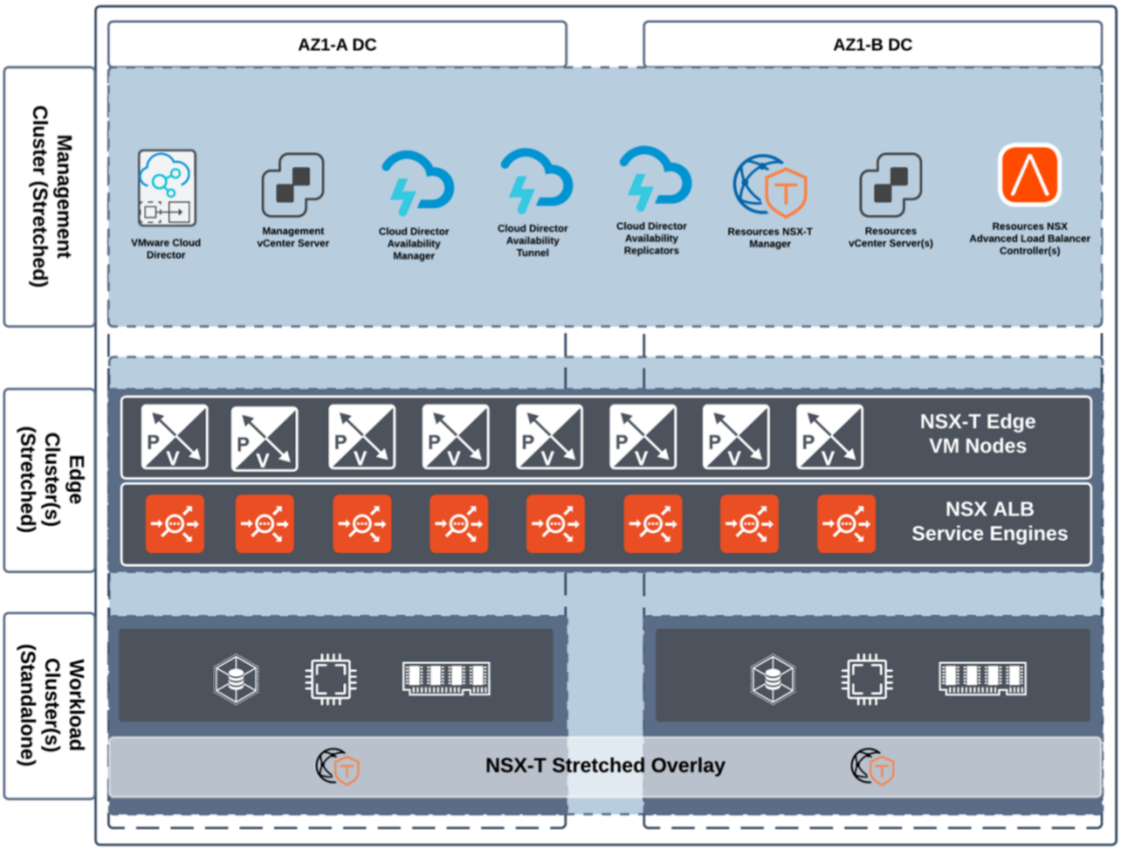
AZ1-A DC
AZ1-B DC
Management Cluster (Stretched)
Edge Cluster(s) (Stretched)
Workload Cluster(s) (Standalone)
VMware Cloud Director
Management vCenter Server
Cloud Director Availability Manager
Cloud Director Availability Tunnel
Cloud Director Availability Replicators
Resources NSX-T Manager
Resources vCenter Server(s)
Resources NSX Advanced Load Balancer Controller(s)
NSX-T Edge VM Nodes
NSX ALB Service Engines
NSX-T Stretched Overlay

Figure 1 – Infrastructure High Level Diagram

The Infrastructure

Based on the requirements the high-level architecture must provide continuous recovery; and to achieve this, we need first to look at the management components, and this is because all the cloud management components must be up and running at any point in time. As such, the management components will have a significant impact on how the recovery process will look like, and those management components include the cloud management platform and its integrations, along with the cloud resources components.

Continuing with management components, like our standard architecture for primary functions separation, there would be two vCenter Servers, the first one will be dedicated for the management components and will incorporate the management cluster which will host all the cloud management components. The vCenter Server will incorporate the cloud resources and will host two types of clusters, the first is the tenant's compute workloads and the second cluster will host the tenant's edge networking workloads (mainly the NSX-T edge VMs and AVI Service Engines.

The key element here, is cluster classification and availability where:

- Management must be a stretched cluster.
- Edge must be a stretched cluster.
- Compute must be standalone clusters (as the intention is to replicate workloads for disaster recovery purposes).

Management and Edge would be two distinct clusters that will utilize hosts from each availability zone in a single region and will be configured to sustain a complete outage of an availability zone by using a stretched vSAN cluster capability. Management and edge clusters must be sized in a way that in case when an entire availability zone is down all VMs will be able to run from the other zone without performance impact (remember that the admission control policy here for vSphere must be 50% to sustain either AZ1A or AZ1B failure).

Compute clusters will have different configuration, where each resource cluster will utilize hosts from only one availability zone, and the storage solution again will be based on vSAN technology but for compute clusters vSAN datastores will be standalone and in the boundary of the respective availability zone.

Since the Tenants networking solution will be based on NSX, a key aspect of the NSX-T design is that hosts from a both availability zones must participate in a single overlay transport zone (stretched via L3). Furthermore, the NSX ALB design must take into consideration the placement of the service engines, as those would be part of the tenant(s) data-plane and must move with the NSX Edge VM nodes in the event of an AZ failure event.

Cloud Resources

The resources vCenter Server will be registered in VMware Cloud Director (VCD) and will provide resources that will be pooled and managed by VCD, in addition to the resources NSX-T Manager and the NSX Advanced Load Balancer (ALB) controllers.

The resource clusters from both availability zones will form a single elastic Provider VDC. Tenants will get portions of compute and storage resources in the Provider VDC in the form of Organization VDCs.

For controlling placement of running VMs and replications VM Placement Policies will be utilized. One policy will be assigned to VMs when they are running, and another policy will be assigned to replications. The first policy will create the VMs in AZ1A cluster and the second policy will direct replication to AZ1B cluster.

Now based on the infrastructure design and configuration, from a networking perspective organization VDC networks will span the VDC across AZ1A and AZ1B and VMs will be able to communicate without any configuration change no matter in which zone they are hosted, in addition to the fact that this will also enable consistent security approach for tenant workloads without dependency in which availability zone they run.

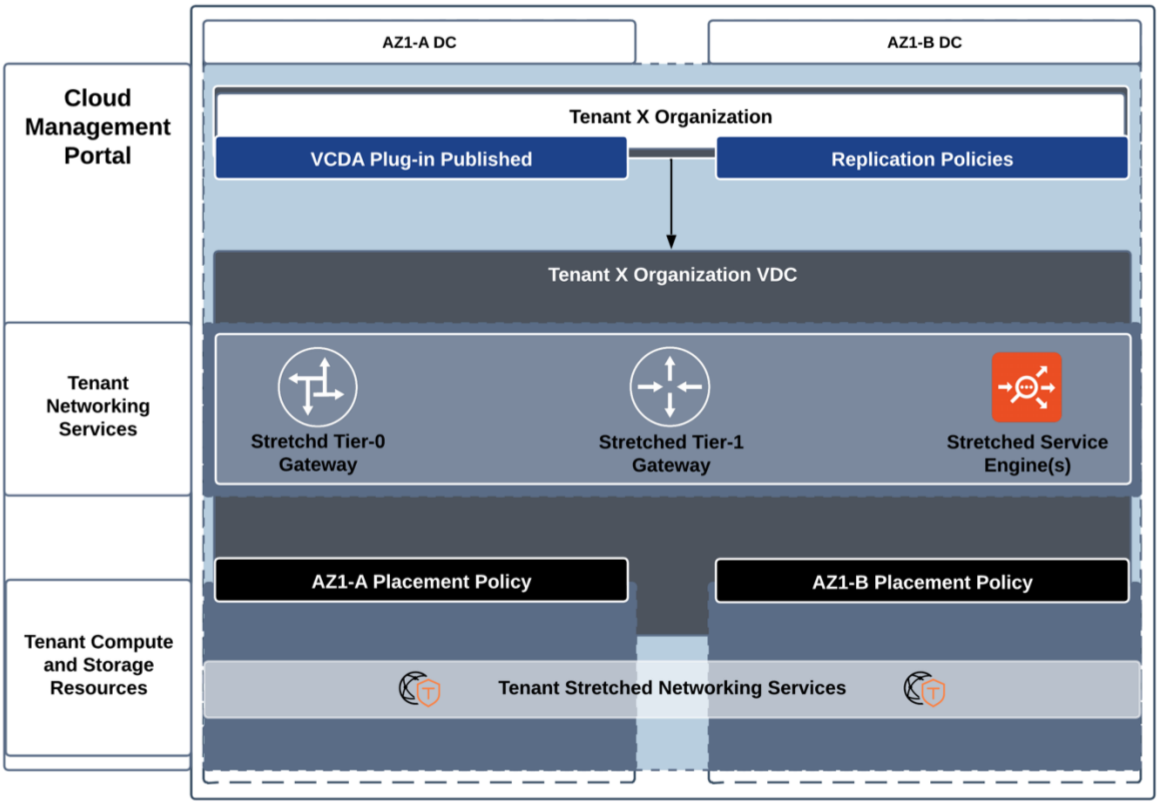

Figure 2 - Service Provider Cloud Resources High Level Diagram

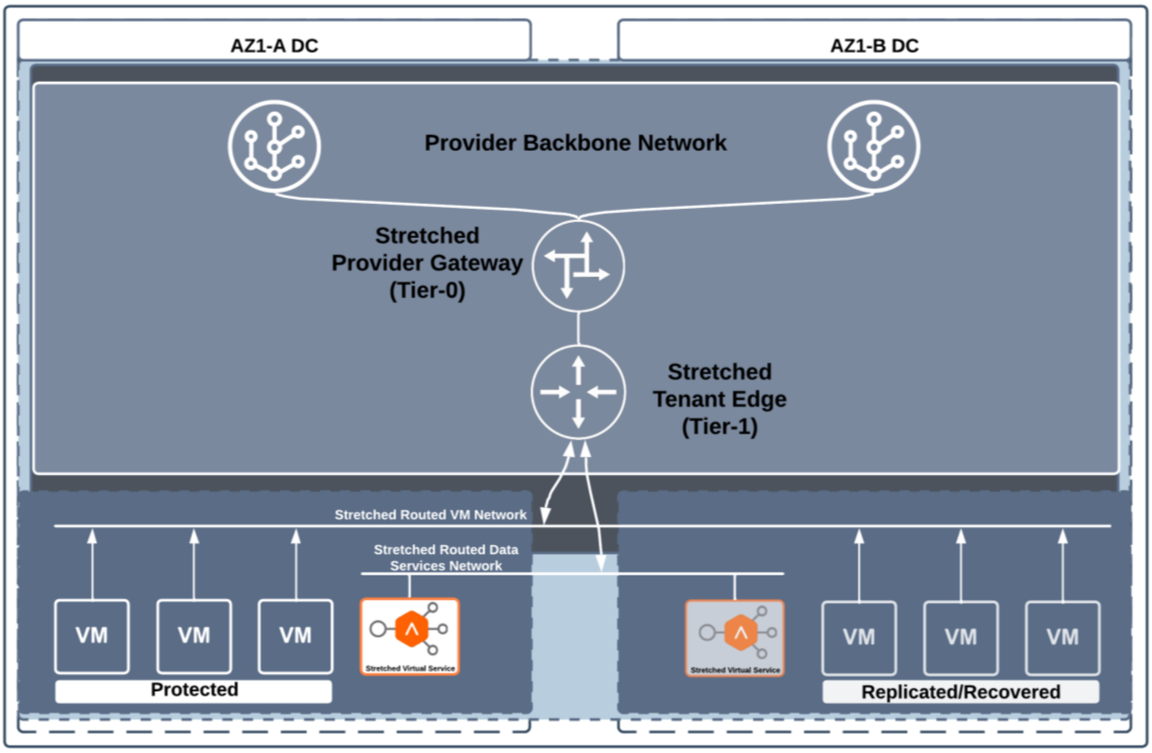
AZ1-A DC
AZ1-B DC
Provider Backbone Network
Stretched Provider Gateway (Tier-0)
Stretched Tenant Edge (Tier-1)
Stretched Routed VM Network
Stretched Routed Data Services Network
VM
VM
VM
Stretched Virtual Service
Protected
Stretched Virtual Service
VM
VM
VM
Replicated/Recovered

Figure 3 – Tenant Cloud Resources High Level Diagram

Design decisions

In this section, we will cover the details of the integration of components with VCDA to achieve the use case requirements; in addition, we will brief the most important decisions for SDDC (vSphere and NSX) and VMware Cloud Director to illustrate what is going to be achieved.

Software Defined Data Centre (SDDC) - VMware vSphere

Prior to working the VCDA magic for this use case, there are a couple of requirements that must be met at the infrastructure level:

1. There is a stretched management cluster that is centralized to managing resources of two availability zones.
2. There is a stretched edge cluster (it's possible to share it with management as well) that is centralized to host:
   a. NSX-T Tier-0 Gateways Edge VM Nodes.
   b. NSX-T Tier-1 Gateways Edge VM nodes of two availability zones.
   c. NSX Advanced Load Balancer Tenants Service Engine VMs.
3. There are two distinct physical data centers with their own workload clusters.
4. A single vCenter Server managing all compute clusters in both availability zones (placed within the stretched management cluster).

Software Defined Data Centre (SDDC) - VMware NSX for Data Centre

1. A single NSX-T Manager managing all transport nodes (Compute, Edge and perhaps management if it's a shared management/edge design) and placed within the stretched management cluster.

2. A stretched NSX-T overlay between the two data centers, while taking into consideration the latency requirements, bandwidth requirements for replication and application communication traffic; as there could be a need to apply QoS if this is a single dedicated link.

3. The network topology and configuration for the Tier-0 gateways with the physical underlay can be either Active/Active or Active/Standby depending on the physical network constraints and design, however it is an assumption that is in-place.

4. An integration of NSX ALB with the resources NSX-T Manager, while making sure that the service engine groups are configured to be placed on the stretched vSphere Edge cluster or shared management/edge vSphere cluster.

Cloud Management Platform - VMware Cloud Director

1. A single VMware Cloud Director deployment managing all cloud resources, where the cells are also placed within the stretched management cluster.
2. A single elastic Provider VDC will be created and clusters from both AZs will be assigned as resource pools to it.
3. VM Placement policies will be used to control where the workloads will be instantiated – one for running VMs and one for workload protections, where you must have at least two placement policies, one for AZ1A clusters and one for AZ1B clusters.
4. Organization VDCs will be configured in Datacenter Groups (DCG) and all Organization networks and gateways will have DCG scope to satisfy Solution Business Requirement SBR001.

Disaster Recovery as a Service - VMware Cloud Director Availability

For some of the VCDA design decisions we'll provide more details about the reasons behind them.

1. A single VCDA instance will be deployed to protect both AZs in a single region.
2. VCDA will be deployed with a dedicated tunnel, manager, and replicator appliances.
3. All VCDA appliances will be deployed in the stretched management cluster – this is a key decision which ensures resilience of the DR service. This decision will enable VCDA appliances to survive outage of entire availability zone and preserve the capability to initiate failover operation.
4. VCDA tunnel and replicator appliances will always run in the availability zone which hosts the resource cluster with the workload protections VM placement policy – this decision will ensure replication traffic will cross the link between both availability zones only once.

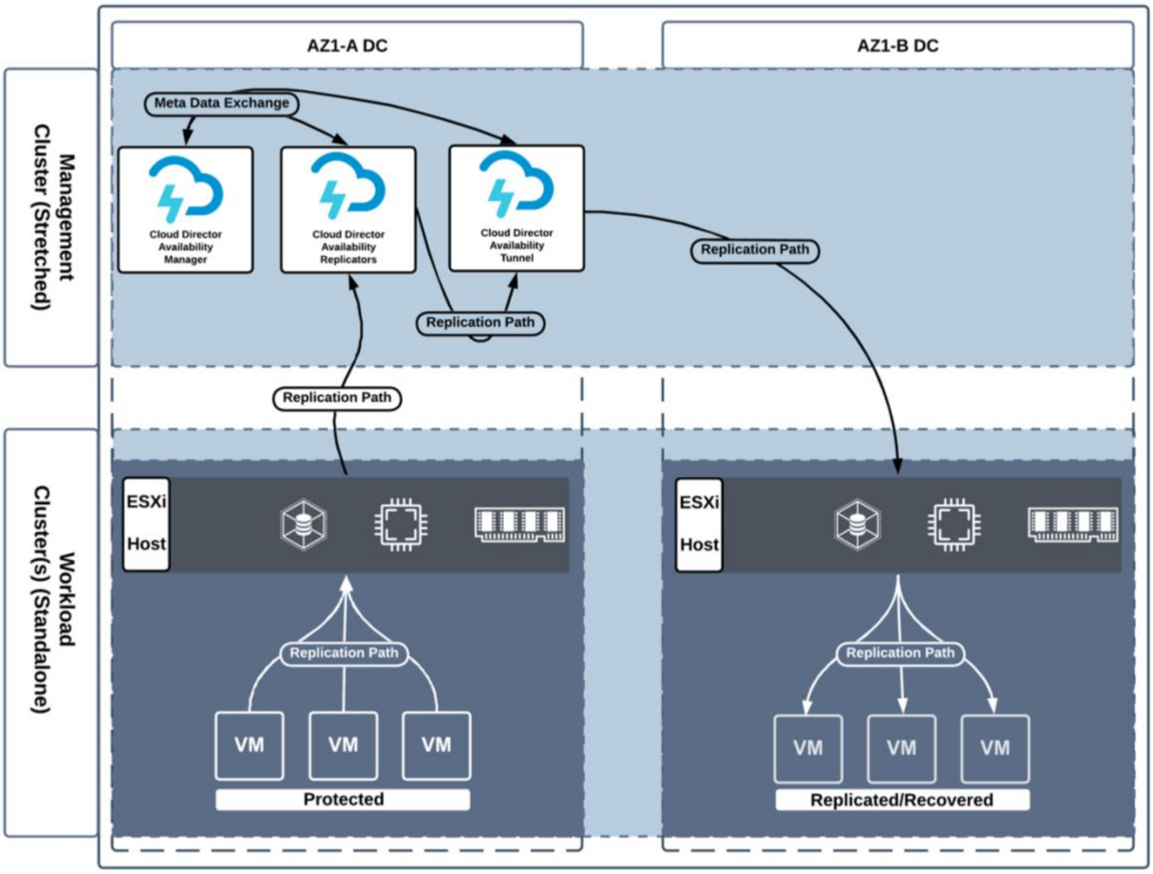

Figure 4 – Cloud Director Availability Placement with Protection Path (Normal) High Level Diagram

Proposed Solution Analysis

Drawbacks of the proposed solution

This solution will always require uncompressed replication traffic to travel over the link between availability zones. This could be a problem if this connection has a limited bandwidth or high latency. In such a scenario it is possible to have a race condition between multiple types of network traffic competing for bandwidth. vSAN for example will not perform as expected and may have unpleasant experience.

To mitigate against this problem, two things must be considered:

- Host VCDA replicators and tunnel in the management cluster (stretched).
- Always group the VCDA tunnel and replicators at hosts from a single site (DRS rules, should and not must).

The replicators usually consume a significant amount of CPU resources. In cases when the management cluster has a small number of hosts and respectively CPUs, this could end up in high demand that is close to the limits. To avoid any negative impact, we recommend avoiding CPU oversubscription in the management cluster.

Alternatives and related risks

The alternative solution could be to host replicators in both resource clusters. Usually, resource clusters have much more hosts and it will be much easier for DRS to load balance CPU demand from VCDA replicators across more physical servers and to scale-out the replicators.

Unfortunately, this solution comes with several significant risks, which will lead to not satisfying the requirements. The risks are related to the conjunction of the proposed solution with VCDA's replicators behavior:

With hosts in both availability zones managed by a single vCenter Server all these hosts will be part of a single LookupService, as such no matter where the VCDA replicators are hosted, they will be able to discover hosts in both availability zones.

This will enable VCDA manager to choose any pair of source and destination replicators for each replication. This means for protecting a VM from AZ1A, VCDA manager will be able to select a source replicator running in AZ1B and destination replicator in AZ1A, and with this in behind, we are up against two risks:

1. With destination replicator hosted in AZ1A if AZ1A is lost due to an outage even if the replica data exists in AZ1B, there will be no way to recover the protected VM, and this is because the destination replicator which should initiate the failover operation will be unavailable.

   Mitigation: Host all replicators in the stretched management cluster which will allow replicators (source or destination) to sustain availability zone outage.

2. The next risk this behavior will introduce is related to the link between availability zones and its utilization, considering the example, the replication traffic will most likely travel through the inter-site link three times:

   a. From the source host in AZ1A to the source replicator in AZ1B.
   b. From the source replicator in AZ1B through the tunnel to the destination replicator in AZ1A.
   c. From the destination replicator in AZ1A to the destination host in AZ1B.

   Such replication flow will create unwanted and unaccepted network traffic between the availability zones and will have an impact on both RPO and RTO.

   Mitigation: Replicators must reside within AZ1A or AZ1B at any given point in time.

Considering the benefits and risks related to this approach it is obvious the risks could have a much bigger negative impact over the entire solution. This is the reason we do not recommend it.

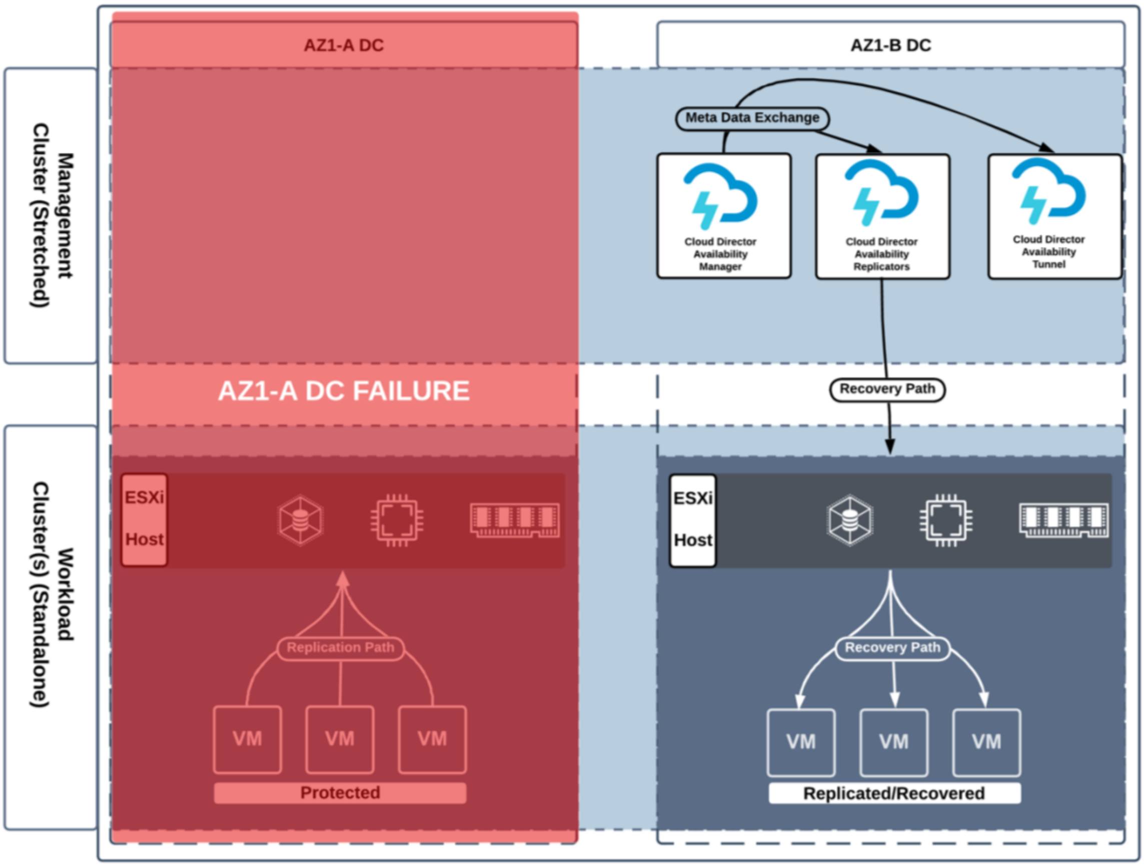
AZ1-A DC
AZ1-B DC
Management Cluster (Stretched)
Workload Cluster(s) (Standalone)
Meta Data Exchange
Cloud Director Availability Manager
Cloud Director Availability Replicators
Cloud Director Availability Tunnel
AZ1-A DC FAILURE
Recovery Path
ESXi Host
ESXi Host
Replication Path
Recovery Path
VM
VM
VM
VM
VM
VM
Protected
Replicated/Recovered

Figure 5 – Cloud Director Availability Placement with Recovery Path (During Failure) High Level Diagram

Solution Business Requirements Versus Solution Design

SOLUTION BUSINESS REQUIREMENTS	TYPE	DETAILS	ANALYSIS
SBR001	Disaster Recovery	# Simplify the recovery process, where recovery of the workloads with no configuration changes at the application level and most importantly its security posture.	# Since we have the cloud resources designed to be available in any AZ, tenant workloads can failover and failback without having to introduce any major change at the Organization VDC level nor at the application level. # Knowing that there is minimal changes required at the organization VDC level, this enhances the RTO of workloads that are being fully failed-over or granularly failed-over, and the focus would be on supporting the application. # Through this solution, even application owners can failover and failback their workloads within a VDC, which also provides agility in terms of disaster recovery plans execution and provides a better RTO for the overall business.
SBR002	Disaster Recovery	# Achieve the lowest possible RTO (Recovery Time Objective) while conducting full or granular workloads failover.	
SBR003	Disaster Recovery	# Tenants would be able to control the failover and failback process and own it end-to-end with the least intervention from the service provider.	

About the Authors

Atanas Stankov is a Senior Solutions Architect in VCPP Engineering focused on DR solutions. In his role he works with cloud providers participating in VCPP program helping them design new DR solutions or improve existing ones.

Abdullah Abdullah is a Staff Consulting Architect in PSO METNA and is a CTO Ambassador. He currently leads business transformation engagements that are focused on VMware's SDDC stack, with a focus on VMware Cloud Service Providers for the past 6 years; and has been in the industry for more than 15 years and has successfully delivered 100+ customer engagements across METNA.

Abdullah is an 11 years vExpert, and a double VCDX #270 in NV and CMA, and is an avid contributor and supporter within the VMware community, where his contribution is focused on the VCDX, vExpert and Education SME programs.

Acknowledgments and Reviewers

Tomas Fojta – Senior Staff Architect, VCPP Engineering

Nikolay Patrikov - Senior Technical Product Manager, VCPP Engineering

Ross Wynne – Senior Staff Consulting Architect

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 vmware.com Item No: vmw-wp-tech-temp-a4-word-101-proof 6/20