VMware vCloud® Architecture Toolkit™
for Service Providers

VMware Hybrid Cloud Extension Architecture Field Guide

Version 2.9
January 2019

Ray Heffer **Double VCDX**
Principal Architect

VMware, Inc.
3401 Hillview Ave
Palo Alto, CA 94304
www.vmware.com

Contents

Executive Summary

The VMware Cloud® Provider Program is a global network of over 4,200 service providers who have built their cloud and hosting services on VMware software solutions. These service providers deliver world-class cloud and hosting services to their customers across the globe.

VMware Cloud Providers are uniquely positioned to offer their services to the market and become a seamless extension of existing VMware enterprise customers on-premises data centers. Managed service providers have traditionally operated data centers hosting multiple tenants, and have developed reference topology models. This platform offers tenant separation while allowing the providers to offer value-added services from common management platforms.

Cloud Providers with the VMware Hybrid Cloud Extension (HCX) service can provide customers with a true hybrid cloud experience for workload portability, agility, disaster recovery and business continuity. Cloud Providers can take the lead with hybridity, abstracting customer on-premises and cloud resources as one seamless cloud. No changes are required on the source network infrastructure, eliminating complexity for tenants of the cloud platform.

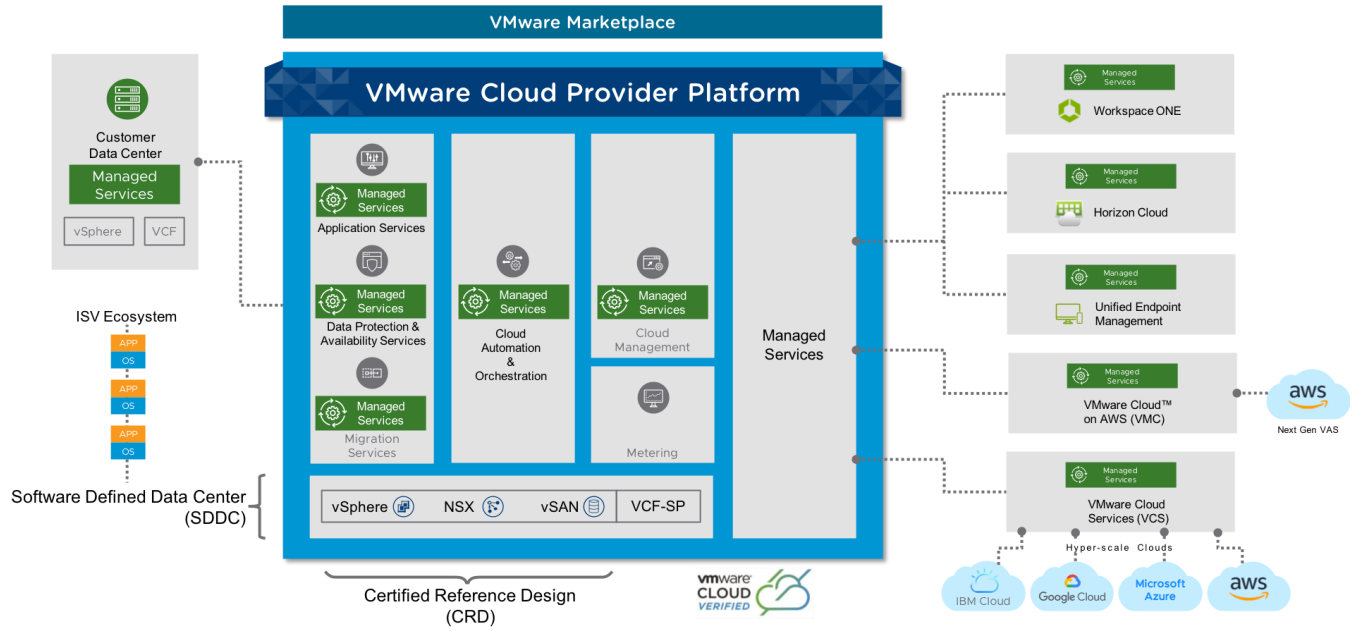

Figure 1. VMware Cloud Provider Platform

What Is Cross-Cloud Mobility?

Cross-Cloud Mobility is the capability to pair any cloud infrastructures together and expect each cloud to act as an extension to the others. The HCX platform becomes the basis on which cross-cloud mobility is provided by leveraging infrastructure services (Any-Any vSphere® Zero downtime migration, seamless Disaster Recovery, enabling Hybrid Architectures, and so on) to provide tangible business values.

What Challenges Does HCX Address?

Although VMware HCX enables migration as the primary use case, there are many solved challenges:

- Migrations often require costly and time-consuming workload assessments.
- Migrations from legacy to next-generation environments are complex due to different versions of the underlying infrastructure.
- Business-critical applications require migrations without downtime.
- Applications often suffer performance degradation during migration.
- The WAN and LAN can be negatively affected with replication and vMotion® occurring over the network.
- Time to migrate the workloads is affected due to network latency and packet loss.
- IP address changes introduce complexity.
- Rollback can often be complicated and affects production.

What Is a Legacy Cloud Environment?

- Often vSphere 5.x with VLANs (no NSX® or Virtual Extensible LAN（VXLAN))
- Older Intel® generation of CPUs that vary (E5, E7, and so on)
- One-offs vSphere clusters for customers
- Older storage technology

What is a Next-Generation Environment?

- VXLAN with NSX
- vSphere 6.x
- Full Software-Defined Data Center (SDDC™) stack including NSX and vSAN™
- Newer Intel CPU, standard across hosts or very limited set of differences
- New data center locations and data center consolidation

Use Cases for HCX

There are many use cases for HCX, not just limited to the common workload migration scenario. Although workload migration often meets key business requirements for many organizations, HCX can also meet other requirements that the flexibility of a true hybrid-cloud demands.

HCX can extend the layer 2 network of a customer on-premises data center to the cloud, including private clouds and VMware Cloud on AWS (Amazon Web Services).

- Distributed Applications—Portability of workloads both on-premises and the cloud
- Customer Hybridity Services—Burst between on-premises and the private cloud
- Any-to-Any vSphere migrations with zero downtime
- Migration of workloads from legacy to next generation hybrid cloud data centers.
- Disaster Recovery as a Service (DRaaS) – for datacenter to datacenter use cases.
- Disaster avoidance measures—In advance of a disaster (for example, extreme weather, natural disasters) for both the provider and customer

HCX Components

VMware Hybrid Cloud Extension is initially deployed as a virtual appliance (HCX Manager), along with a vCenter® plug-in that enables the HCX feature. After the HCX Manager virtual appliance is deployed, it is responsible for deploying the other appliances required for HCX. Mirror appliances are deployed at both the source and target (cloud provider) locations. These deployments provide the capabilities outlined below.

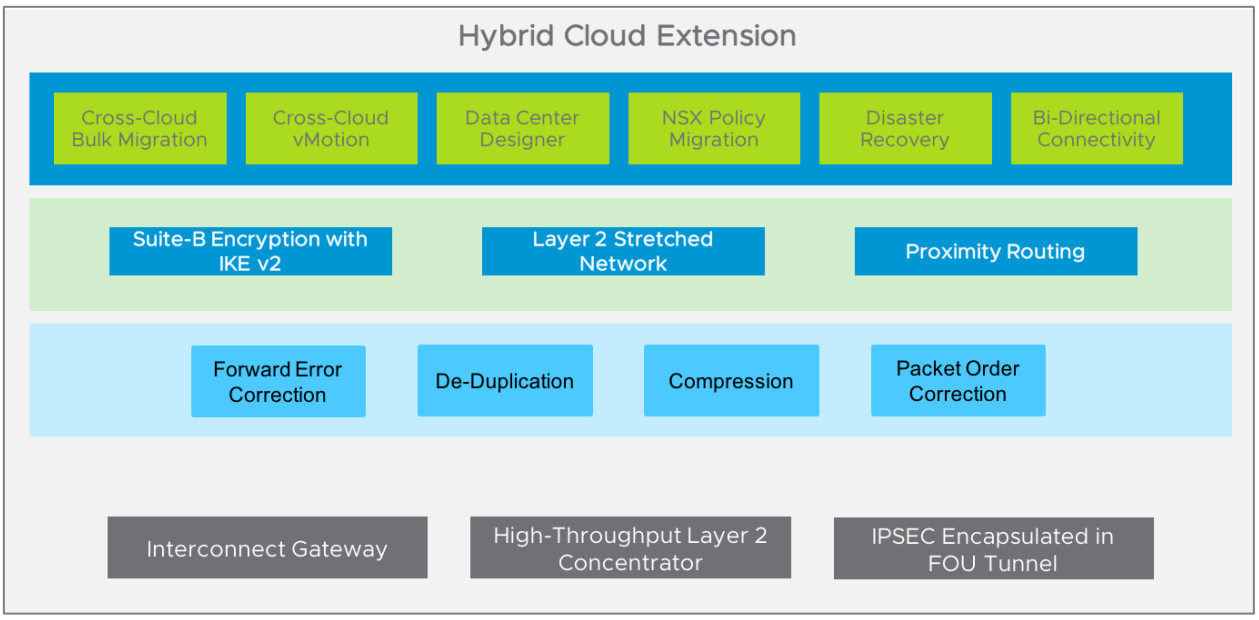

Figure 2. HCX Capabilities

HCX Manager

The HCX Manager is an open virtual appliance (OVA) downloaded from the VMware HCX service after logging in to VMware Cloud Services (https://console.cloud.vmware.com). It provides a management plug-in for vCenter that bootstraps the HCX services and manages virtual appliance deployment and connectivity. There can only be one HCX Manager per vCenter.

After HCX Manager has been deployed and activated with connect.hxc.vmware.com, the administrator will select the data center location (for example, Raleigh, NC) and complete the vCenter Single Sign-On (SSO) configuration. As a cloud provider, this step has already been completed at one or more cloud provider data centers. The cloud provider locations will be referred to as Cloud Target Sites.

The next step is to deploy the HCX Interconnect service virtual appliance that is deployed by HCX Manager at both the source and destination sites, and an Internet Protocol Security (IPSEC) VPN tunnel is established across the two sites. The WAN Optimization service is also deployed in this fashion, and provides secure connectivity between the source on-premises site and the cloud provider (target site).

The Network Extension service is not required, but highly recommended because it provides a layer 2 extension where the same IP subnet extends to the target site. This service provides great value with workload mobility and demonstrates the real value of the hybrid cloud.

The HCX virtual appliances are:

- HCX Manager
- HCX Interconnect service (CGW)
- Network Extension service (L2C), required for L2 extension
- WAN Optimization service

HCX Interconnect Appliance

The HCX interconnect service, also referred to as the WAN Edge, provides resilient access over the Internet and interconnects multiple private lines to the target site. This service provides strong Suite B encryption and traffic optimization between the source and target sites. It also simplifies secure pairing of site and management of HCX components.

The Interconnect appliance provides connectivity between the source and target data centers by using site pairing, and bootstraps the bi-directional hybrid connection, providing resilient connectivity over multiple circuits, such as Internet, Multi-Protocol Label Switching (MPLS) and Direct Connect links.

Internet connectivity for most enterprise customers might provide significantly higher throughput than other private circuits. A preferred path of connectivity can be configured, and customers can select other links as secondary or tertiary.

- Supports: vSphere® Distributed Switch™ (vDS) (all versions), Cisco Nexus® 1000v (all versions)
    - vSphere Standard Switch (vSS) - Not supported at this time
- No NSX installation (or standalone edge) required on-premises
    - If NSX is present, this adds the ability to stretch VXLANs
- Minimum 100 Mbps network connection
    - Latency not a major concern
- Supports Maximum Transmit Unit (MTU) emulation (Jumbo-frame emulation over WAN)
- Traffic Fairness, Elephants (large flows) & Mice (small flows)

The WAN Edge provides WAN Optimization (compression and de-duplication), intelligent routing and IPSEC with Suite B encryption (using certificates, and not pre-shared keys). Features such as vMotion and vSphere Replication are securely proxied behind this abstraction layer with a fully encrypted connection.

Although vSphere 6.0 provides cross-vCenter vMotion, the two vCenter servers must be linked (for initiating vMotion from the GUI) and reside on the same SSO domain. Other considerations such as Fault Domain and routing requirements might not be suited to tenant and cloud provider environment. Using HCX these constraints do not need to be considered.

Scenario 1: Virtual and Physical Workloads

The scope of this document is specific to vSphere environments, whether that is on-premises, VMC on AWS or vCloud Director® for Cloud Providers. However, there might be situations where a customer has both virtual and physical workloads. If the VLANs where a physical server is located can be presented to vSphere, a layer 2 network extension can remain in place and enable the physical server to communicate with virtual machines (VMs) at the target site. This extension effectively bridges the VLAN on the source, for example, VLAN 100, to VXLAN 10000 at the cloud provider side. It is still a single broadcast domain, so the default gateway IP address is the same at the source and target site.

Scenario 2: Multiple VLANs

Most examples within this document discuss moving VMs that reside on a single VLAN; however, it is very likely that customers have hundreds of VLANs where VM workloads reside.

In this situation, you want to avoid deploying hundreds of High-Throughput Layer 2 Concentrators so that VLANs can be grouped on a single concentrator. For example, VLANs 234, 300-600, and 624 are stretched on a single concentrator across the WAN. If the source site has NSX deployed, it is also much more efficient to stretch VXLAN from the source to VXLAN on the target site. This is even possible if NSX versions differ because NSX domains are not federated. This is the role of the concentrator.

Scenario 3: VLAN to VXLAN

In this scenario HCX is tapping directly to the broadcast domain. It's not sending the traffic through vCenter, the VM is directly sitting on the network at the source and it's also sitting on the same network at the destination on an internal interface. The layer 2 concentrator has code that bridges and manages ARP (Address Resolution Protocol) suppression and ensures that there is loop detection.

WAN Optimization Appliance

This is the second service to install by the bootstrap process. The WAN optimization service provides compression and de-duplication across the WAN. WAN optimization improves performance characteristics of the private lines or internet paths by leveraging techniques like data de-duplication and line conditioning. This optimization makes performance closer to a LAN environment.

VMware HCX requires a minimum throughput of 100Mbps and excels on the Internet connections where workloads can be securely and optimally migrated.

You have wrapped IPSEC into a User Datagram Protocol (UDP) encapsulation that allows intelligent routing to find the path of lowest latency. Brown outs can be avoided on the WAN or the Internet because you can take advantage of the entropy on the WAN and bundle many tunnels. This is 5-tuple WAN encapsulation.

- WAN optimization service provides compression and de-duplication across the WAN.
    - vMotion uses an average of 250Mbps with HCX.
- WAN Optimization provides forward error correction and ensures that the performance is optimal.

Network Extension Appliance

The network extension service provides high-throughput connectivity with integrated proximity routing that unlocks seamless mobility and simple disaster recovery plans across sites. This service enables the mobility of VMs seamlessly and keeps the same IP and MAC address across sites without the need for NSX at the source site. Proximity routing can eliminate the traffic trombone effect, and vMotion-aware proximity routing ensures that vMotion tasks are not aborted due to spikes in network latency.

HCX is bi-directional. When a network is extended from the source site to the target site, it will deploy the HCX layer 2 concentrator on the source, and a mirror image of the concentrator will be deployed on the destination to prepare for network to be stretched.

- Stretches the Layer 2 network (VXLAN or VLAN) to the target site (cloud provider).
- Emulates Jumbo-frames on a WAN with smaller MTU.
- Elephant or Mice flow detection and inter-CPU balancing
- 4-6 Gbps per VLAN, scaling out for further performance
- Each micro-flow scales to at least 1Gbps.
- Scale-out cluster can fully utilize multiple 10Gbps lines.

The Layer 2 Network Extension is not a requirement, but it is recommended because it reduces the complexity of workload migration. The L2 extension taps into the on-premises vDS or Cisco Nexus 1000V, leveraging sync ports and bridging VLANs and VXLAN from source to destination. The network extension also provides MTU emulation over the WAN for jumbo-frames.

Customers are not expected to be networking experts, and there is typically no requirement to make any changes on the source site network infrastructure. Note that HCX does not support the extension of the management network(s) on which the HCX components are deployed.

Research suggests that the majority of network flows within the data center are short-flow (Mice); however, the majority of packets between data centers belong to long-flow (Elephants). The short-flow packets are associated with latency-sensitive and high burst applications, whereas long-flow are typical large transfers of data and high-throughput is required.

Multi-site VLAN Stretched Topology

If you are extending VLAN 100 from the source data center, HCX will orchestrate the creation of the same named VM on the VXLAN on the destination site and will provide a bridge between the data centers using the Layer 2 Concentrator.

The same VLAN can be stretched to different cloud endpoints using a V (multi-point) topology.

Source > Destination A
Source > Destination B

High-Throughput Layer 2 Concentrator

The high throughput Layer 2 Concentrator is a scale-out architecture that provides UDP encapsulation overlay over an L2 extension. This architecture provides multi-gigabit throughput for an extended VLAN across the WAN, of approximately 4-6Gbps. Each flow can reach up to 1Gbps, although multiple appliances can be deployed to increase throughput on the network links.

IPSEC encryption is one layer, not per circuit or connection with IPSEC wrapped with UDP encapsulation. The Network Extension service ensures that the port groups are extended at the destination (for example, a L2 extension where the same IP subnet extends to the target site).

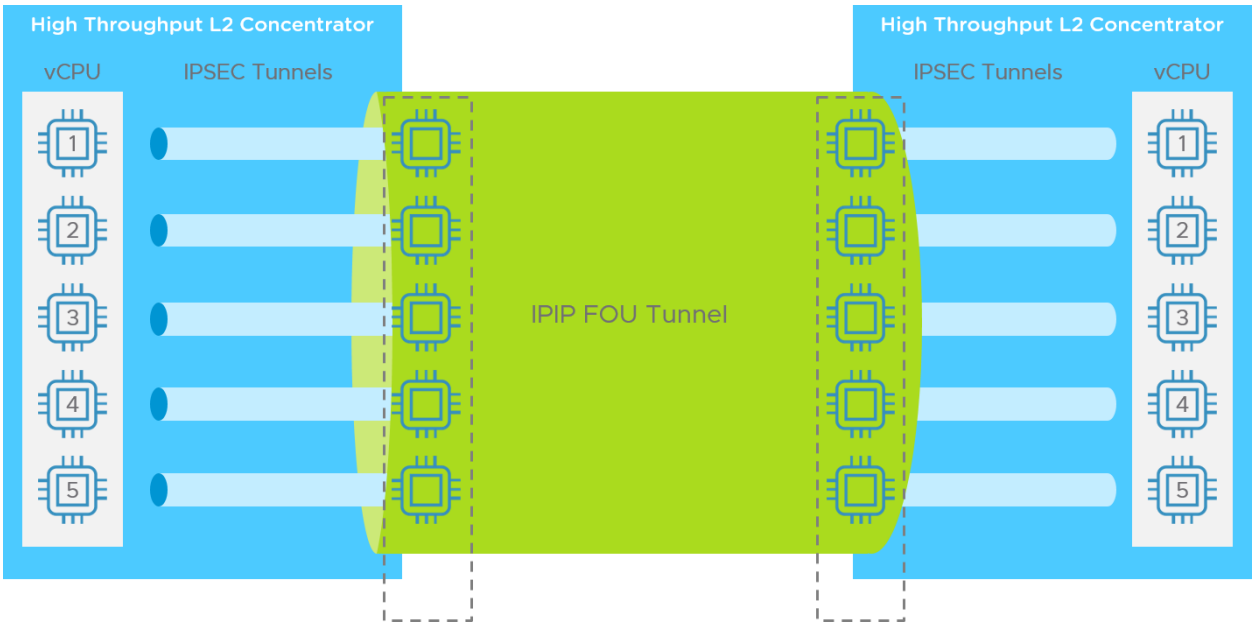

Figure 3. High-Throughput Layer 2 Concentrator

Proximity Routing

Proximity routing is a function that the High-Throughput Layer 2 Concentrator performs. The concentrator is a virtual appliance that is managed by the HCX manager virtual appliance.

- Forces routing symmetry
- Enables stateful intermediate devices (FW/IPS)
- Migration and Disaster Recovery (DR) aware
- vMotion aware Proximity Routing

Proximity routing optimizes connectivity both to and from HCX migrated VMs by integrating the HCX Network Extension and Migration state changes with the NSX Dynamic Routing configuration at the HCX Enabled Cloud.

It also allows VMs that are extended over a layer 2 network extension to route optimally (egress) via the target cloud sites first hop gateway. By dynamically injecting VM routes into the routing protocols, ingress traffic from the local and target data centers will use an optimal (non-hairpinning or non-tromboning) path to reach the extended VM, while ensuring that all flows remain symmetric.

vMotion with HCX Proximity Routing

vMotion based migration requires zero downtime and is compatible across multiple versions of vSphere (5.5 or higher). It is the same vMotion protocol you use today, but it is proxied through the HCX pipeline so there doesn't need to be direct connectivity between the source ESXi and destination ESXi hosts.

The HCX protocol proxy performs flow control for vMotion, so if there are network latency spikes, the vMotion won't abort. Performing vMotion without HCX, even using a WAN optimizer might be troublesome due to the fact that vMotion is very aware of TCP back-off, and it can be affected by connections being closed which will terminate vMotion.

The vMotion proxy is created in both the source and target cloud provider site and represents a host. This allows `hostd` and `vpxa` to be leveraged as if the proxy were an ESXi host. HCX is streaming vMotion traffic through the proxy to the target cloud provider site. This is what allows for the same vMotion to work across the WAN.

Note `hostd` is a management service daemon that runs on an ESXi host and is responsible for managing host level operations. `vpxa` is a vCenter agent that runs on the ESXi host and is responsible for communication between the host and vCenter.

Workload Migrations

There are two types of migration for VMs with Hybrid Cloud Extension: vMotion and replication-based migration (Bulk Migration).

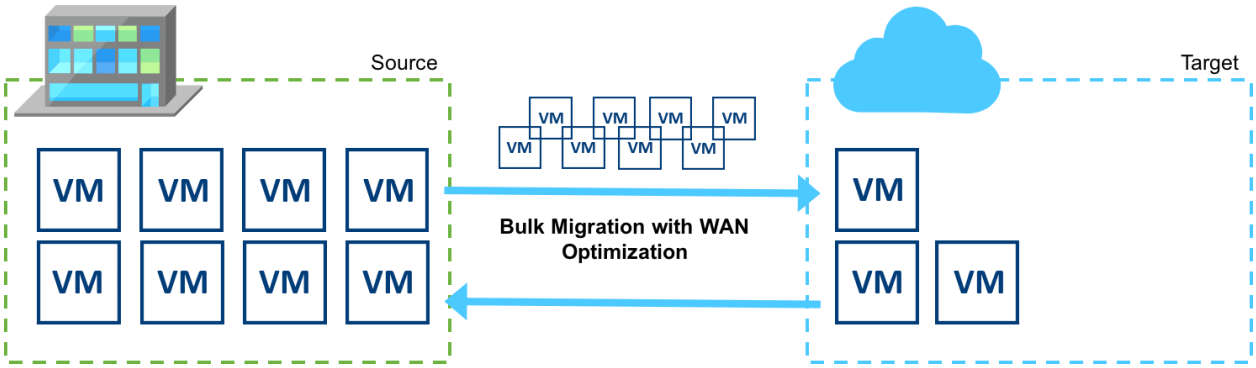

Figure 4. Workload Migration

The 80/20 Rule

When you plan for large-scale workload migrations, the 80/20 rule is applicable. 20% of workloads are typically mission-critical and require zero downtime, whereas the remaining 80% of workloads can be migrated with a maintenance window to allow for a reboot. Bulk migration allows workloads to move at scale, and provides options to schedule the network swing or failover. Maintenance windows can be configured by using the schedule failover option.

vMotion to the Cloud

For mission-critical applications, vMotion can be used from the source site (on-premises) to the target site (cloud) with zero downtime. HCX allows vMotion to be performed securely across the Internet. The HCX Interconnect solves such issues as poor line conditions by pooling multiple paths to connect to the destination cloud.

Results with cloud providers adopting HCX have shown impressive statistics. One example of a 25GB-VM running on VMware Cloud Foundation™ (VCF) was migrated with vMotion from Mexico to Paris, France. During this test, the vMotion finished in 13 minutes with latency on the network of around 139ms. The same workload was moved with vMotion from Mexico to Toronto in 8 minutes with a latency of around 55ms.

WAN optimization methods, such as de-duplication, are crucial for workload migrations. Even if migrating between two cloud data centers and leveraging the cloud provider backbone, there are still multiple paths to the target. If a path fails over or a latency is introduced on a link, the impact on the vMotion is minimal thanks to HCX interconnect technology.

Local (LAN based) vMotion also has some limitations. If one host is running the Sandy Bridge CPU chipset and the other host is using Skylake, cluster level Enhanced vMotion Compatibility (EVC) is required to maintain CPU instruction compatibility. For more information, see the knowledge base article "Enhanced vMotion Compatibility (EVC) processor support": https://kb.vmware.com/s/article/1003212

Using vMotion with HCX, CPU flags are injected allowing the VMs to migrate with no downtime or any change to the source or destination cluster. Think of this as per-VM EVC. After the VMs are migrated, the VMs can stay running on a new version of the hardware, and adopt the new instruction set on the next reboot or remain in a backward compatibility mode.

Bulk Migrations

Bulk migrations do not use vMotion, but leverage host-based replication. HCX provides protocol-aware security for proxies and extends layer 2 network. The VM is replicated while the source VM is still running, with a failover step that reboots the VM at the target site (cloud provider) and a network swing.

During a migration, the VM is running while being replicated through the HCX Interconnect pipeline. All traffic is encrypted, compressed and de-duplicated and the Layer 2 Network Extension ensures that the VM on each site can use the same IP address.

Scheduling a Failover for Bulk Migration

When a failover window is configured, a replica VM is created on the target site and changed blocks are sent across to the target site. The replica VM continues to sync until the failover time. The source VM is shut down. Final blocks are sent to the target site and the VM is rebooted.

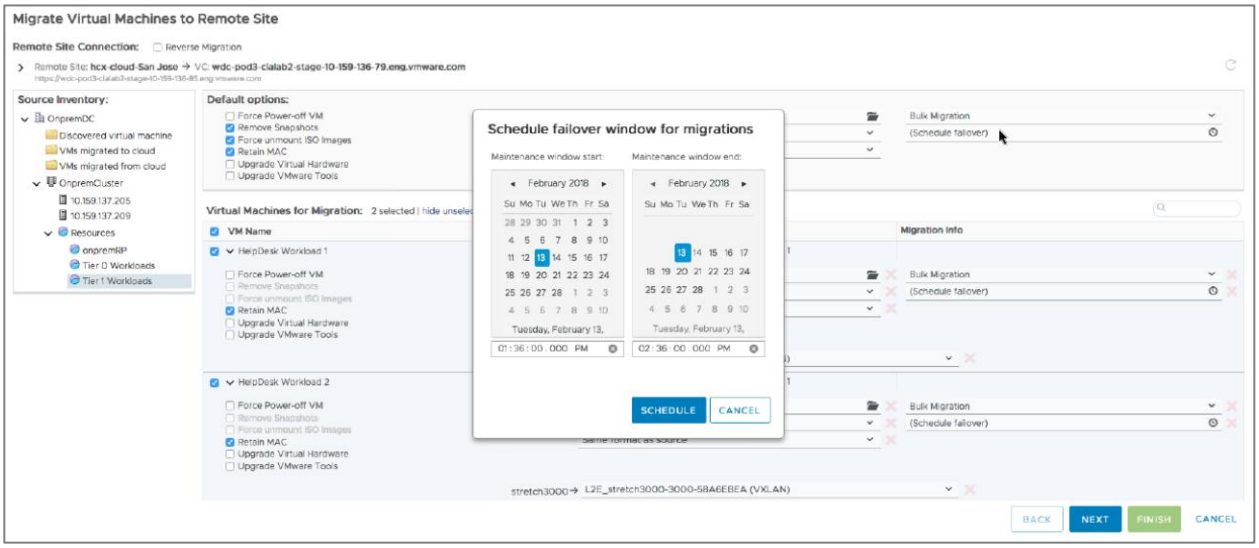

Figure 5. Scheduling a failover window for migrations

Network Swing

The on-premises network is extended to the cloud with HCX by using layer 2 extension, including VLAN or VXLAN, and legacy versions of vSphere 5.0 or higher. Both Cisco Nexus 1000V or VDS switches are supported.

While the VMs are being migrated, both of the default gateways at the source and target sites are available. However, a swing will need to take place so that the edge gateway at the target site will take over as the default gateway. This process is known as the network swing.

The network swing will remove the layer 2 extension (de-coupling the two broadcast domains), and it will connect the VLAN or VXLAN to the NSX edge on the destination side. This edge will take over as the default GW. Perform the network swing during a maintenance window.

The network swing is part of the HCX workflow. However, customers may choose to keep the layer 2 extension in place for an extended period of time. The layer 2 extension does not need to be temporary.

Disaster Recovery

Mobility through migration of workloads is one aspect of HCX, but the protection of workloads is also possible with its DR capabilities.

Typically, there are five key pain-points that customers face with DR planning:

- DR plans are lacking or put together last minute.
- DR plans can lack any prior risk assessment and lack RPO or RTO.
- The DR plan is outdated and doesn't cater for newer infrastructure.
- DR planning hasn't been fully tested.
- Skills required to execute DR plan are not always available (for example, Contractors, staff turn-over).

Cloud providers can address these common issues by offering a true state of hybridity with HCX and seamlessly extending the customer on-premises network. VMware HCX can be used to provide active-active protection for recovery of both VMs and infrastructure from a customer's on-premises vSphere environment to the provider's hosted vSphere or vCloud Director environment.

Compatibility and Interoperability

Cloud Providers can migrate workloads from vSphere version 5.0 to version 6.5. HCX is backward compatible back to vSphere version 5.0 for bulk migrations and vSphere 5.5 for vMotion based migrations.

Site Pairing

Customers deploying HCX at the source site are required to download an OVA-based virtual appliance from the HCX Cloud portal. HCX is deployed to vSphere and a vCenter plug-in is installed to enable the HCX functionality. After deployed, the HCX instance (vCenter) is registered with the cloud provider site URL. The registration will establish a new site pairing.

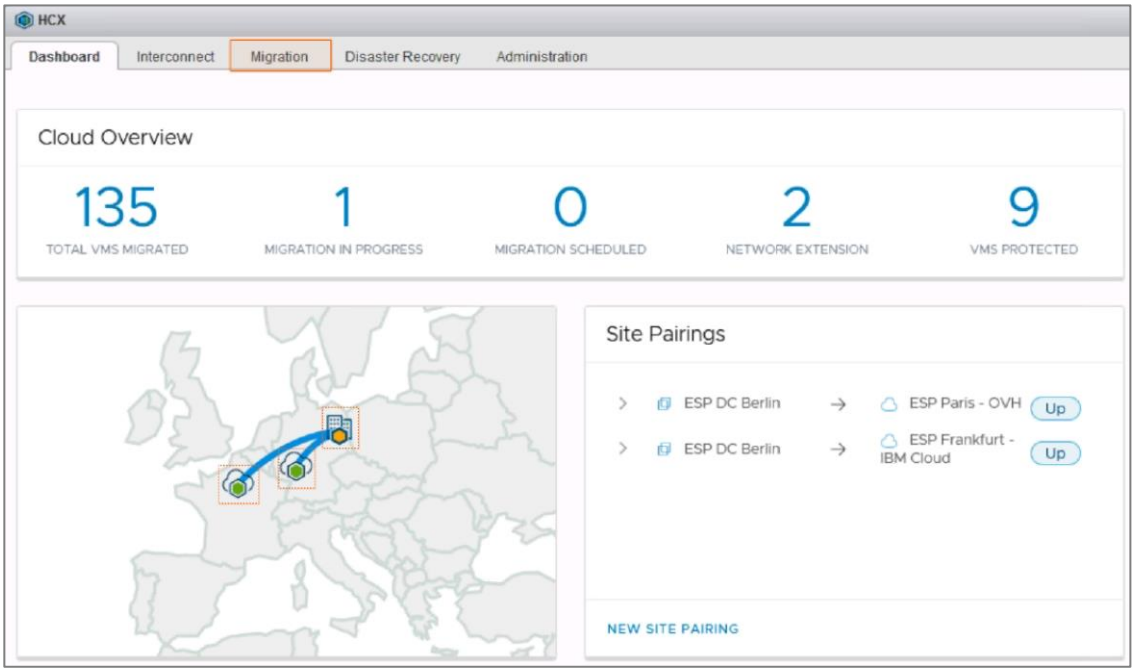

Figure 6. HCX Dashboard and Site Pairings

When HCX creates a site pairing and registers with the Cloud site, three services are deployed as virtual appliances. The appliances automatically bootstrap with the HCX target site (for example, cloud provider). When the configuration finishes, the services start at the destination site and an IPSEC VPN tunnel is established across the two sites. This process is known as the site pairing.

HCX keeps inventory of vCenter objects including hosts and is able to differentiate between HCX proxy-hosts and real hosts. It excludes HCX proxy-hosts from user operations. For example, users will not see an option of proxy-host when they try to move a VM from VMC back to on-premise.

Design Best Practices

Firewalls and Ports

When preparing to deploy HCX, it is important to ensure that the proper firewalls settings are configured to allow the required connectivity for HCX components between environments. The HCX managers require communications over TCP Port 443. HCX Interconnect and Extension appliance require communications over UDP port 500 and 4500 between sites. Please refer to Appendix A for more details regarding HCX port communications.

Network Swing and Proximity Routing

Default gateways initially point to an on-premises gateway. After the VMs move as part of the migration, you can swing the default gateway to the destination site.

A hybrid scenario might occur where you don't want to perform a network swing for all VMs in one fell swoop. In this case proximity routing can be used. For VMs that have already moved over to the destination side, it might not be desirable to be routed through the source infrastructure default gateways.

Migration Between Legacy and Next Generation Hardware

As mentioned previously, one use case of HCX is the workload migration from a legacy to the next-generation architecture. This migration often results in varying CPU chipset architecture between the older and newer hosts, for example, migrating from the Sandy Bridge CPU architecture to Skylake.

Instead of using EVC to maintain CPU instruction compatibility (a cluster level change), HCX can inject CPU flags on a per VM basis. After the VMs are migrated, the VMs run on a newer version of the hardware and adopt the new instruction set on the next reboot.

Monitoring the Cloud Infrastructure

As an infrastructure administrator, it is important to monitor Cloud Resources Usage on both the source site and target site via the HCX Dashboard. These cloud resources include CPU, Memory and Storage of the source and target data centers. Activity logs keep a record of each migration and are useful for auditing purposes.

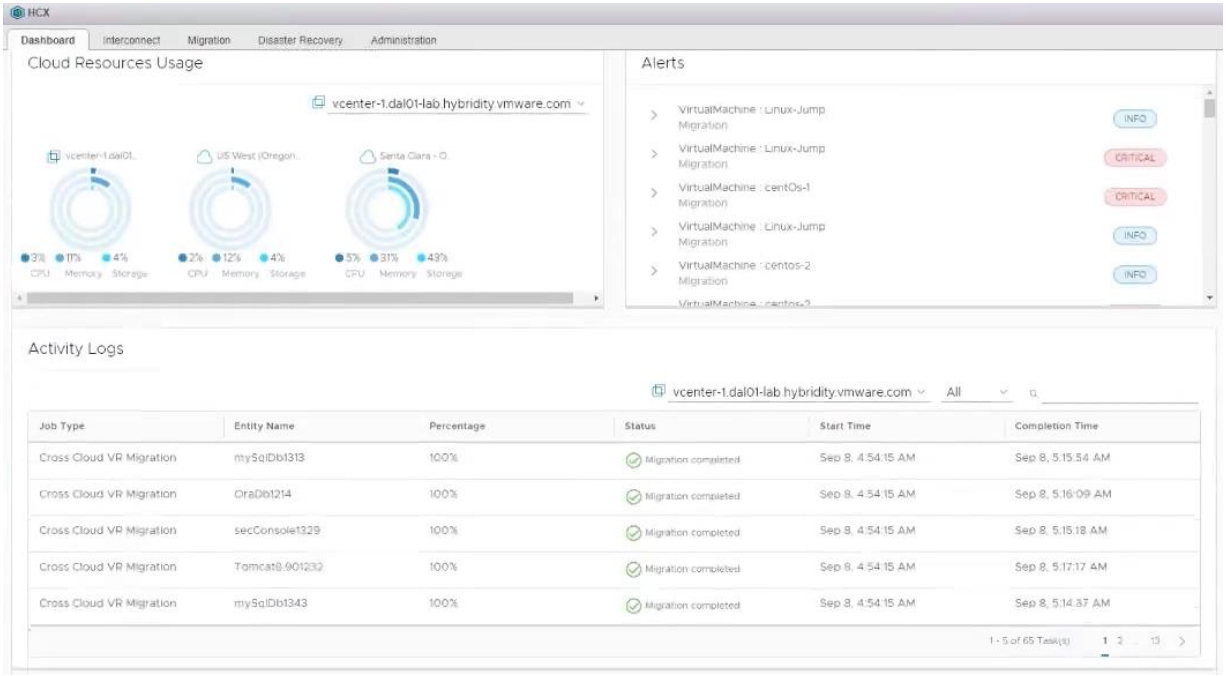

Figure 7. Monitoring Cloud Resource Usage with HCX

vCenter alarms and alerts can be used for compliance and audit purposes, in addition to integration with vSphere SSO at the source and destination sites.

A vRealize Operations management pack is available for HCX. This management pack is recommended for migration planning and helps you understand if you can increase the migration intensity.

Consider the following factors:

- When was the last migration wave?
- How long did it take?
- How much data was transferred?

Use these statistics to improve migration planning. Check how many workloads you performed in your allotted timeframe. You can increase the number of workloads being migrated in the wave.

When monitoring VMs, vRealize Operations can also provide in-guest monitoring. The statistical history of VMs is not lost when it moved to the cloud. This history allows the customer to see the performance differential on-premises versus the cloud.

vMotion

vMotion workloads with HCX uses the same vMotion protocol, but are proxied through the HCX pipeline. vMotion can work across the WAN and significantly benefits from WAN optimization. Take Mexico to Paris as an example. Although this scenario isn't common practice, it shows what can be accomplished. Customer On-boarding accelerated.

Managed Infrastructure and Roles

Commonly, traditional workload migrations enlist the services of a professional service organization to carry out application and workload assessments and make planning for mission-critical workloads.

Roles

System admin—Global administrator (backend flow, no user access)

Tenant admin—Tenant administrator (vCD) insulated from backend operations

Legacy Environments

In an ideal world, all workloads are equal and don't present any restrictions or constraints. However, it is far from reality and there are always at least one VM workload that fits into the "legacy" category.

One example is the VMs that use Raw Device Mappings (RDMs). In this case the L2 Extension bridges the source and target sites to extend the VLAN without any IP address overlap issues. While the legacy workload remains the source site, this allows for the remaining workloads to be moved to the cloud.

HCX REST API

The HCX API Explorer is available to both customers (HCX Enterprise) and cloud providers (HCX Cloud) by accessing the following URL: https://<HCX-MGR-IP>/hybridity/docs/index.html. The APIs are available for the following functions:

HCX Cloud (vCenter and vCloud Director)

- Appliance Management
- Audit Logs
- Disaster Recovery
- Interconnect
- Networking
- Platform
- Tech Support

HCX Enterprise (Tenant)

- Appliance Management
- Audit Logs
- Cloud Registration
- Disaster Recovery
- Migration
- Network Extension
- Platform
- Tech Support

Author and Contributors

Authors and contributors of this guide include:

- Ray Heffer, Principal Architect, VCDX #122 (Twitter @rayheffer)
- Harold Simon, Staff Cloud Solutions Architect, VCDX #80 (Twitter @harold_simon80)
- Gabe Rosas, Staff Engineer II, VMware HCX Technologies (Twitter @gabe_rosas)

Appendix A: Network Ports

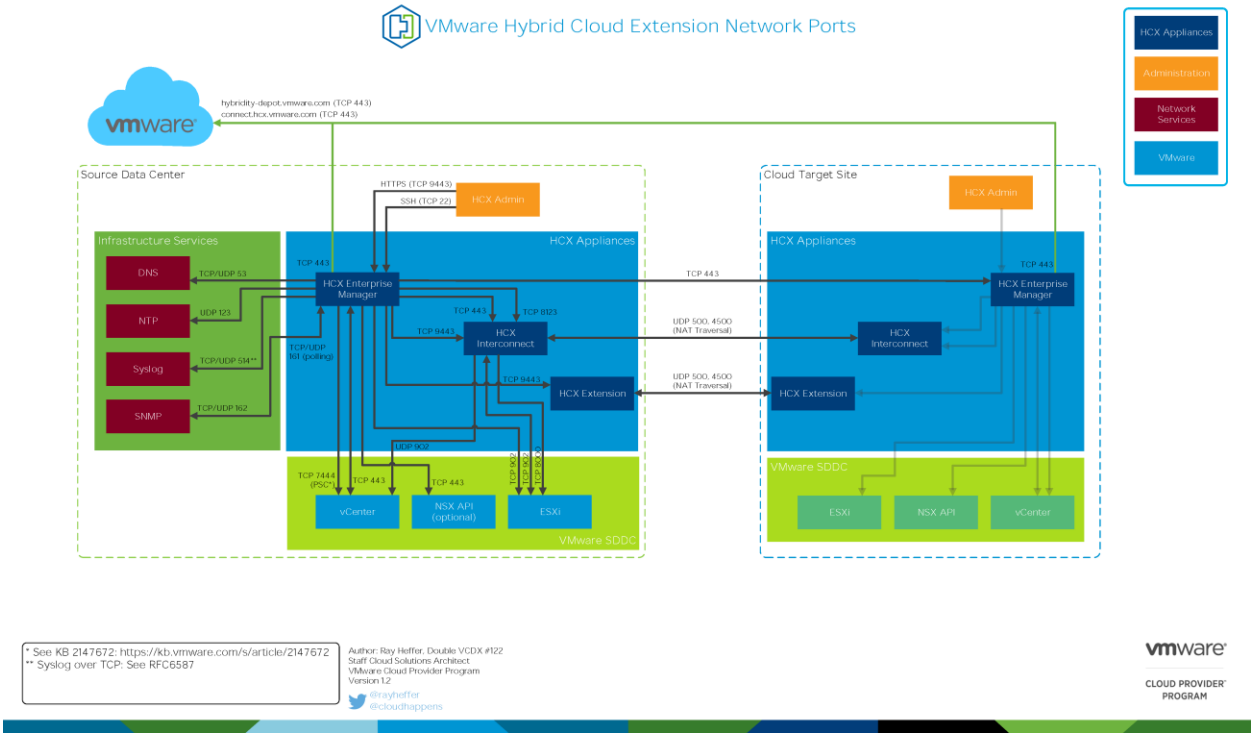

Figure 8. HCX Network Ports

The high-resolution diagram is available at: https://blogs.vmware.com/vcat/2018/05/vmware-hybrid-cloud-extension-hcx-network-ports.html

Note Do not extend the management network (where HCX management interfaces are). Management is required to be on a separate network from the stretched VLANs.