

vmware® PRESS



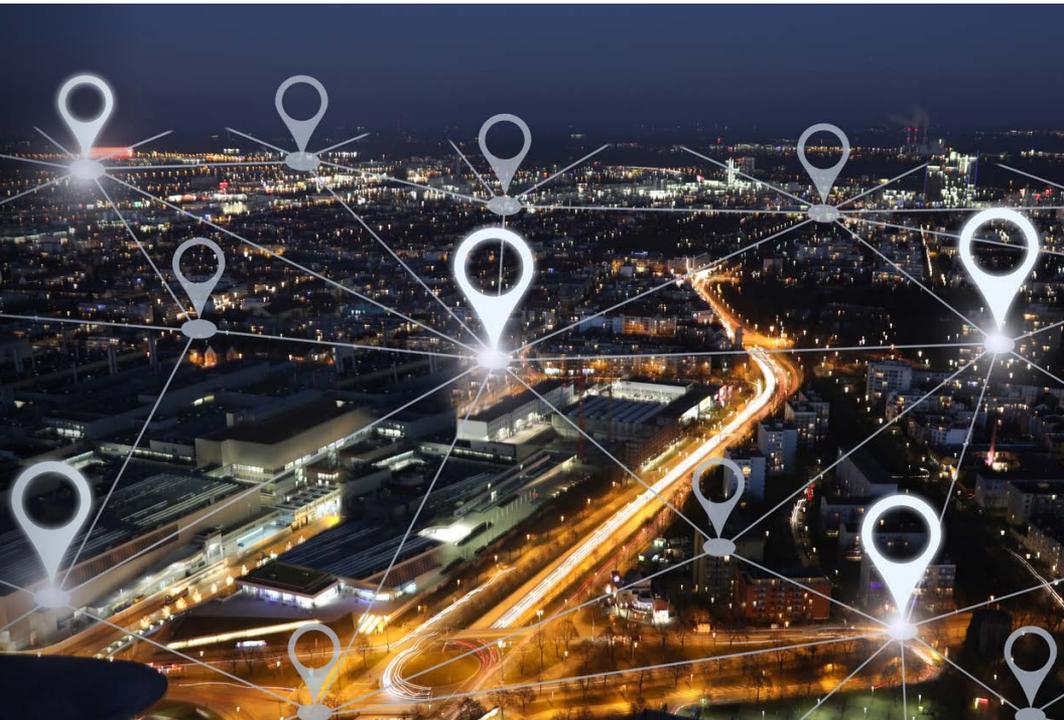
# VMware NSX® Multi-site Solutions and Cross-vCenter NSX Design

Day 1

**Humair Ahmed**

**With Contributions From Yannick Meillier**

Foreword by Justin Giardina, CTO, Iland





# **VMware NSX® Multi-site Solutions and Cross-vCenter NSX Design**

Day 1

**Humair Ahmed**  
**With Contributions From Yannick Meillier**

Foreword by Justin Giardina, CTO, Iland

## **VMWARE PRESS**

### **Program Managers**

Katie Holms  
Shinie Shaw

### **Technical Writer**

Rob Greanias

### **Reviewers and Content Contributors**

Marcos Hernandez  
Anderson Duboc  
Gustavo Santana  
Angel Villar Garea  
Andrew Voltmer  
Scott Goodman

### **Designer and Production Manager**

Michaela Loeffler  
Sappington

### **Warning & Disclaimer**

Every effort has been made to make this book as complete and as accurate as possible, but no warranty or fitness is implied. The information provided is on an “as is” basis. The authors, VMware Press, VMware, and the publisher shall have neither liability nor responsibility to any person or entity with respect to any loss or damages arising from the information contained in this book.

The opinions expressed in this book belong to the author and are not necessarily those of VMware.

**VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA  
Tel 877-486-9273 Fax 650-427-5001 [www.vmware.com](http://www.vmware.com).**

Copyright © 2018 VMware, Inc. All rights reserved. This product is protected by U.S. and international copyright and intellectual property laws. VMware products are covered by one or more patents listed at <http://www.vmware.com/go/patents>. VMware is a registered trademark or trademark of VMware, Inc. and its subsidiaries in the United States and/or other jurisdictions. All other marks and names mentioned herein may be trademarks of their respective companies.

# Table of Contents

Content Contributors.....	XIII
Additional Contributors.....	XIII
Preface.....	XV
Foreword.....	XVI
Chapter 1 - Introduction.....	1
Chapter 2 - Multi-site and Traditional Challenges .....	5
Definition of Multi-Site.....	5
Why Multi-site.....	8
Traditional Multi-site Challenges.....	9
Chapter 3 - NSX for Multi-site Data Center Solutions.....	15
About NSX.....	15
Multi-site with NSX.....	18
NSX Multi-site Solutions.....	19
1.) Multi-site with Single NSX/VC Instances and Stretched vSphere Clusters (vSphere Metro Storage Cluster).....	19
2.) Multi-site with Single NSX/VC Instances and Separate vSphere Clusters.....	21
3.) Multi-site with Cross-VC NSX.....	24
4.) Multi-site with L2VPN.....	25
NSX Multi-site Solutions Summary and Comparison.....	29
Chapter 4 - Cross-VC NSX Overview .....	33
About Cross-VC NSX.....	33
VMware Cross-VC NSX Use Cases.....	36
Use Case 1: Workload Mobility .....	36
Use Case 2: Resource Pooling .....	37
Use Case 3: Unified Logical Networking and Security Policy.....	38
Use Case 4: Disaster Recovery .....	39
Cross-VC NSX Terminology.....	41
Architecture and Key Concepts.....	42
NSX Manager Roles .....	42
Universal Controller Cluster (UCC).....	53
Components Communication in the Cross-VC NSX Architecture	54
Universal Objects Architecture Considerations .....	55
UDFW Rule Creation Example .....	58
ULS Creation Example .....	59
Controller Disconnected Operation (CDO) Mode.....	60

<b>Chapter 5 - Understanding VMware Cross-VC NSX Networking and Security .....</b>	<b>71</b>
Cross-VC NSX Switching and Routing.....	71
Universal Logical Switch (ULS).....	71
Universal Distributed Logical Router (UDLR) .....	72
Edge Services Gateway (ESG).....	74
Cross-VC NSX L2 Bridging Between Logical and Physical Network ..	75
Handling Broadcast, Unknown Unicast, Multicast (BUM) traffic.....	76
ARP Suppression.....	78
BUM Replication Modes.....	79
Unicast.....	79
Hybrid .....	80
Multicast .....	80
Cross-VC NSX Security.....	82
DFW Sections .....	84
UDFW Rule Objects.....	86
Apply To.....	95
Service Composer .....	97
Third Party Security Services .....	99
<b>Chapter 6 - Cross-VC NSX Implementation &amp; Deployment Considerations .....</b>	<b>103</b>
Physical Underlay Network Consideration.....	104
Cross-VC NSX Deployment Example and Options.....	106
Platform Services Controller (PSC).....	107
Management and UCC Component Deployment and Placement.....	110
Universal Controller Cluster (UCC).....	112
Secondary NSX Manager.....	116
UDLR Universal Control VM Deployment and Placement .....	118
Edge Component Deployment and Placement.....	120
ESG Stateful Services .....	121
Graceful Restart .....	122
Local Egress.....	123

Chapter 7 - Cross-VC NSX Deployment Models.....	125
1. Multi-site with Multiple vCenters.....	127
a. Active/Passive North/South with Routing Metric or Local Egress.....	127
Topology Details.....	127
Routing Details.....	131
Traffic Flow.....	139
b. Active/Active North/South with Local Egress.....	147
Global Server Load Balancing (GSLB).....	162
Route Injection.....	164
2. Multi-site with Single vCenter.....	164
a. Active/Passive Site Egress with Routing Metric or Local Egress.....	164
b. Active/Active Site Egress with Local Egress.....	167
Summary.....	170
Index.....	173

# List of Figures

Figure 1.1	NSX Logical Networking.....	2
Figure 1.2	NSX Provides Multiple Solutions .....	3
Figure 2.1	Traditional Multi-Site Deployment .....	6
Figure 2.2	Multi-Site Deployment with IPSEC VPN .....	7
Figure 2.3	Multi-site Deployment with L2VPN.....	7
Figure 2.4	Multi-Site Data Center Solution Using L2 Spanning Across Dark Fiber.....	9
Figure 2.5	Multi-Site Data Center Solution Using VPLS Managed Service.....	10
Figure 2.6	Multi-Site Data Center Solution Using Hardware Based Overlay (OTV) .....	12
Figure 3.1	NSX Multi-site .....	15
Figure 3.2	NSX Network Logical View .....	16
Figure 3.4	NSX With vMSC.....	20
Figure 3.5	NSX with Separate vSphere Clusters.....	22
Figure 3.6	Cross-VC NSX Deployment.....	24
Figure 3.7	NSX L2VPN For L2 Extension.....	26
Figure 3.8	NSX L2VPN Used For L2 Extension Between Sites And To Migrate Workloads.....	27
Figure 3.9	NSX L2VPN Supports A Combination Of Networks.....	28
Figure 3.10	NSX L2VPN Topologies .....	29
Figure 4.1	VMware NSX Deployment - Single Site, Single vCenter .....	34
Figure 4.2	Pre-NSX 6.2 Cross-VC NSX.....	35
Figure 4.3	VMware Cross-VC NSX Deployment - Multi-Sites, Multiple vCenters.....	36
Figure 4.4	Workload Mobility Across vCenter Domains with Consistent Networking and Security Policies for the Application.....	37
Figure 4.5	Resource Pooling And Better Utilization Of Idle Capacity Across vCenter Domains and Sites .....	38
Figure 4.6	Security Policy Automation and Consistent Security Across vCenter Domains .....	38
Figure 4.7	SRM DR Solution Leveraging Cross-VC NSX .....	40
Figure 4.8	Multi-Site NSX Deployment With Multiple vCenters.....	43
Figure 4.9	Manually Initiating Universal Synchronization Service (USS) Syncs .....	44
Figure 4.10	USS on Primary NSX Manager Replicating Universal Objects.....	45
Figure 4.11	Assigning Primary Role To Standalone NSX Manager.....	46
Figure 4.12	Confirming Primary Role Assignment To Standalone NSX Manager.....	47
Figure 4.13	Standalone NSX Manager Made Primary NSX Manager.....	47

Figure 4.14	Adding a Secondary NSX Manager .....	48
Figure 4.15	Providing Secondary NSX Manager Login Credentials.....	49
Figure 4.16	Secondary NSX Manager Successfully Registered with Primary NSX Manager .....	51
Figure 4.17	Removing a Secondary NSX Manager From Primary NSX Manager.....	51
Figure 4.18	Confirming Removal of Secondary NSX Manager.....	52
Figure 4.19	Secondary Role On NSX Manager Successfully Removed	52
Figure 4.20	Cross-vCenter NSX Component Communication.....	54
Figure 4.21	Adding a Secondary NSX Manager .....	55
Figure 4.22	USS Only Runs On Primary NSX Manager.....	56
Figure 4.23	USS is Always Stopped On Secondary NSX Manager(s) ....	57
Figure 4.24	Local And Universal Transport Zone Being Utilized .....	57
Figure 4.25	UDFW Creation Process .....	58
Figure 4.26	ULS Creation Process and LS Creation Process .....	59
Figure 4.27	Configuring Segment IDs For Local And Universal Objects.....	60
Figure 4.28	Enabling CDO Mode At Transport Zone Level.....	61
Figure 4.29	CDO Mode Enabled On Universal Transport Zone.....	62
Figure 4.30	CDO Logical Switch not Visible Under 'Logical Switches' Tab .....	62
Figure 4.31	New Universal Logical Switch Selects Next Available VNI 900001 .....	63
Figure 4.32	No CDO Mode, Controller Cluster Up, With Two VMs On The Same Host And Same Universal Logical Switch.....	64
Figure 4.33	NSX Manager Central CLI Displaying VTEP Table on Controller for VNI 900002.....	64
Figure 4.34	NSX Controller Cluster Down, Communication Between VMs On Universal Logical Switch VNI 900002 Continues To Work .....	65
Figure 4.35	NSX Controller Cluster Down, Communication Between VMs On Universal Logical Switch VNI 900002 Continues To Work .....	66
Figure 4.36	NSX Manager Central CLI Displaying VTEP Table on Controller for VNI 900002.....	66
Figure 4.37	'Host 1' Has the VTEP Entry for 'Host' 2 for VNI 900002...	67
Figure 4.38	'Host 2' has the vTEP Entry for 'Host 1' for 'VNI 900002 ...	67
Figure 4.39	NSX Controller Cluster Down and VM on Universal Logical Switch VNI 900002 on 'Host 1' vMotions to 'Host 2' with no Data Plane Disruption.....	67
Figure 4.40	NSX Controller Cluster Down and VM on Universal Logical Switch VNI 900002 on 'Host 1' vMotions to 'Host 2' Causing Data Plane Disruption.....	68

Figure 5.1	ULSs Created In Universal Transport Zone And Local LSs Created In Local Transport Zone .....	72
Figure 5.2	UDLR Being Deployed.....	73
Figure 5.3	Logical Interfaces Created on The UDLR and Providing Connectivity to ULSs.....	73
Figure 5.4	Cross-VC NSX Deployment for Disaster Recovery Solution.....	74
Figure 5.5	Cross-VC NSX Deployment with ESGs in HA Mode.....	75
Figure 5.6	NSX L2 Bridge Used For Communications Between Physical and Virtual Workloads .....	76
Figure 5.7	Creating A Universal Transport Zone And Selecting The Replication Mode.....	77
Figure 5.8	Deploying a Universal Logical Switch.....	78
Figure 5.9	Enabling Multicast Addressing.....	81
Figure 5.10	Consistent Security Policies Across vCenter Domains with Cross-VC NSX.....	82
Figure 5.11	Leveraging NSX REST API at Primary Site to Get Consistent Security Across Sites.....	83
Figure 5.12	UDFW Providing Micro-segmentation Across vCenter Boundaries.....	83
Figure 5.13	Universal Section within DFW Configuration.....	84
Figure 5.14	Adding new UDFW section .....	85
Figure 5.15	Universal Section Always On Top On Secondary NSX Managers.....	85
Figure 5.16	Application Must be Entirely at Site 1 or Site 2 to Leverage New Matching Criteria.....	87
Figure 5.17	Creating A Universal IP Set .....	88
Figure 5.18	Three Universal IP Sets Created on The Primary NSX Manager .....	88
Figure 5.19	Universal IP Sets Synced by USS To Secondary NSX Managers.....	89
Figure 5.20	Including Universal 'Web IP Set' as Part of Universal 'Web Security Group' .....	89
Figure 5.21	Three Universal Security Groups Created For Respective Universal IP Sets.....	90
Figure 5.22	Creation and Sync of Universal Security Tags .....	90
Figure 5.23	Setting Unique ID Selection Criteria on Primary NSX Manager .....	91
Figure 5.24	Flow for Utilizing Universal Security Tags in Active/Standby Deployments .....	92
Figure 5.25	Creating Universal Security Group Which Can Match on UST and VM Name.....	92
Figure 5.26	Within USG, Selecting VM Name for Matching Criteria.....	93

Figure 5.27	Within USG, Selecting Universal Security Tag for Matching Criteria.....	93
Figure 5.28	USG with Matching Criteria of UST Being Used in a Universal Security Policy.....	94
Figure 5.29	Entire Application Must be at Same Site When Using UDFW Rules Leveraging USGs.....	95
Figure 5.30	Leveraging Universal Security Group with ApplyTo in UDFW.....	96
Figure 5.31	Including ULS in Local Security Group.....	97
Figure 5.32	DR Solution Leveraging Cross-VC NSX with Local Security Policies.....	98
Figure 5.33	Cross-VC NSX Deployment Using Palo Alto Networks Security with Separate Panoramas at each Site.....	99
Figure 5.34	Cross-VC NSX Deployment Using Palo Alto Networks Security with Separate Panoramas at each Site.....	100
Figure 6.1	Physical Network Becomes Underlay Transport for Logical Networking.....	105
Figure 6.2	VMkernel interface for Transport Network.....	105
Figure 6.3	Example Cross-VC NSX setup.....	107
Figure 6.4	Installing PSC From the vCenter Server Appliance ISO.....	108
Figure 6.5	Installing vCenter From the vCenter Server Appliance ISO.....	108
Figure 6.6	Connecting vCenter to PSC.....	109
Figure 6.7	Enhanced Link Mode Allows for Central Management of Multiple vCenter Domains.....	110
Figure 6.8	Assigning Primary Role to NSX Manager.....	111
Figure 6.9	NSX USS Running on Primary NSX Manager.....	111
Figure 6.10	NSX Manager Promoted to Primary Role.....	112
Figure 6.11	Deploying A Controller.....	113
Figure 6.12	Management vCenter Managing all vCenters and Respective NSX Managers.....	114
Figure 6.13	Separate Management vCenter For Multi-site Deployment with Cross-VC NSX.....	114
Figure 6.14	vCenter and NSX Manager installed on Mgmt Cluster It's Managing.....	115
Figure 6.15	Universal Controller Cluster Deployed at Site 1.....	115
Figure 6.16	Secondary NSX Manager Registered with the Primary NSX Manager.....	116
Figure 6.17	Secondary NSX Manager Successfully Registered with Primary NSX Manager.....	117
Figure 6.18	Secondary NSX Manager with Successful Connectivity to UCC.....	118
Figure 6.19	UDLR ON Primary NSX Manager with Universal Control VM Deployed.....	119

Figure 6.20	ESG Deployed Across Two Sites for North South Traffic Resiliency .....	120
Figure 6.21	Multiple ESGs in ECMP Mode Deployed Across Two Sites .....	121
Figure 6.22	Multi-site, Multi-vCenter Deployment With Active/Passive N/S and One-Arm Load Balancer.....	122
Figure 7.1	Cross-VC NSX Providing Different Deployment Models for Multiple Tenants .....	127
Figure 7.2	Multi-site with Multiple vCenters & Active/Passive Site Egress.....	128
Figure 7.3	Multiple ESGs in ECMP Mode Deployed Across Two Sites .....	129
Figure 7.4	HA ESGs with Stateful Services Deployed At Two Sites....	130
Figure 7.5	NSX Multi-Site Cross-VC Setup With BGP .....	131
Figure 7.6	BGP Weight Attribute Used to Prefer Routes to ESG 1 at Site 1.....	132
Figure 7.7	Traceroute From VM on Web Universal Logical Switch at Site 1 To Physical Workload Confirms ESG At Site 1 is Being Used for Egress.....	133
Figure 7.8	Traceroute From Physical Workload to VM on Web Universal Logical Switch at Site 1 shows ESG 2 Is Being Used.....	135
Figure 7.9	NSX Logical Network Summary Routes Redistributed at Site 1 ESG .....	135
Figure 7.10	NSX Logical Network Summary Routes Not Redistributed at Site 2 ESG.....	136
Figure 7.11	Summary routes for NSX permitted in the prefix list at Site 1 ESG Interface 1.....	137
Figure 7.12	Summary routes for NSX permitted in the prefix list at Site 1 ESG Interface 2 .....	137
Figure 7.13	Summary routes for NSX denied in the prefix list at Site 2 ESG Interface 1.....	138
Figure 7.14	Summary routes for NSX denied in the prefix list at Site 2 ESG Interface 2 .....	138
Figure 7.15	Cross-VC NSX setup - Multi-vCenter, Multi-Site, Active/Passive Egress.....	140
Figure 7.16	Universal Control VM Informs UCC of Best Forwarding Paths .....	141
Figure 7.17	UCC Distributing Forwarding Information to ESXi hosts across All Sites.....	142
Figure 7.18	ESXi Hosts Across All Sites Use Site 1 Egress.....	143
Figure 7.19	Site 2 ESGs Used For Site 2 Egress Upon Site 1 ESG/Upstream Connectivity Failure.....	144

Figure 7.20	Cross-VC NSX Deployment with Multiple UDLRs – ESGs at both Sites Being Utilized.....	145
Figure 7.21	Cross-VC NSX Deployment for Bi-directional DR.....	146
Figure 7.22	Enabling Local Egress Upon UDLR Creation.....	147
Figure 7.23	Universal Control VMs Learns Routes From Site-Specific ESGs.....	149
Figure 7.24	Universal Control VMs Sends Best Forwarding Paths with Associated locale ID to UCC.....	150
Figure 7.25	UCC Uses locale ID to Distribute Forwarding Information to ESXi Hosts with Matching locale ID .....	151
Figure 7.26	Site-specific Active/Active North/South Egress Enabled by Local Egress .....	152
Figure 7.27	Changing locale ID At The Cluster Level.....	153
Figure 7.28	Default locale ID Inherited From Local NSX Manager .....	153
Figure 7.29	Upon Partial Application Failover, Site 1 N/S Egress Still Used.....	154
Figure 7.30	Upon Full Application Failover or Edge Failure, Site 2 N/S Egress Used.....	155
Figure 7.31	Changing locale ID At The UDLR Level.....	156
Figure 7.32	Changing locale ID At The Static Route Level .....	157
Figure 7.33	Multi-site with Multiple vCenters & Active/Active Site Egress.....	158
Figure 7.34	Deploying Universal Control VM at Primary Site .....	159
Figure 7.35	UDLR With Status of Deployed at Primary Site .....	160
Figure 7.36	Local Egress Configured But Universal Control VM Not Yet Deployed at Secondary Site .....	160
Figure 7.37	Universal Control VM not Deployed at Secondary Site .....	160
Figure 7.38	Local Egress Configured And Universal Control VM Deployed at Secondary Site .....	161
Figure 7.39	Cross-VC NSX Active/Active Deployment Leveraging F5 Networks GSLB .....	163
Figure 7.40	Load Balancing Methods available via F5 Networks GSLB .....	163
Figure 7.41	Multi-Site Solution with One vCenter and No Universal Objects.....	165
Figure 7.42	vMSC Multi-Site Solution with One vCenter and No Universal Objects.....	166
Figure 7.43	Multi-site Solution With One vCenter and Local Egress With Static Routes.....	168

## List of Tables

Table 3.1	NSX Multi-site L2 Extension Option Comparaison .....	30
-----------	--	----

# About the Authors



**Humair Ahmed**  
**Senior Technical Product Manager**  
**VMware Inc.**

Humair Ahmed is a Senior Technical Product Manager in VMware's Network & Security Business Unit (NSBU). Humair has expertise in network architecture/design, multi-site and cloud solutions, disaster recovery, security, and automation. Humair is currently focused on cloud and end-to-end multi-site solutions that enable workload mobility, resource pooling, consistent multi-site security, and disaster recovery.

Humair holds multiple certifications in development, systems, networking, virtualization, and cloud computing and has over 16 years of experience across networking, systems, and development. Previously at Force10 Networks and Dell Networking, Humair specialized in automation, data center networking, and software defined networking (SDN). He has designed many reference architectures for Dell's new products and solutions, including Dell's first reference architecture with VMware NSX®.

Humair has authored many white papers, reference architectures, deployment guides, training materials, and technical/marketing videos while also speaking at industry events like VMworld and participating in think tanks.

In his spare time, Humair writes on the VMware Network Virtualization Blog and authors a popular technology blog - <http://www.humairahmed.com> - focused on networking, systems, and development.

You can contact Humair on Twitter @Humair\_Ahmed.

# Content Contributors



**Yannick Meillier**  
**NSX Solutions Architect**  
**VMware Inc.**

Yannick Meillier is an NSX Solutions Architect in the VMware Customer Success Team helping customers and partners with NSX designs, architectures, and implementations. Yannick joined VMware in 2012 and has been working on VMware virtual networking solutions since then starting with vShield 5.0, vCloud Networking and Security, and now NSX. Yannick holds a PhD in Atmospheric Physics with a background in electrical engineering and in numerical weather forecasting modeling.

# Additional Contributors

**Jonathan Morin**  
**Ray Budavari**

# Acknowledgements

I would like to thank my family and friends for supporting me throughout my career, across many endeavors, and during the process of writing this book. I also want to give a big thank you to reviewers Yannick Meillier and Jonathan Morin; thank you for putting in the time to look over the details and provide feedback.

Thank you to all contributors and folks who provided feedback from the VMware Network and Security Business Unit (NSBU).

Additionally, I would also like to thank all the VMware NSX customers out there, many of whom I have learned a great deal from and who have helped provide valuable feedback throughout the years as NSX evolved.

Finally, thank you to Katie Holms and Shinnie Shaw who helped get this project started and kept on-track. Thank You!

A handwritten signature in black ink that reads "Humair Ahmed". The signature is written in a cursive style with a long, sweeping underline that extends to the right.

Humair Ahmed

# Preface

Traditional solutions have been ineffective at addressing many challenges faced when deploying multi-site data center architectures. *NSX Multi-site Solutions and Cross-vCenter NSX Design Day 1* details both challenges and solutions, discussing how VMware NSX® provides a better holistic solution for multi-site. Specifically, this book focuses on multi-site solutions with VMware NSX for vSphere.

No longer are logical networking and security constrained to a single VMware vCenter® domain; Cross-vCenter NSX enables logical networking and security across multiple vCenter domains/sites and provides enhanced solutions for specific use cases.

*NSX Multi-site Solutions and Cross-vCenter NSX Design Day 1* opens with an outline of several NSX solutions available for multi-site data center connectivity, then digs deeper into the details of the Cross-vCenter NSX multi-site solution. Cross-vCenter NSX use cases, architecture, functionality, deployment models, design, and failure/recovery scenarios are discussed in detail.

This document is targeted toward virtualization and network architects interested in using VMware NSX® network virtualization solution in a VMware vSphere® environment to provide multi-site solutions.

# Foreword

The network components of the data center have been in a constant flux over the last decade. Starting my career in the 90's and primarily focusing on networking in the beginning of the 21st century, I witnessed the introduction of numerous network technologies within the data center. Early on, we were constantly navigating the maze of fault tolerance and segmentation through various hardware-based solutions. As time went on and the computing power of network hardware increased, we started to see the dawn of segmenting services within the hardware's network operating system and also allowing features such as virtual contexts to be introduced.

I began my career with iland back in 2008. At that time, we began to migrate most of our physical footprints over to virtual, leveraging ESX 3.5 and vCenter 2.5, creating our first Infrastructure as a Service (IaaS) cloud. Back then, we did not have the luxury of products such as NSX and software-defined networking. During this time, we also launched our first Disaster Recovery as a Service (DRaaS) offering. Although we had VMware Infrastructure 3 at the time, we were still limited by hardware networking solutions.

Starting in 2011, we began our first journey into software-defined networking with VMware's vShield technology. Using a component of vShield called VCDNI, it allowed us to create virtual network segments using mac-in-mac encapsulation technology. VCDNI allowed us to scale to millions of tenant networks, overcoming issues such as VLAN exhaustion. At that time, we also utilized the vShield Edge technology to not only reduce the requirement for hardware based firewalls, but also better segment customer environments and do this all via an API, so deployments could be done programmatically and in a repeatable nature.

Since 2011, we have continually followed the software-defined networking path with VMware. From vShield, we migrated to the vCloud Networking and Security product to enhance our offerings and also move to the VXLAN standard. Over time, as more features were introduced in the product and the API, we continued to migrate legacy hardware based services like SSL offload and routing to the vCloud Networking and Security stack.

In 2015, we then began to migrate our vCloud Networking and Security installations over to NSX. With NSX, leveraging the in-kernel capabilities like distributed routing and distributed firewall, we are able to handle even the most demanding applications. Further, the network services available on the NSX Edge allowed us to offer additional capabilities per tenant such as NAT, load balancing, edge firewall, and VPN. Following the natural progression of features, we were also able to continue using the API to allow more functionality to our customers, like SSL offload and stretched networks to on-prem via L2VPN.

Starting with the Net-X API in vCloud Networking and Security and now all of the great features in NSX, iland has been able to develop an industry leading cloud solution to provide complete security and multi-site solutions such as DRaaS and IaaS for our customer base. Being able to programmatically integrate features like logical networks, distributed routing, micro segmentation, and third-party security ecosystem solutions, we successfully created a secure multi-site cloud offering. Previously, having to do this with hardware-based solutions and little or no API support was cumbersome and not flexible.

Once we began working with the NSX team on the Cross-vCenter functionality, we were very pleased to see that customers could have multiple options and flexibility in deploying end-to-end solutions for multi-site and hybrid cloud. The dream of having an integrated networking solution that not only seamlessly bridged multiple data centers but also maintained policy and security posture across both had finally become a reality.

I am very excited to see this book come to fruition and share the experience we had working with Humair. It is a joy to see this technology in action and the industry's reaction to it; we are proud and grateful to be able to be part of it.

**Justin Giardina, CTO**  
**iland Internet Solutions**  
<https://iland.com>

Justin Giardina is the Chief Technology Officer and oversees all aspects of iland's global technical operations and strategy including design, implementation, and support. Under Justin's leadership, iland has established a global cloud infrastructure footprint and been first to market with innovative public, private and hybrid cloud solutions for production applications, virtual desktop and lab environments as well as business continuity and disaster recovery.

With more than 20 years' experience in datacenter and network operations, Justin speaks regularly on several topics including security, network and server virtualization, resource optimization and performance. He is a member of the Cisco and VMware Partner Technical Advisory Boards, elite groups of technical experts that provide new ideas and constructive feedback to help develop product and service offerings that better meet the needs of customers and partners. Justin also volunteers in his spare time as a systems administrator for several Open Source Community projects. Prior to joining iland, Justin led network engineering and system administration teams for companies in the consulting and petrochemical industries.





# Introduction

VMware NSX is a network virtualization technology that decouples networking services from the underlying physical infrastructure. In doing so, NSX enables a new software-based approach to networking that provides the same operational model that virtual machines (VMs) brought to compute virtualization. Virtual network topologies can now be easily created, modified, backed-up, and deleted. By replicating the physical networking constructs in software, VMware NSX provides similar benefits to what server virtualization did with VMs: increased efficiency, productivity, flexibility, agility, and cost savings.

NSX's software-based approach for networking and security addresses several problems of traditional solutions, including challenges with security, automation, and application continuity.

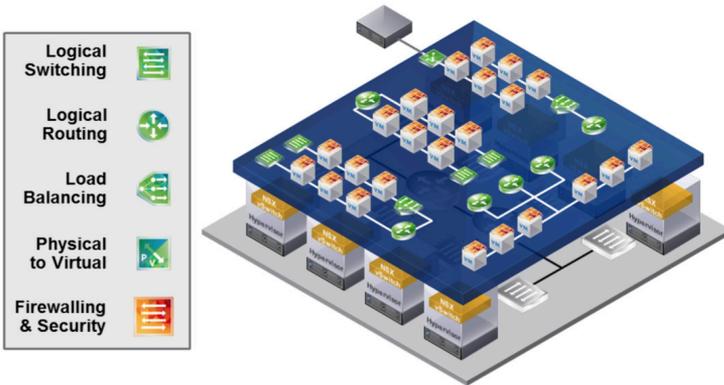
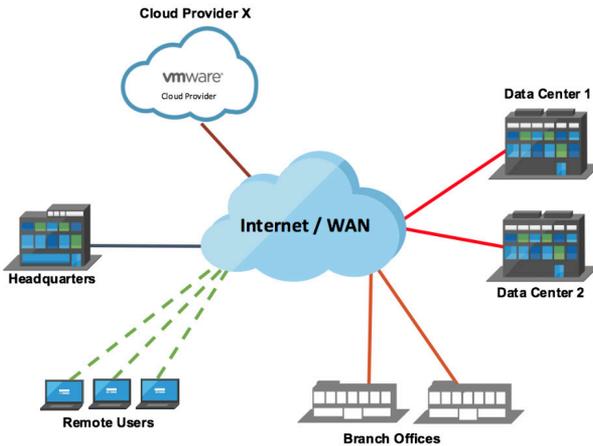


Figure 1.1 NSX Logical Networking

This book focuses on multi-site data center solutions with NSX. Companies have many different connectivity objectives and requirements for their multi-site solutions. An organization may need to establish connectivity between data centers, to cloud environments provided by a VMware Cloud Provider, to branch offices, or simply to remote users.

VMware NSX provides solutions for all of these requirements:

1. Multi-Data Center Connectivity: single-VC multi-site with and without stretched clusters, multi-VC single-site and multi-site with cross-VC NSX, IPSEC VPN, L2VPN
2. Public Cloud: IPSEC VPN, L2VPN, Cross-VC NSX
3. Remote User Access: SSL VPN
4. Branch Offices: IPSEC VPN, L2VPN



**Figure 1.2** NSX Provides Multiple Solutions

The proper deployment solution depends on several factors including bandwidth and latency between sites, MTU requirements, shared versus dedicated administrative domains, etc.

This book focuses on multi-site data center solutions providing L2 extension over L3 and multi-site security allowing for extended workload mobility. These solutions include the enablement of active/active data centers as well as active/standby deployment models for use cases like disaster avoidance (DA) and disaster recovery (DR).

Although multiple options for multi-site deployments with NSX are discussed and compared, this book specifically focuses on a feature called Cross-vCenter NSX. Cross-VC NSX allows the extension of logical networking and security across vCenter domains and/or sites. This book starts by outlining several NSX solutions available for multi-site data center connectivity and then digs deeper into the technical details of multi-site with Cross-VC NSX. The reader should have a fundamental understanding of VMware NSX as the solutions discussed herein are an advanced topic and share the same foundational design principles.

Before looking at how NSX provides robust multi-site solutions, it's important to understand where the need for multi-site topologies comes from, the challenges faced by traditional multi-site solutions, and how NSX addresses these issues.

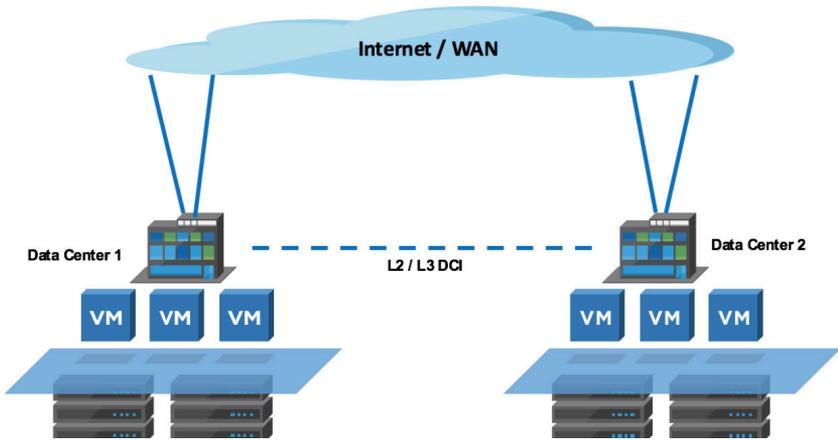


# Multi-site and Traditional Challenges

## Definition of Multi-Site

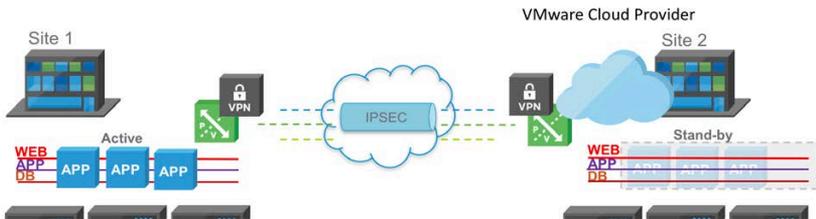
As a first step, it's important to arrive at a common definition for the term multi-site. In the context of this book, it specifically refers to multiple data centers. In the traditional sense, multi-site data center solutions consist of two or more data centers connected via an L2 or L3 data center interconnect (DCI). This DCI can be owned by the customer or leased from a service provider. It can be a dedicated connection, or shared connection where multiple customers are running secure traffic over the same network. DCI connectivity may consist of L2/L3 over dedicated dark fiber, MPLS, L2/L3 VPNs, or other technologies discussed in the upcoming sections.

These data centers also have their own connectivity to the Internet/WAN. Multiple different configurations are possible, ranging from basic single-homed connections to dual multi-homed architectures that provide additional resiliency for north/south traffic at each site.

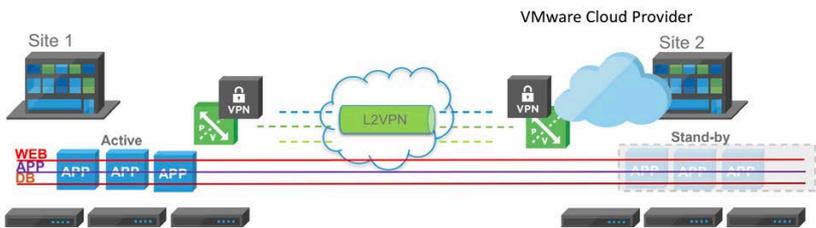


**Figure 2.1** Traditional Multi-Site Deployment

If one of the sites is an environment provided by a VMware Cloud Provider, it is also possible that the connectivity between data centers would simply be provided by L2 or L3 VPN tunnels over the Internet. Examples of these environments are shown in Figures 2.2 and 2.3.



**Figure 2.2** Multi-Site Deployment with IPSEC VPN



**Figure 2.3** Multi-site Deployment with L2VPN

In such multi-site deployments there can be active workloads at both sites to fully utilize resources at both data centers. Another option would have workloads active at only one site with standby workloads at the other site; this is commonly used for DR purposes.

The next section looks at the reasons why customers and organizations desire multi-site solutions and the challenges they are confronted with.

# Why Multi-site

Some reasons an organization may have multiple data center and/or sites interconnected via a multi-site solution include:

- Mobility of workloads between data centers may be desired/required
- Disaster avoidance/disaster recovery may be desired/required
- Organization growth
- A merger between organizations may result in multiple data center sites
- Multiple data centers in different geographical locations to serve local offices/customers
- Migration to a new data center
- Resource pooling not only due to organizational growth and mergers but also for compliance reasons, business requirements, and because of idle capacity being available at specific data center sites

Any of these reasons could lead to an organization to either have a multi-site data-center solution in place (e.g., brownfield) or to plan for the deployment of one (e.g., greenfield). In most cases, the expectation is to have active workloads across multiple sites to fully utilize resources at all locations. An exception may be if a site is used solely for disaster recovery. However, in most cases, even the disaster recovery site will be running active workloads to better utilize idle resources as opposed to being leveraged purely as a cold site.



When active workloads are running across both data center sites in a multi-site data center solution, it is called “active/active multi-site.”



When active workloads are running at one site and are standby at another, it is called “active/standby multi-site.” It is also possible to use the active/protected site as the recovery/standby site for applications active at another site. For example, an application could be active at site A and standby at site B, while another application could be active at site B and standby at site A. This kind of deployment is called “bi-directional DR.”

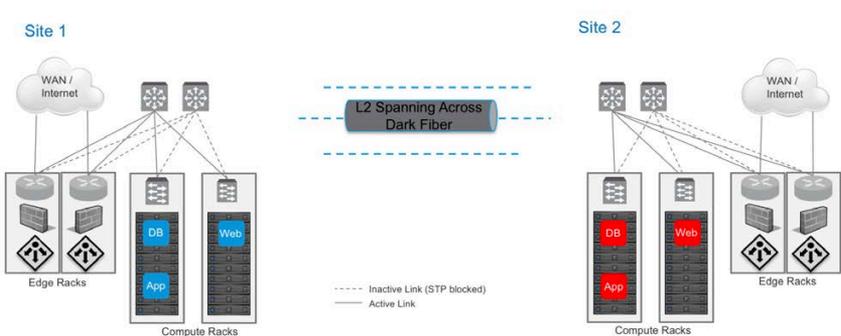
# Traditional Multi-site Challenges

Traditional multi-site challenges to be addressed when moving an application or workload across sites for purposes of workload mobility, disaster avoidance/disaster recovery, or resource pooling scenarios, consist of:

- Changing the application's IP address.
- Reconfiguring the physical network to accommodate the L2 and L3 requirements of the application.
- Replicating the security policies applied to the applications which would either need to be recreated at the secondary site or manually synced. Security policies would also need to be updated with the new applications' IP addresses.
- Performing configuration updates on physical devices (e.g., load balancer)
- Implementing additional updates and reconfigurations for any ACLs, DNS, application IP dependencies, etc.

To overcome such challenges and provide workload mobility while maintaining the respective applications' IP addresses and security policies, traditional solutions have simply focused on extending L2 across data centers. Such solutions have typically been network focused and failed to provide a holistic solution that efficiently addresses customer challenges and concerns. Some of these traditional solutions and their shortcomings are detailed below.

## 1.) L2 Spanning Across Dark Fiber



**Figure 2.4** Multi-Site Data Center Solution Using L2 Spanning Across Dark Fiber

Issues with spanning VLANs across sites as shown in this solution include:

- Increasing failure domain between sites by spanning physical L2/VLANs
- Introducing bridging loops, causing instability across both sites
- Allowing broadcast storms to propagate across all sites, causing a large-scale global outage
- Lack of agility or scalability – newly provisioned tenant workloads and applications will require additional network configurations while resources will continue to be limited to 4096 VLANs
- Management of Spanning Tree Protocol (STP), creating additional overhead and complexity
- Risk of STP misconfiguration causing convergence-related instabilities
- Inefficiencies related to STP blocking ports in L2 fabrics – loss of bandwidth and load balancing of traffic across redundant paths. The result is an underutilized network with 50% wasted bandwidth capacity. Alternatives to avoid STP involve complex proprietary protocols that are hardware dependent.
- Operationally challenging and hard to maintain
- Focused only on network and L2 extension
- Per device configuration with lack of automation/flexibility

## 2.) Managed Service: Virtual Private LAN Service (VPLS)

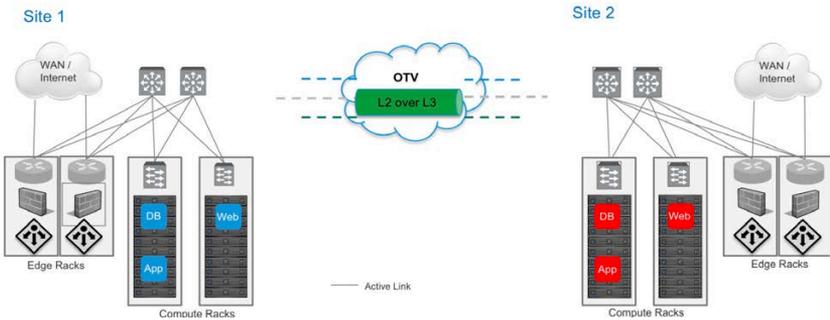


**Figure 2.5** Multi-Site Data Center Solution Using VPLS Managed Service

Although a VPLS managed service provided by an ISP can provide multipoint L2 connectivity across L3 and alleviate some of the issues tied to spanning L2 over dedicated fiber, other issues arise:

- Expensive (capital and operational)
- A VPLS bridge does not require STP to prevent loops, however, local STP misconfiguration/re-convergence can still cause instability at each site.
- Underutilized networks at each site with 50% wasted bandwidth and local STP overhead and complexity. Inefficiencies due to STP blocked ports locally - loss of bandwidth and load balancing of traffic across redundant paths.
- Bridging loops can still occur locally and cause instability.
- Alternatives to avoid STP locally involve complex proprietary protocols that are hardware dependent.
- A typical VPLS deployment needs to learn customer MACs within all service provider MPLS routers (PEs) which can cause MAC explosion issues. To avoid this PBB (provider backbone bridging), a MAC-in-MAC encapsulation, is used; however, this creates additional complexity in the service deployment.
- Not agile or scalable - lead times can take several months. New workloads/applications will require additional network configuration. Scalability is still limited to 4096 VLANs.
- Operationally challenging and hard to maintain
- Focused only on network and L2 extension
- Per device configuration with lack of automation/flexibility

### 3.) Hardware Based Overlay Approach (e.g., OTV)



**Figure 2.6** Multi-Site Data Center Solution Using Hardware Based Overlay (OTV)

Issues with hardware based overlay approaches like Overlay Transport Virtualization (OTV) include:

- Expensive (capital and operational)
- Hardware dependent requiring additional licenses and possible replacement of existing switches if not supported on specific platforms
- Proprietary and complex
- Operationally challenging and hard to maintain
- Focused only on network and L2 extension
- Per device configuration with lack of automation/flexibility
- Problematic for extending to new data centers acquired through mergers that do not run the required hardware





# NSX for Multi-site Data Center Solutions

## About NSX

VMware NSX network virtualization provides a full set of logical network and security services decoupled from the underlying physical infrastructure. Distributed functions such as switching, routing, and firewalling not only provide L2 extension across sites, but also enhanced distributed networking and security functions.



**Figure 3.1** NSX Multi-site

NSX's distributed networking and security functions are provided as kernel-level modules for the VMware ESXi™ hypervisor. Other services such as NAT, VPN, load balancing, and DHCP server/relay are centralized services delivered via NSX Edge Services Gateway (ESG) appliances. This book does not cover the basics of NSX but rather discusses the advanced use cases and capabilities of NSX for multi-site solutions. Multi-site is an advanced topic; this book discusses how NSX simplifies the implementation of such deployments and reviews details of its enhanced multi-site support capabilities.

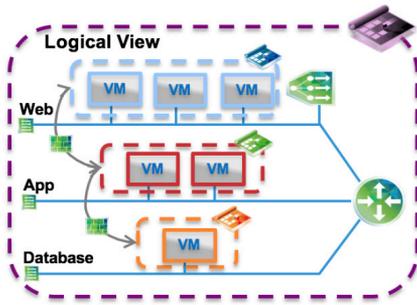


Figure 3.2 NSX Network Logical View

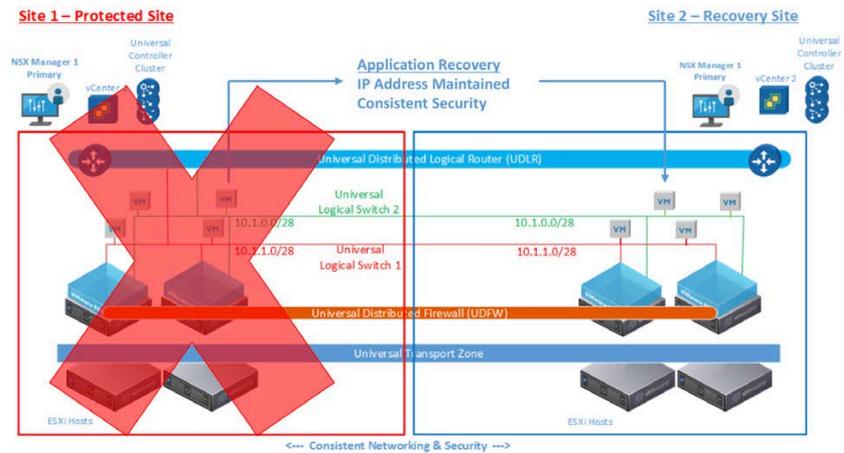
NSX’s inherent ability to extend L2 over L3 network boundaries through its use of the standards-based VXLAN overlay protocol can also be applied across data centers. Thanks to NSX, there is no longer a need to span physical L2 segments and manually replicate security policies across data centers to achieve workload mobility or provide disaster recovery services.

Additionally, the following desirable characteristics are achieved:

- Consistent logical networking and security across sites
- Presentation of the same IP address space across sites
- Enhanced security with micro-segmentation across sites (e.g., security policies at the vNIC-level)
- Ability to use higher-level security constructs like security tags and VM names which provide more context than traditional IP address based policies
- Works on top of any IP network
- Software-based, hardware-agnostic solution that can be deployed on top of any existing network infrastructure
- Flexibility in types of deployment

- Inherent automation capabilities (e.g., workloads automatically identified and placed in correct security groups with automatic enforcement of correct security policies)
- As VXLAN is a standards-based overlay, different vendors/organizations can provide rich integration with third party products
- Integration with other VMware software defined data center (SDDC) components and solutions (e.g., vSphere, the VMware vRealize® Automation™ Suite, OpenStack, Horizon View)

By providing consistent logical networking and security across sites via logical switches and distributed firewalls, VMware NSX allows application IP addresses to remain the same as workloads are migrated to or recovered at another site. The security/firewall policies for these applications are also centrally managed, ensuring consistency across sites with no need to manually sync security policies when an application is migrated. This concept is illustrated in Figure 3.3.



**Figure 3.3** Application recovered at Site 2 maintains IP address and Security Policies are Consistent Across Sites so no need to manually Sync security Policies

With these capabilities available, NSX provides enhanced solutions for DR and multi-site deployments. Working from this foundation, the remainder of this chapter examines the details of implementing a multi-site data center solution with NSX.

## Multi-site with NSX

NSX offers many options for multi-site data center connectivity to allow workload mobility or disaster recovery across sites. These solutions include L2VPN-based site-to-site connectivity, stretching of logical networks across multiple sites by a single NSX instance, and stretching of logical networks across vCenter domains within a single site or across multiple sites (Cross-VC NSX).

This book details four solutions that provide active/active or active/standby multi-site data center connectivity, each with multiple deployment options. These solutions focus on providing consistent networking and security services across data centers. They employ L2 extension over L3 to enable use cases for workload mobility, resource pooling, disaster avoidance, full/partial application failure, and disaster recovery.

Each solution can be further deployed in either an active/active or active/standby ingress/egress model. The connectivity into and out of the SDDC can route all north/south traffic through one data center (i.e., active/passive) or through the data center where a given workload resides (i.e., active/active). There are scenarios where one deployment model makes more sense than the other, which will be discussed later in the book.

One great benefit of NSX is the flexibility it provides; with everything done in software; hardware dependencies are removed and there is no lock-in to a specific topology. It is easy to move things around, change connectivity, deploy additional devices, and enable additional features. This book will discuss how the same application can span multiple data centers and operate in an active/active ingress/egress configuration. This type of deployment is sometimes desired for optimized traffic flows from clients to web services. It is also useful for high availability cases where the same application is active at both sites and can handle a complete site failure with little or no impact to services.

The four deployment options mentioned in this book consist of:

1. Multi-site with single NSX/VC instances and stretched vSphere clusters (vSphere Metro Storage Cluster)
2. Multi-site with single NSX/VC instances and non-stretched clusters
3. Multi-site with Cross-VC NSX (multiple VC domains and NSX instances)
4. Multi-site with L2VPN

This book focuses specifically on option three - **Multi-site with Cross-VC NSX**. It goes into detail on the architecture, key concepts, and deployment models for Cross-VC NSX. For reference and comparison purposes, each NSX multi-site solution providing L2 extension across sites is also covered with specific details about implementation, benefits, and requirements that may dictate a specific solution deployment.

## NSX Multi-site Solutions

### 1.) Multi-site with Single NSX/VC Instances and Stretched vSphere Clusters (vSphere Metro Storage Cluster)

Any time vSphere clusters are stretched across sites, the configuration is called vSphere Metro Storage Cluster (vMSC). Since the vSphere clusters are stretched across sites, the same datastores must be present and accessible at both sites, mandating a storage replication solution such as EMC VPLEX.

A vMSC deployment does not require NSX, but deploying NSX with vMSC provides all the benefits of logical networking and security: flexibility; automation; consistent networking across sites without spanning physical VLANs; consistent security across sites; and micro-segmentation across sites. The connectivity between data centers can be entirely routed thanks to VXLAN permitting the extension of L2 over L3.

Figure 3.4 shows NSX deployed with vMSC.

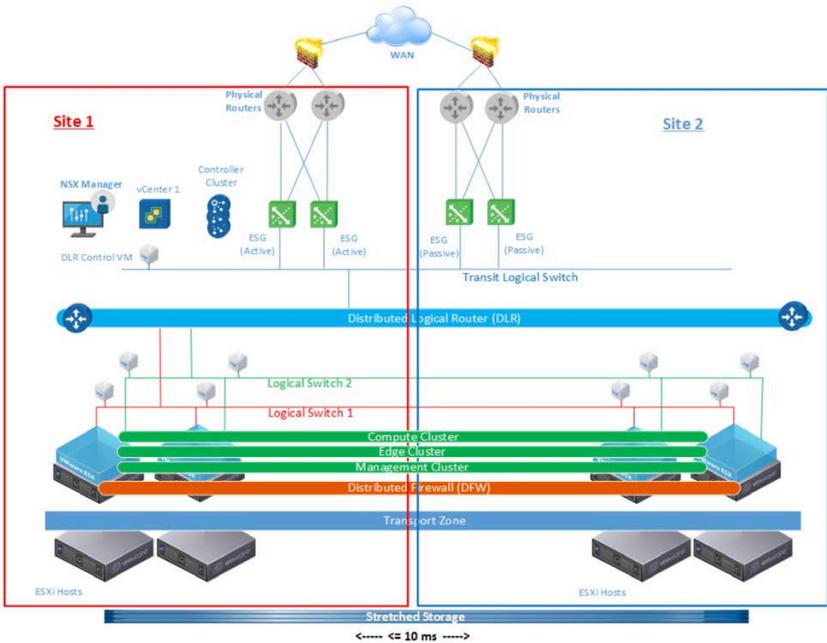


Figure 3.4 NSX With vMSC

The following items are of special interest in the NSX with vMSC deployment model:

- A vMSC-based solution has a single vCenter managing the management, Edge, and compute clusters that are stretched across both sites
- A stretched storage solution (e.g., EMC VPLEX) is required for vMSC so that the same datastore can be presented to and available at both sites
- In vSphere 6, the maximum latency requirement for vMotion between sites is 150 ms RTT. In a vMSC deployment, the latency requirement between sites is dictated by the storage vendor. In the case of EMC VPLEX, the requirement is 10 ms RTT maximum latency
- With vMSC deployments, the distance requirements are within a metro area - typically within the same city or the same building - due to synchronous storage replication latency requirements

- Since the vSphere clusters are stretched across sites, the native vSphere high availability (HA) and Distributed Resource Scheduling (DRS) features for providing high availability and resource balancing within clusters can be leveraged across sites
- This solution inherently provides DA through vSphere DRS and basic DR through vSphere HA
- All management and control plane components (e.g., vCenter, VMware NSX® Manager™, VMware NSX® Controller™ cluster) normally run at one site and can be restarted at the other through vSphere HA. Having a set of three NSX Controllers spread across sites is ill-advised as it adds no further resiliency benefits and risks split-brain problems. Although possible as a solution, it is not necessary to span physical L2 for management. Upon failover, the same L2 segment can be made available at the failover site with additional orchestration (e.g., advertising the network segment)
- Logical networking and security across sites via NSX provides VM mobility and micro-segmentation without the need to span physical L2 VLANs
- An MTU of 1600 for VXLAN is required for the network connectivity between sites
- Since vSphere clusters, logical networking, and security span across sites, there must be a common administrative domain

Figure 3.4 shows NSX Edge Services Gateways (ESGs) deployed at both sites. ESGs at site 1 are used for egress for both sites, while ESGs at site 2 are passive for egress traffic. This active/passive egress functionality is controlled by setting a preferred routing metric for site 1 egress routes. It is also possible to deploy an active/active egress solution using universal objects with the local egress feature. In this case, only static routing is supported in a single vCenter design. In both cases, active workloads can reside at each site. The active/passive egress model is preferred for simplicity and to avoid asymmetric traffic flows. The ESGs at a single site can also be deployed in HA mode if stateful services are required.

## 2.) Multi-site with Single NSX/VC Instances and Separate vSphere Clusters

In this solution, vSphere clusters are not stretched across sites. Instead, they are local to each site, so stretched storage is not required and the maximum latency requirement between sites falls back to that of vMotion – 150 ms RTT for vSphere 6.0. This latency requirement is in line with the latency requirement of the NSX control plane which is also 150 ms.

Figure 3.5 shows an NSX deployment with separate vSphere clusters. Each site has local storage, and the maximum latency requirement between sites is 150 ms RTT.

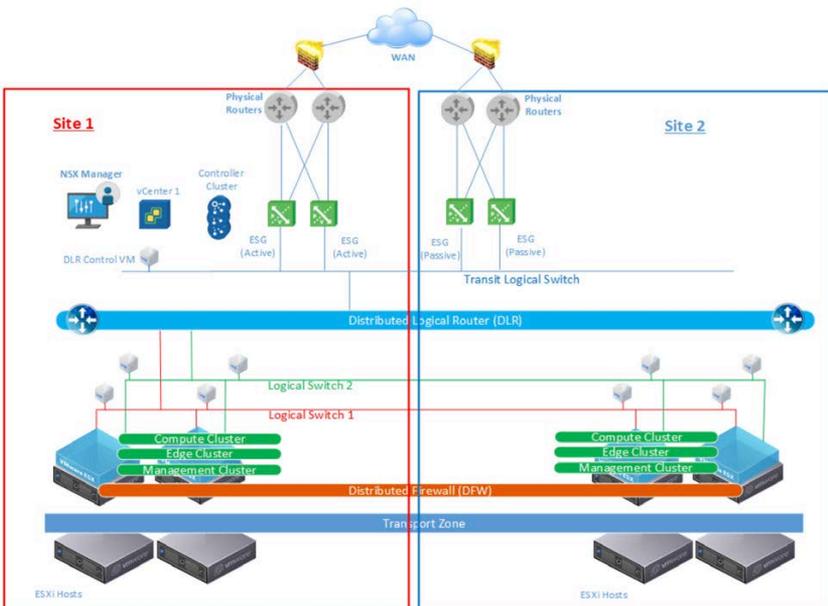


Figure 3.5 NSX with Separate vSphere Clusters

A few things to note in the NSX with separate vSphere cluster deployment displayed in Figure 3.5:

- NSX with separate clusters at each site uses a single vCenter. The compute, Edge, and management clusters are local to each site (e.g., not stretched across sites)
- Storage is local to each site. The latency requirement between sites falls back to that of vMotion (e.g. 150 ms RTT)

- An MTU of 1600 for VXLAN is required for the network connectivity between sites
- Since storage is local to each site, there is no storage latency requirement between sites. This allows the sites to be further geographically dispersed
- The vSphere clusters are not stretched across sites, so cluster-level technologies such as vSphere HA and DRS cannot be utilized across sites
- Since logical networking and security are spanned across sites, VM mobility and micro-segmentation are possible without stretching physical L2 VLANs
- Since logical networking and security are spanned across sites, a common administrative domain is required
- Figure 3.5 shows NSX ESGs deployed across both sites. ESGs at site 1 are used for egress for both sites, while ESGs at site 2 are passive devices. This active/passive egress functionality is controlled by setting a preferred routing metric for site 1 egress routes

It is also possible to deploy an active/active egress solution. This requires the local egress feature that is only available with the universal distributed logical router provided by Cross-VC NSX. In this single vCenter design with local egress, only static routing is supported between the universal distributed logical router (UDLR) and equal cost multipath (ECMP) Edges. In both cases, active workloads can reside at both sites. The active/passive model is preferred for simplicity and to avoid asymmetric routing issues. The ESGs at a single site can also be deployed in HA mode if stateful services are required.

### 3.) Multi-site with Cross-VC NSX

Cross-VC NSX was introduced in NSX 6.2. Cross-VC NSX allows spanning of logical networks and security policies across multiple vCenter domains while maintaining the use of dedicated vCenter servers per site as shown in Figure 3.6. In this solution, vSphere clusters are not stretched across sites; instead, the NSX logical networking and security domains span vCenter domains. A stretched storage solution is not required or used between the multiple vCenter domains; the datastores in each vCenter domain have different managed object reference IDs (moref ID) and would be seen as different datastores.

Figure 3.6 shows a Cross-VC NSX deployment. Each site has local storage and the latency requirement between sites is the same as in the previous single VC/NSX multi-site deployment model (e.g., maximum RTT of 150 ms).

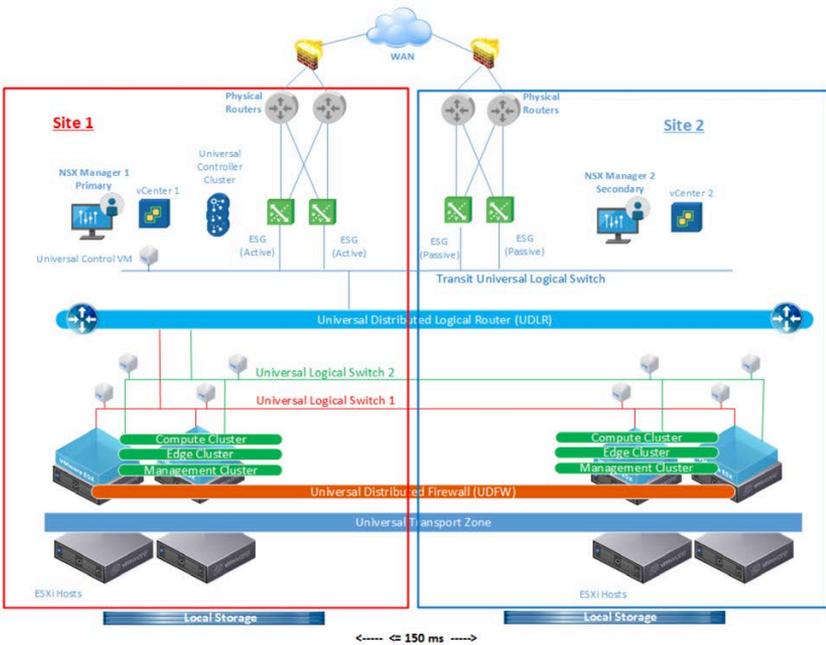


Figure 3.6 Cross-VC NSX Deployment

The Cross-VC NSX deployment displayed in Figure 3.6 has the following characteristics:

- The compute, edge, and management clusters are local to each site (e.g., not stretched)

- Storage is local to each site. The latency requirement between sites falls back to that of long distance vMotion (150 ms RTT)
- Since storage is local to each site, there is no storage-specific latency requirement between sites, allowing them to be further geographically dispersed
- Since the vSphere clusters are not stretched across sites, cluster-level technologies such as vSphere HA and DRS cannot be utilized across sites; however, they can be used locally
- For live migrations, an enhanced vMotion must be performed to move the workload and change its associated storage. This is required because the live migration crosses vCenter boundaries and storage is local to vCenter
- Since logical networking and security span sites, VM mobility and micro-segmentation is possible across sites without spanning physical L2
- Cross-VC NSX allows for spanning of logical networking and security across vCenter domains as well as across multiple sites
- An MTU of 1600 for VXLAN is required for the network connectivity between sites
- Since logical networking and security spans sites, there must be a common administrative domain
- The diagram in Figure 3.6 shows NSX ESGs deployed at both sites. ESGs at site 1 are used for egress by both sites while ESGs at site 2 are standby devices for egress traffic. This active/passive egress functionality is controlled by setting a preferred routing metric for site 1 egress routes

It is also possible to deploy an active/active egress solution using the local egress feature. In both cases, static and dynamic routing is supported and active workloads can reside at both sites. With the local egress implementation, a universal control VM would reside at both sites. The active/passive model is preferred for simplicity and to avoid asymmetric flows. The ESGs at a single site can also be deployed in HA mode instead of ECMP if stateful services are required.

## 4.) Multi-site with L2VPN

The Edge Services Gateway provides L2VPN capabilities for simple L2 extension across data centers. Common use cases for L2VPN include workload migration, cloud bursting, and onboarding.

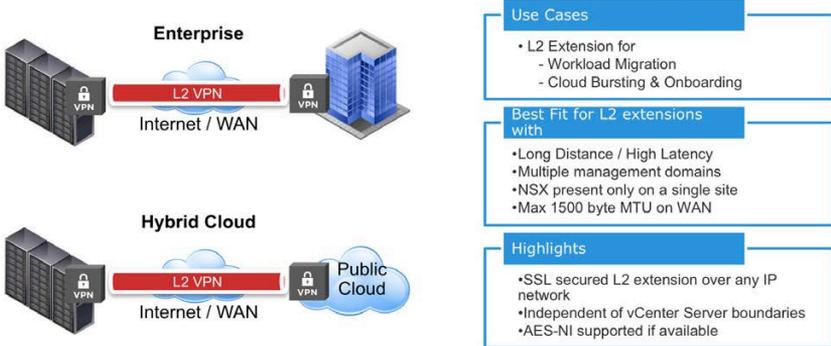
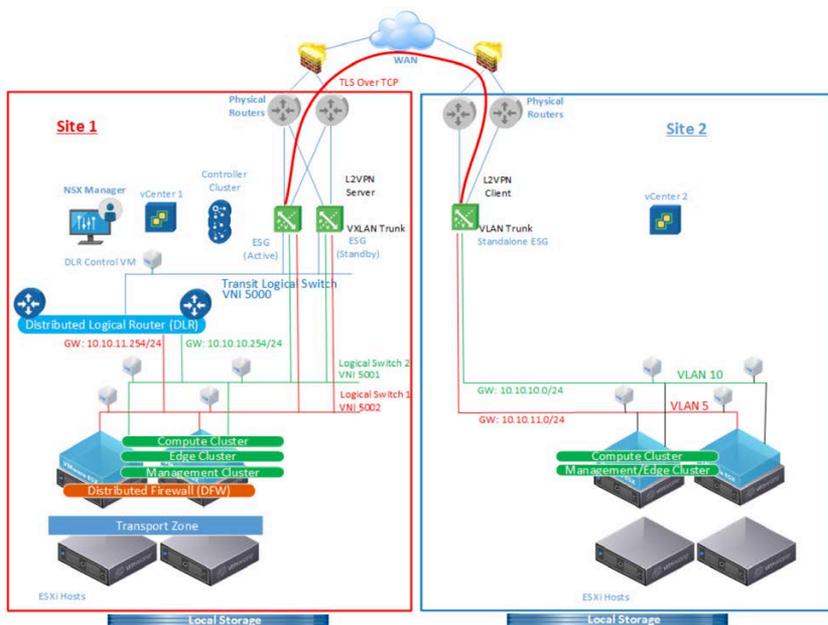


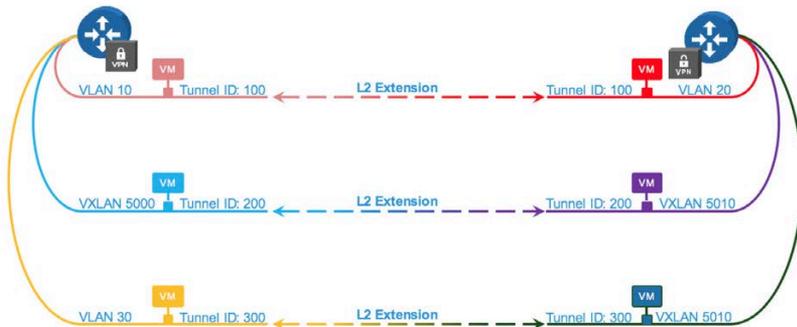
Figure 3.7 NSX L2VPN For L2 Extension

Figure 3.8 displays NSX's L2VPN service extending L2 across two sites. Workloads on physical VLANs at site 2 are being migrated to logical networks at site 1. The DLR at site 1 is the gateway for both the logical networks at site 1 as well as the physical networks at site 2. The logical switches at site 1 also connect to the ESG for L2 extension to site 2 via L2VPN. L2VPN is bridging between VXLAN and VLAN backed networks.



**Figure 3.8** NSX L2VPN Used For L2 Extension Between Sites And To Migrate Workloads

As shown in Figure 3.9, NSX L2VPN supports a combination of network bridging options. The L2 extension between sites can be VLAN-to-VLAN (same or different VLANs), VXLAN-to-VXLAN (same or different VXLANs), and VLAN-to-VXLAN. A unique identifier called the **Tunnel ID** is configured on both the ESG server and ESG client. This identifier is used to provide the L2 extension across sites. All traffic flows across a single TLS-over-TCP connection where the traffic is compressed and multiplexed.



**Figure 3.9** NSX L2VPN Supports A Combination Of Networks

A few things to note in the NSX L2VPN deployment of Figure 3.8:

- NSX is only deployed at site 1; it is only required for the L2VPN server side. The client side is deployed as a standalone ESG L2VPN client appliance. The standalone client only supports L2 and cannot be used as a gateway or other services offered by an ESG
- The NSX L2VPN server is deployed in HA mode. The L2VPN client can also be deployed in HA mode with NSX deployed on-prem or with the standalone ESG L2VPN client. HA mode with the standalone ESG L2VPN client is supported starting NSX 6.4
- NSX licensing is required only for the server side, not for the client
- Although there is no specific latency requirement between sites, high latency would have a negative impact on application performance
- There is no 1600 MTU requirements across sites when bridging VXLAN networks
- Since there is no stretched storage or storage latency requirement between sites, the sites can be geographically dispersed
- The administrative/management domains are unique to each site, so migrations are always cold migrations
- The DLRs at site 1 are the gateways for both the logical networks at site 1 and the physical networks at site 2. The gateway could have also been set on the L2VPN ESG or physical routers/L3 switches
- The ESGs at site 1 are used for north/south egress for both sites. The NSX L2VPN also supports egress optimization by using the same gateway IP on both sites. ARP filtering on the gateway

address is used to avoid flapping related to those duplicate IPs. This allows for local north/south egress at both sites

- NSX L2VPN provides a good solution for L2 extension if the 1600 MTU requirement for VXLAN cannot be met for some of the other solutions. NSX L2VPN works by default with 1500 MTU and can support up to 9000 MTU
- NSX ESG supports up to ten interfaces and up to 200 sub-interfaces per edge
- One NSX ESG L2VPN server can support up to five clients

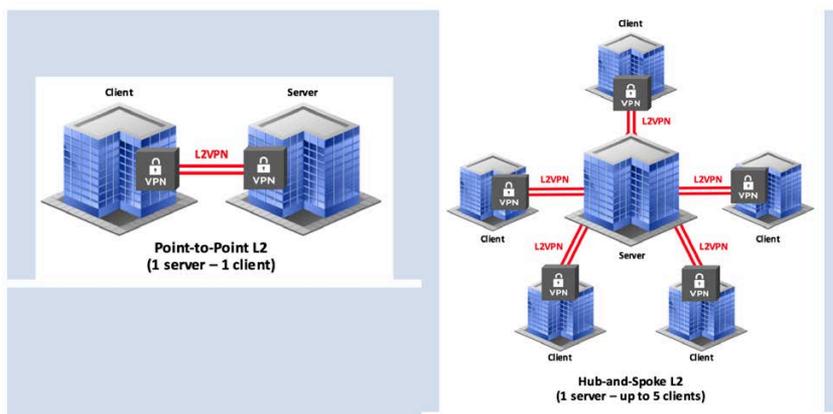


Figure 3.10 NSX L2VPN Topologies

## NSX Multi-site Solutions Summary and Comparison

Selecting the proper multi-site NSX deployment model depends on several factors, including bandwidth between entities, latency between sites, data center interconnect (DCI) technology in use, MTU, and administrative domain boundaries. Table 1 shows a side-by-side comparison of the different NSX deployment options for multi-site data centers. It also provides guidance about which solution may be best for a given environment based on customer needs.

**Table 3.1** NSX Multi-site L2 Extension Option Comparison

	<b>NSX with vMSC</b>	<b>NSX with Separate Clusters</b>	<b>Cross-VC NSX</b>	<b>L2VPN</b>
Scope	Metro	Geo	Geo	Geo
DCI Network Latency	Storage Driven ≤10 ms (EMC VPLEX) ≤5 ms for most storage vendors	vSphere 6 vMotion Driven ≤150 ms In line with NSX Control Plane latency of ≤150 ms	vSphere 6 vMotion Driven ≤150 ms In line with NSX Control Plane latency of ≤150 ms	No latency requirement, but high latency will result in poor application performance
Network MTU	1600 MTU (VXLAN)	1600 MTU (VXLAN)	1600 MTU (VXLAN)	1500 MTU (default)
DCI bandwidth / connectivity	≥10 Gbps	≥1 Gbps (driven by vMotion requirements)	≥1 Gbps (driven by vMotion requirements)	≤1 Gbps
Throughput	~line rate VXLAN (kernel module)	~line rate VXLAN (kernel module)	~line rate VXLAN (kernel module)	not line rate; see NSX documentation for current specs; up to 5 clients per server
Administrative Domain	Common	Common	Common	Separate
Storage	Stretched Storage (Metro)	Independent Storage	Independent Storage	Independent Storage
Features	Seamless pooling across datacenters; can leverage vSphere HA and DRS capabilities with consistent networking and security across sites	VM mobility across sites; resource pooling; consistent networking and security across sites	VM mobility across sites and vCenters; resource pooling; consistent networking and security across sites and vCenters; Cross-VC NSX can provide an enhanced DR solution	NSX at one or both ends; administrative domain across sites can be different; L2 extension to data center or cloud for workload migration or cloud bursting

With an understanding of the different options available from NSX for multi-site L2 extension and deployments established, the remainder of the book will focus on the Cross-VC NSX multi-site deployment model.



# Cross-VC NSX Overview

## About Cross-VC NSX

Cross-VC NSX was introduced in NSX 6.2. This book discusses Cross-VC NSX specifics as of release NSX 6.3.5 along with the relevant changes and additions provided by NSX 6.4.

VMware NSX is a network virtualization technology that decouples the networking and security services from the underlying physical infrastructure. By replicating traditional networking hardware constructs and moving the network intelligence to software, logical networks can be created efficiently over any basic IP network transport. The software-based approach to networking provides the same benefits to the network as what server virtualization provided for compute.

Although earlier versions of NSX provided flexibility, agility, efficiency and other benefits of network virtualization, prior to NSX 6.2 the logical networking and security constructs were constrained to the boundaries of a single vCenter domain.

Figure 4.1 shows NSX logical networking and security deployed within a single site and across a single vCenter domain.

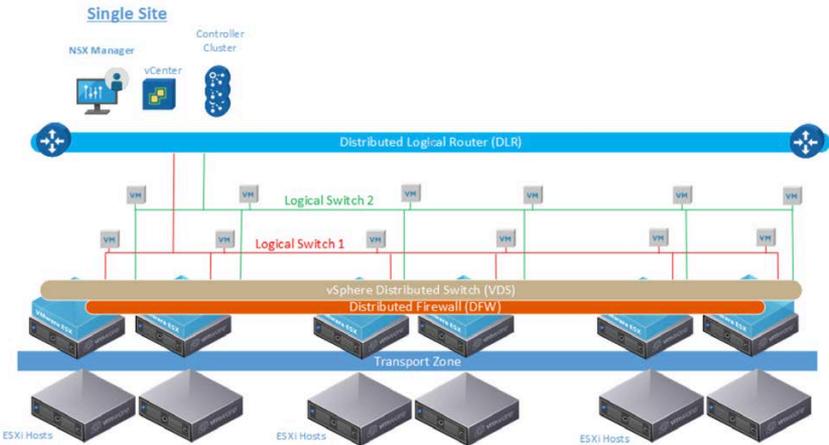


Figure 4.1 VMware NSX Deployment - Single Site, Single Vcenter

Although it was possible to use NSX with one vCenter domain and stretch logical networking and security across sites, the benefits of network virtualization with NSX were still limited to one vCenter domain. Figure 4.2 shows multiple vCenter domains located at different sites. Each domain has a separate NSX Controller with isolated logical networking and security due to NSX's 1:1 mapping with vCenter.

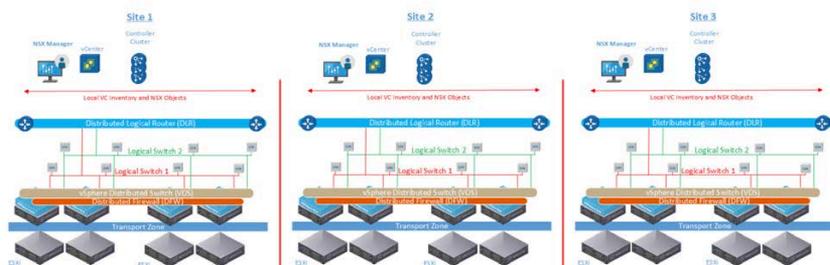


Figure 4.2 Pre-NSX 6.2 Cross-VC NSX

The Cross-VC NSX feature introduced in VMware NSX 6.2 allows for NSX networking and security to be implemented across multiple vCenter domains. NSX still has a 1:1 mapping with vCenter, but now logical switches (LS), distributed logical routers (DLR), and distributed firewalls (DFW) can be deployed across multiple vCenter domains as universal objects.

These Cross-VC NSX objects are called universal objects. The universal objects are similar to local objects such as distributed logical switches, routers, and firewalls, except they have a universal scope (e.g. they span multiple vCenter instances).

Combined with vSphere 6.0 support for Cross-VC vMotion, Cross-VC NSX ensures consistent networking and security across sites and vCenter domains when a VM is migrated to a different site and/or vCenter domain.



With Cross-VC NSX, in addition to the site local objects, users can implement universal objects such as universal logical switches, universal distributed logical routers, and universal distributed firewalls across a multi-vCenter environment. That environment can be within a single site or across multiple sites.

The benefits of NSX networking and security stretching across multiple vCenter domains as shown in Figure 4.3 become immediately clear. Logical networking and security can be enabled for application workloads that span multiple vCenters and multiple physical locations.

Typically, customers that have multiple sites have multiple vCenters across different sites, so Cross-VC NSX provides a perfect solution. VMs can be vMotioned across vCenter boundaries while maintaining consistent networking and security policy enforcement. Manual modification or provisioning of networking and security services at the

new site is not required. The workloads' IP addresses and security policies remain consistent upon vMotion events or as a result of disaster recovery events, regardless of vCenter or site.

NSX control and automation is extended across vCenter boundaries, whether those boundaries are within the same data center or across multiple. This opens the door for many use cases that can now be applied across multiple vCenter domains which may also be distributed across multiple sites: VM/workload mobility, resource pooling, consistent multi-site security, disaster avoidance, partial/full application failure, and disaster recovery.

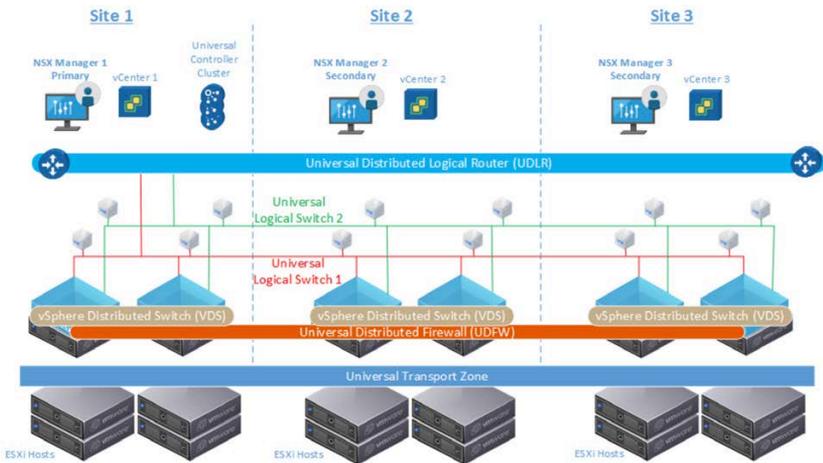


Figure 4.3 VMware Cross-VC NSX Deployment - Multi-Sites, Multiple Vcenters

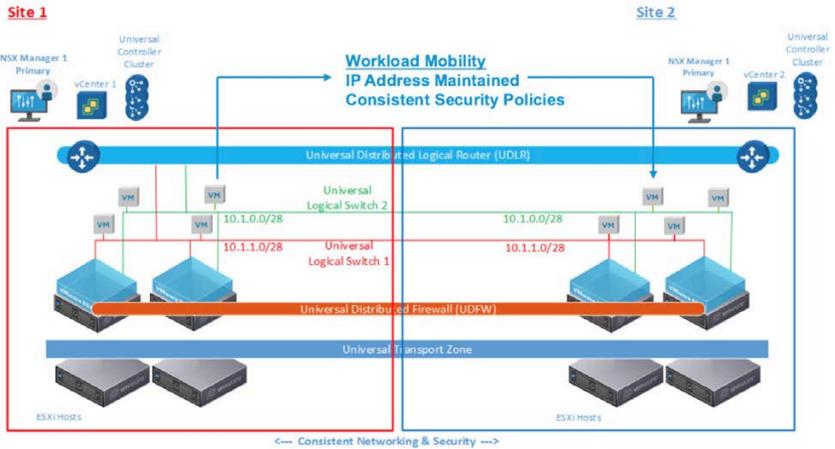
## Cross-VC NSX Use Cases

This section briefly discusses several important Cross-VC NSX use cases. References for additional use case details are provided at the end of the list.

### Use Case 1: Workload Mobility

Since logical networking and security can span multiple vCenter domains and multiple sites, Cross-VC NSX allows for enhanced workload mobility. This may not only span multiple sites, but also multiple vCenter domains across active/active data centers. Workloads can move between vCenter domains/sites on demand for tasks such as data center migration, data center upgrades/security patches, or

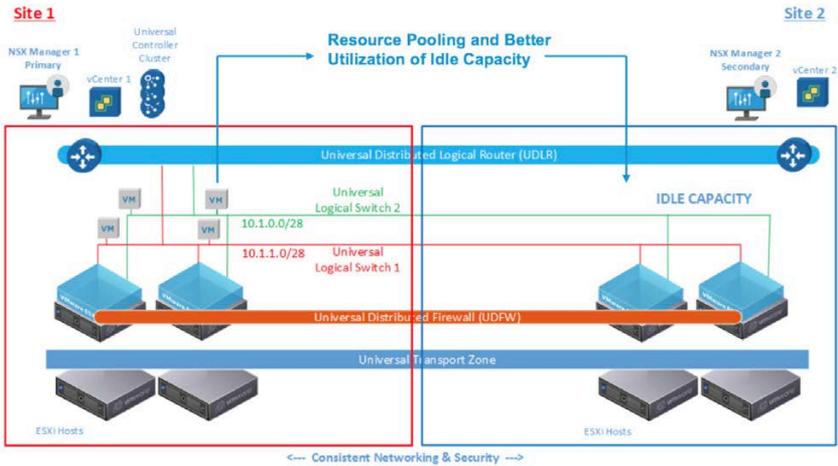
disaster avoidance. Workload mobility boundaries are no longer constrained by artificial vCenter boundaries as the IP addresses and security policies for the workloads remain consistent across vCenter domains and sites.



**Figure 4.4** Workload Mobility Across vCenter Domains with Consistent Networking and Security Policies for the Application

## Use Case 2: Resource Pooling

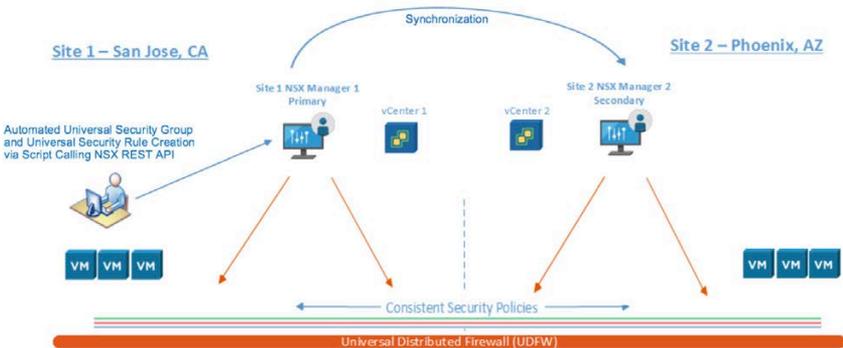
By enabling logical networking and security across multiple vCenters, Cross-VC NSX enables access to and pooling of resources from multiple vCenter domains. This allows for better utilization of resources across multiple vCenter domains and sites. If resources such as storage, CPU, or memory are low at one site, the workload can be deployed or migrated to another site while still utilizing the same logical networking constructs and security policies. Resources are no longer isolated based on vCenter boundaries and idle capacity in other vCenter domains/sites can be leveraged for better overall utilization.



**Figure 4.5** Resource Pooling And Better Utilization Of Idle Capacity Across Vcenter Domains and Sites

### Use Case 3: Unified Logical Networking and Security Policy

Cross-VC NSX provides centralized management of universal networking and security constructs via the primary NSX Manager. Users are no longer required to manually replicate security policies across different domains and sites. Configurations on the primary NSX Manager result in consistent application security policies being pushed across all sites. This is illustrated in Figure 4.6.



**Figure 4.6** Security Policy Automation and Consistent Security Across vCenter Domains

Security is discussed in more detail in the Cross-VC NSX Security section of this book.

## Use Case 4: Disaster Recovery

By spanning logical networking and security across sites, Cross-VC NSX inherently provides features desired by disaster recovery. Applications can be restarted at the recovery site upon a DR event while maintaining their IP addresses. Furthermore, security enforcement for the applications is maintained due to the universal nature of the distributed firewall and security policies.

Using the local egress feature, NSX provides advanced control over egress traffic with site, cluster, host, and even DLR-level granularity. This allows the selection of which set of ESGs are used for egress. This can be a powerful tool for various partial failure scenarios.

Additionally, Cross-VC NSX also improves the disaster recovery benefits provided by VMware Site Recovery Manager™ (SRM) or other disaster recovery orchestration tools. SRM has a 1:1 relationship to vCenter server and utilizes an active/standby model. All workloads are running on the active site with placeholder VMs at the standby recovery site.

NSX is tightly integrated with SRM. When using storage policy-based protection groups (SPPGs), the source and destination networks for the VMs can be automatically mapped when using Cross-VC NSX universal logical switches. There is no need to perform manual mapping as all networking and security services are synchronized across sites, providing huge benefits for highly dynamic environments. NSX also ensures that workloads maintain their IP addresses when recovered. NSX also offers other DR advantages, such as the ability to launch audit-driven test disaster recovery plans that will keep the original workloads running while isolating the recovered workloads. This prevents application connectivity disruption caused by test networks and IP duplicates.

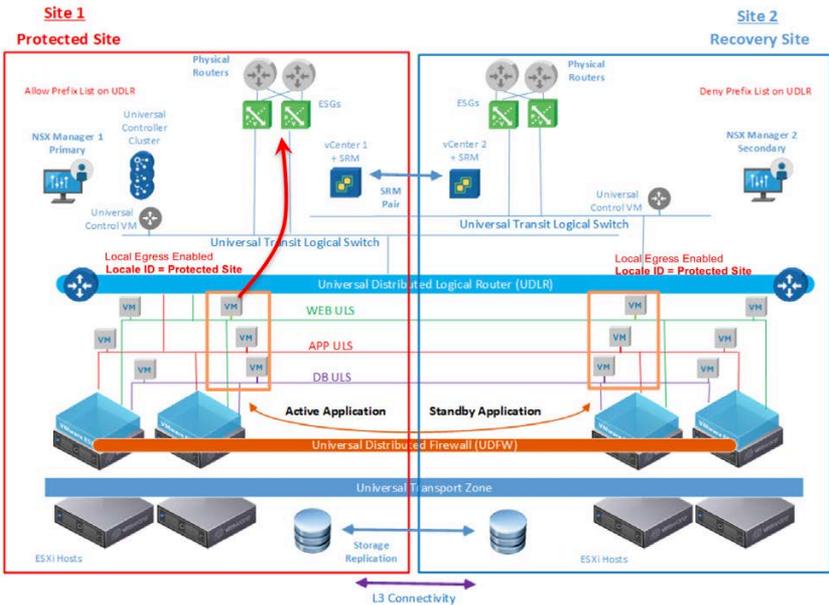


Figure 4.7 SRM DR Solution Leveraging Cross-VC NSX

For more detailed information on this specific use case, see the following resources:

- “Disaster Recovery with NSX and SRM” Whitepaper (<https://communities.vmware.com/docs/DOC-31692>)
- VMware Network Virtualization Blog: “Enhanced Disaster Recovery with Cross-VC NSX” (<http://blogs.vmware.com/networkvirtualization/2016/04/enhanced-disaster-recovery-with-nsx-and-srm.html>)
- VMware Network Virtualization Blog: “VMware NSX and SRM: Disaster Recovery Overview and Demo” (<https://blogs.vmware.com/networkvirtualization/2017/01/vmware-nsx-srm-disaster-recovery-overview-demo.html>)
- VMware Network Virtualization Blog: “Disaster Recovery with VMware NSX-V and Zerto” (<https://blogs.vmware.com/networkvirtualization/2017/08/disaster-recovery-vmware-nsx-v-zerto.html>)

A demonstration of VMware NSX and SRM working together is available on the VMware NSX YouTube channel: “VMware NSX and SRM – Disaster Recovery and Overview Demo” (<https://youtu.be/TjnUvezJ4LU>).

# Cross-VC NSX Terminology

The Cross-VC NSX functionality added in NSX 6.2 introduced new terminology, architecture, and concepts.

**Local Objects:** Objects local to a single vCenter instance. Local objects are labeled with '**Global**' scope within the NSX UI

**Universal Objects:** Objects that can span multiple vCenter domains. Universal objects are labeled as '**Universal**' within the NSX UI

**Primary NSX Manager:** Used to deploy and configure NSX universal objects. Only one NSX Manager can be primary. A primary NSX Manager can be demoted to a secondary NSX Manager

**Secondary NSX Manager:** Universal objects cannot be created on the secondary NSX Manager(s) but universal objects created on the primary NSX Manager are synchronized to the secondary NSX Manager(s). There can be a maximum of seven secondary NSX Managers (15 in NSX 6.4). A secondary NSX Manager can be promoted to a primary NSX Manager

**Universal Synchronization Service (USS):** Runs on the primary NSX Manager to replicate the universal objects to the secondary NSX Managers

**Universal Control Cluster (UCC):** Maintains information about local and universal logical objects. Examples include universal logical switches, universal logical routers, and local logical switches and logical routers that are local to a vCenter domain

**Universal Transport Zone (UTZ):** Defined from the primary NSX Manager, a UTZ is the span of universal logical objects across vSphere clusters. Clusters that ought to participate in universal logical networking across multiple vCenters must be configured to be part of this universal transport zone. There can only be one UTZ in a Cross-VC NSX deployment

**Universal Logical Switch (ULS):** Same as a local LS but able to span multiple vCenter domains. Allows for L2 connectivity across multiple vCenters. An LS is universal if it is deployed in a UTZ

**Universal Distributed Logical Router (UDLR):** Same as a DLR but able to span multiple vCenter domains. Allows L3 connectivity for universal logical switches

**Universal Distributed Firewall (UDFW):** Distributed firewall spanning vCenter boundaries and providing a consistent security model across all vCenter domains/sites

**Universal Firewall Rules:** DFW rules that are configured under the universal section of the DFW and apply across vCenter boundaries

**Universal Network and Security Objects:** The universal section of the DFW supports the following network and security objects: universal IPsets, universal MAC sets, universal security groups, universal services, universal service groups, universal security tags for active/standby scenarios, and match based on VM name for active/standby scenarios

**Universal security tags (USTs)** and match based on VM name were introduced in NSX 6.3. Universal security groups can contain matching criteria based on UST and VM name; however, the new matching criteria can only be used in active/standby scenarios. Therefore, the application can only be active at one site. The new matching criteria are not supported for policies applied to applications which span sites. This is discussed in more detail in the Cross-VC NSX Security section

**Local Egress:** Allows control of which routes are provided to ESXi hosts based on an identifier called 'locale ID'. The UCC learns the routes from a particular site/vCenter domain and associates the learned routes with the locale ID. If local egress is not enabled, the locale ID is ignored and all ESXi hosts connected to the universal distributed logical router receive the same routes. Local egress can be enabled only when creating a UDLR

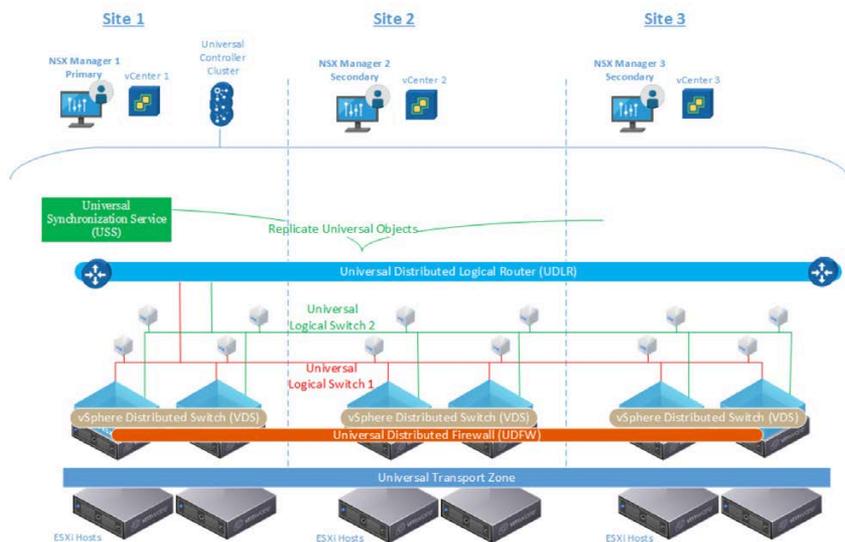
## Architecture and Key Concepts

### NSX Manager Roles

As with standalone NSX, Cross-VC NSX maintains a 1:1 relationship between NSX Manager and VMware vCenter Server instances. One NSX Manager is given a role of primary, with the others joining the Cross-VC NSX environment with a secondary role. Up to eight NSX Manager/vCenter pairs are supported in NSX 6.3; starting with NSX 6.4, up to 16 NSX Manager/vCenter pairs are supported.

The primary NSX Manager is used to deploy the UCC in the vCenter domain that NSX is linked to. This provides the control plane for the Cross-VC NSX environment. The secondary NSX Managers do not have their own controller clusters, leveraging instead the UCC deployed at the primary NSX Manager's site for both local and universal control plane and objects.

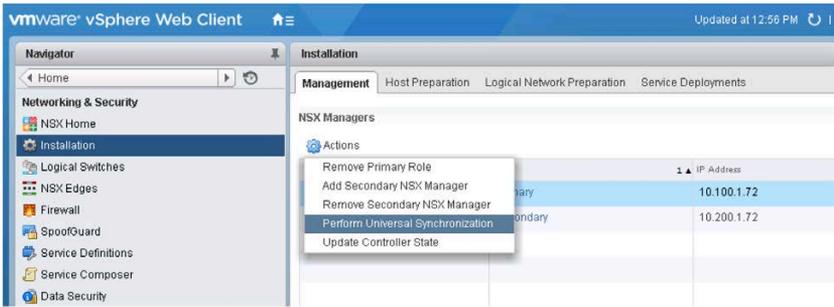
As shown in Figure 4.8, the primary NSX Manager will use the USS to replicate universal objects to the secondary NSX Managers. The UCC resides only on the primary NSX Manager. Local objects on the secondary NSX Managers are created via the respective secondary NSX Managers.



**Figure 4.8** Multi-Site NSX Deployment With Multiple Vcenters

As mentioned above, the USS manages synchronization of universal objects created on the primary NSX Manager with all registered secondary NSX Managers. By default, the USS service is stopped on an NSX Manager and started only when the primary role is assigned. When secondary NSX Managers are registered with the primary, an internal system user account that has access to the universal objects is created on the secondary NSX Manager. This internal user account is used for synchronization of the universal objects by the USS. It also has read-only access to the universal objects so that a secondary NSX Manager can be promoted in the event of loss of the primary NSX Manager.

The USS synchronization updates are triggered whenever universal objects are created or modified. It is also run every 30 seconds to ensure consistency across the multi-vCenter environments. It can be initiated manually on the primary NSX Manager via the GUI, as shown in Figure 4.9.



**Figure 4.9** Manually Initiating Universal Synchronization Service (USS) Syncs

The USS updates perform a full sync operation, determining the differential between primary and secondary NSX Managers to synchronize state. Figure 4.10 shows the USS running on a primary NSX Manager.

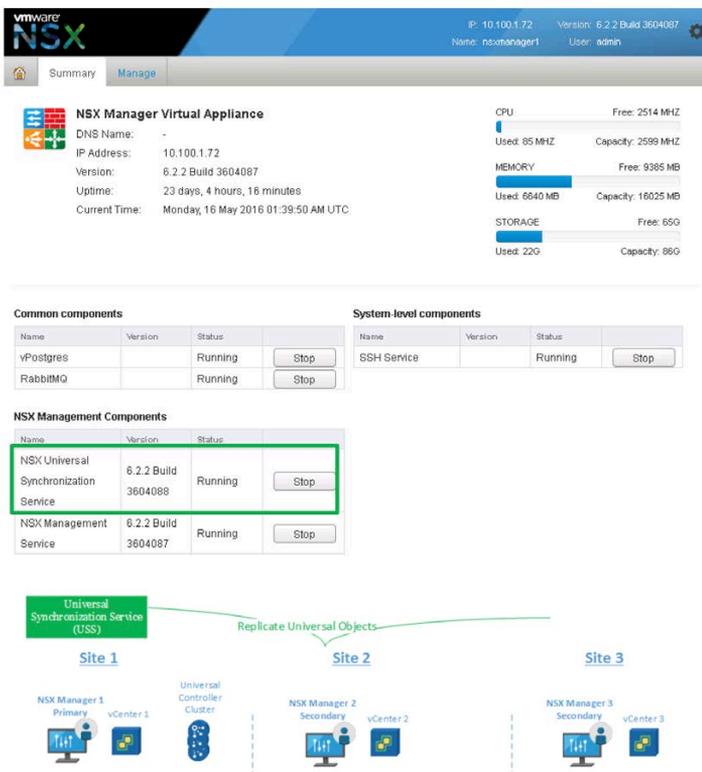


Figure 4.10 USS on Primary NSX Manager Replicating Universal Objects

Further information on the different roles that an NSX Manager can assume is provided below.

**Standalone Role:** The standalone role is the default role for NSX Manager. In this role it is not participating in a Cross-VC NSX deployment. All of the objects are local. This was the only role for NSX Manager prior to the addition of the Cross-VC NSX functionality in NSX 6.2. If an NSX Manager needs to be removed from a Cross-VC NSX deployment, it must first be put in a transit role and all universal objects must be deleted before changing it to a standalone role.

**Primary Role:** There is only one primary NSX Manager in a Cross-VC NSX deployment. The primary NSX Manager is used to deploy the UCC at the local site. The primary and all secondary NSX Managers communicate with the UCC for control plane functionality.



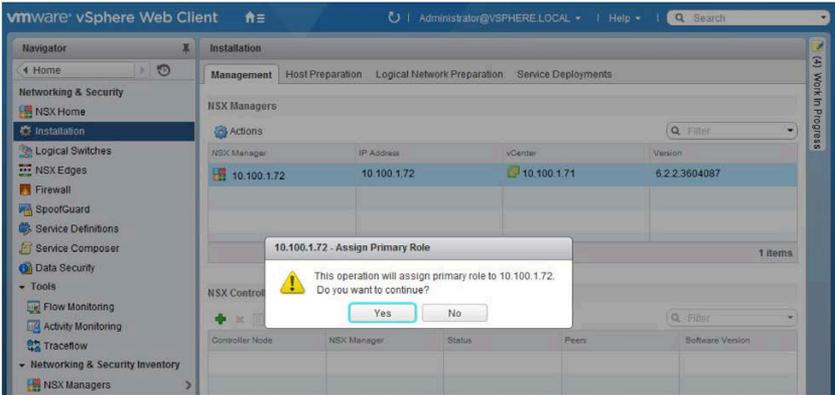


Figure 4.12 Confirming Primary Role Assignment To Standalone Nsx Manager

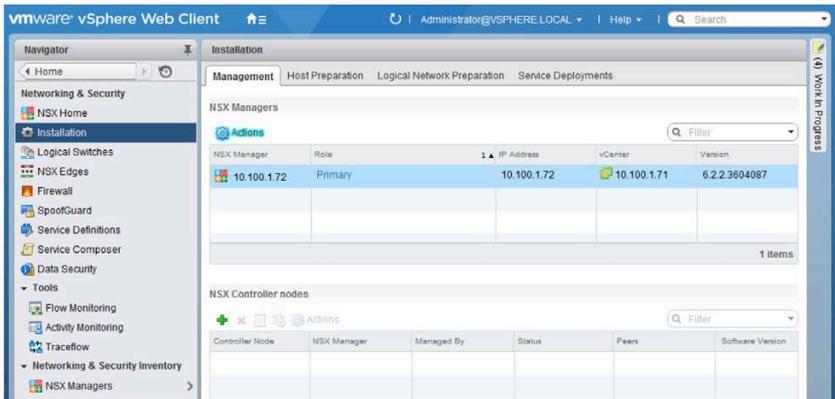


Figure 4.13 Standalone NSX Manager Made Primary NSX Manager

**Secondary Role:** Only one NSX Manager in a Cross-VC NSX deployment can be primary; the remaining ones - up to seven additional NSX Managers pre NSX 6.4, fifteen in NSX 6.4 - must be secondary. The secondary role is assigned when a standalone NSX Manager is registered with the primary NSX Manager; this is shown in Figures 4.15 and 4.16.

Although universal objects can only be created from the primary NSX Manager, local objects can still be created using the respective local NSX Manager.

The universal objects and security policies are synced to the secondary NSX Manager(s) via the USS on the primary NSX Manager. An internal user with access to the universal objects exists on each NSX Manager and helps with creating the universal objects locally once initiated by the USS on the primary NSX Manager. The secondary NSX Manager has read-only access; it can view universal objects but cannot create or edit them.

Figures 4.14–4.15 show a standalone NSX Manager being promoted to secondary by selecting the primary NSX Manager, clicking the **Actions** button, and selecting **Add Secondary NSX Manager**. In this example, three universal controllers – the required value and maximum number supported – have already been deployed from the primary NSX Manager prior to adding a secondary NSX Manager. This supports an easy scale-out model by starting with one vCenter/NSX Manager domain, then adding others as the organization scales.

As shown in Figure 4.14, when registering a secondary NSX Manager to the primary NSX Manager, the **Add Secondary NSX Manager** option must be selected.

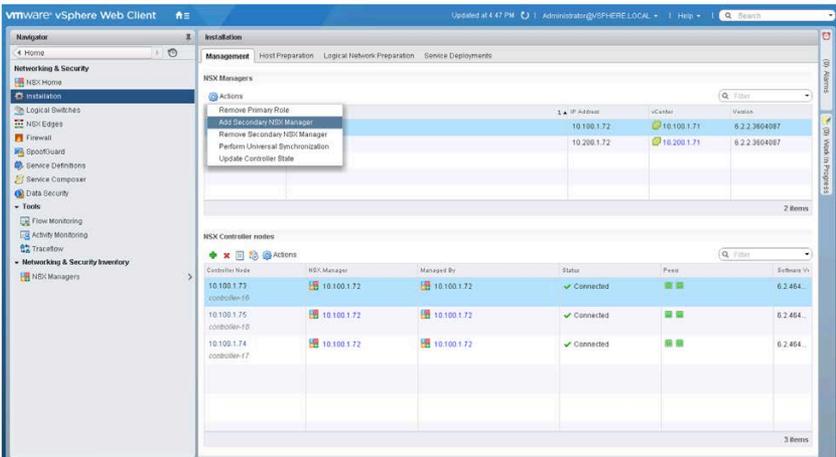


Figure 4.14 Adding a Secondary NSX Manager

Figure 4.15 shows that the credentials for the secondary NSX Manager must be provided to successfully register it with the primary NSX Manager:

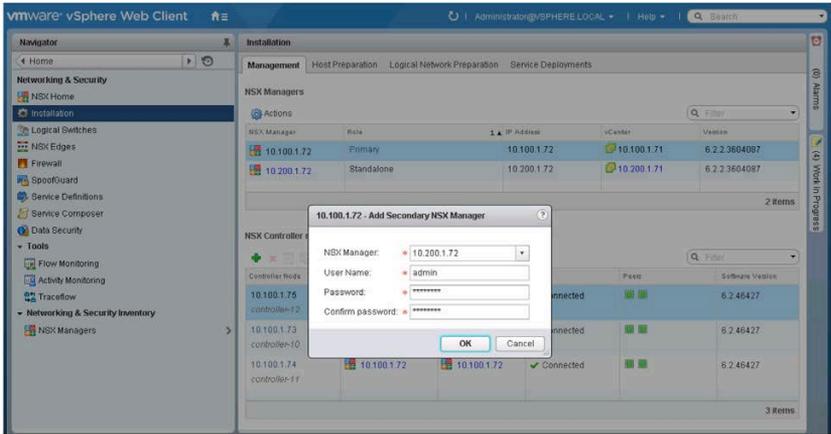


Figure 4.15 Providing Secondary NSX Manager Login Credentials



If they exist, the local controllers at the secondary sites should be deleted. When an NSX Manager at a secondary sites is registered with the primary NSX Manager, the UCC configuration is copied over to the secondary NSX Manager and its respective ESXi hosts. The **Update Controller State** option from the **Actions** button options should be selected for each of the secondary NSX Managers once they are registered with the primary NSX Manager. At this point all NSX Manager domains are using the UCC at the primary site.



In cases of primary site failure, a secondary NSX Manager must be manually promoted to a primary NSX Manager and the UCC must be deployed at the new primary site. This can be done through the NSX Manager GUI. The secondary NSX Manager must first be disconnected from the primary NSX Manager (**Installation->Management->Actions->Disconnect from Primary NSX Manager**), then promoted to primary (**Installation->Management->Actions->Assign Primary Role**).

When moving the NSX Manager from standalone mode to primary mode, it is also recommended to select the **Update Controller State** option from the **Actions** button options to ensure all controller information is updated. The **Update Controller State** action pushes the current VXLAN and DLR configuration as well as including universal

objects from the NSX Manager to the controller cluster.



In a brownfield environment where multiple NSX Manager domains are already installed, the controllers at the primary site do not have to be deleted. The NSX Manager can simply be made primary and the **Update Controller State** option from the **Actions** button options should be selected.

See the Universal Objects and Cross-VC NSX Communication section and ULS Creation Example section for additional details and considerations that need to be taken into account when creating universal objects.

Once the secondary NSX Manager is registered with the primary NSX Manager, three more controllers appear under the NSX Controller Nodes section as shown in Figure 4.16; however, no additional controllers were actually created. The additional rows displayed show the status and connectivity of the controllers as seen by the other NSX Managers as a result of the single pane of view provided by enhanced linked mode. Only the universal controller cluster at the primary site is used. The universal controller cluster must always be deployed in the vCenter domain to which the primary NSX Manager is linked.

Looking at the NSX Manager column in Figure 4.16, the bottom three controllers are linked with the primary NSX Manager at IP address 10.100.1.72 while the top three controllers are linked with the secondary NSX Manager at IP address 10.200.1.72. Under the NSX Controller nodes column, the IP addresses for the controllers are the same across both vCenter/NSX Manager domains, confirming there are only three actual controllers.

The primary NSX Manager rows have the green box status icons under the **Peers** column; this can be used to quickly identify the rows corresponding to the universal controller cluster and primary NSX Manager domain connectivity.

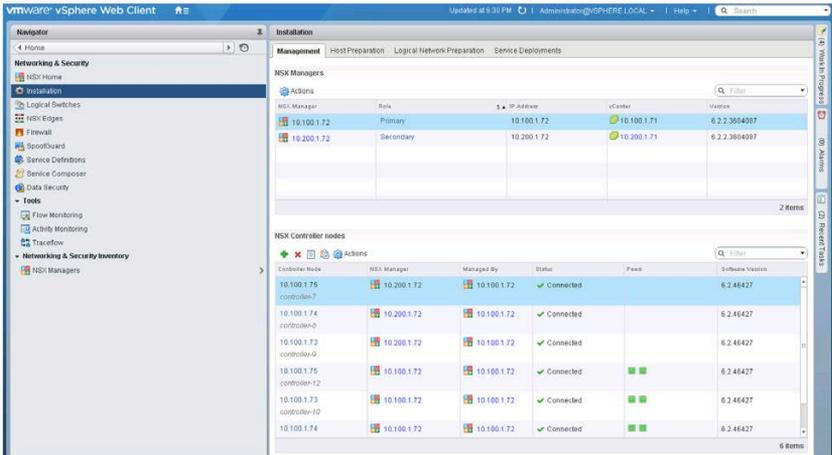


Figure 4.16 Secondary NSX Manager Successfully Registered with Primary NSX Manager

**Transit Role:** The transit role is a temporary role used when configuration changes are required for an NSX Manager being demoted from a primary or secondary role. There is no separate option to move to a transit role; it is done automatically when removing a primary or secondary role from the respective NSX Manager.

Figure 4.17 shows a secondary NSX Manager (10.200.1.72) being removed from a primary NSX Manager (10.100.1.72).

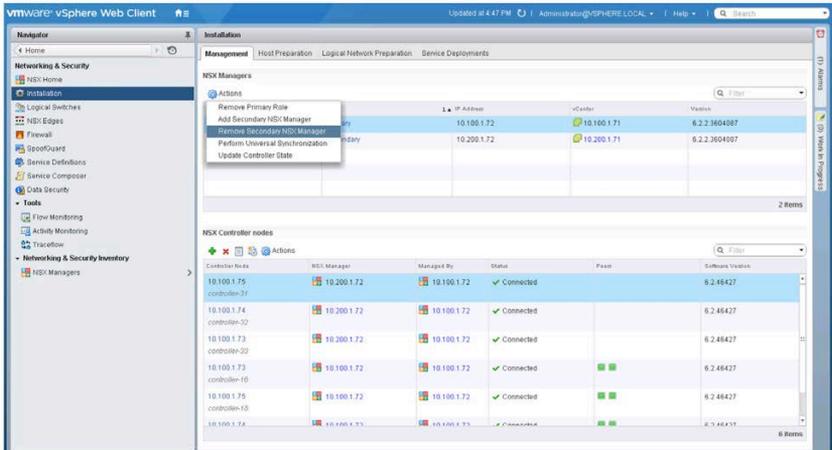


Figure 4.17 Removing a Secondary NSX Manager From Primary NSX Manager

Sometimes the respective secondary NSX Manager may not be accessible due to connectivity issues. The **Perform operation even if NSX Manager is inaccessible** option allows the secondary NSX Manager to be removed even if the secondary NSX Manager is inaccessible.

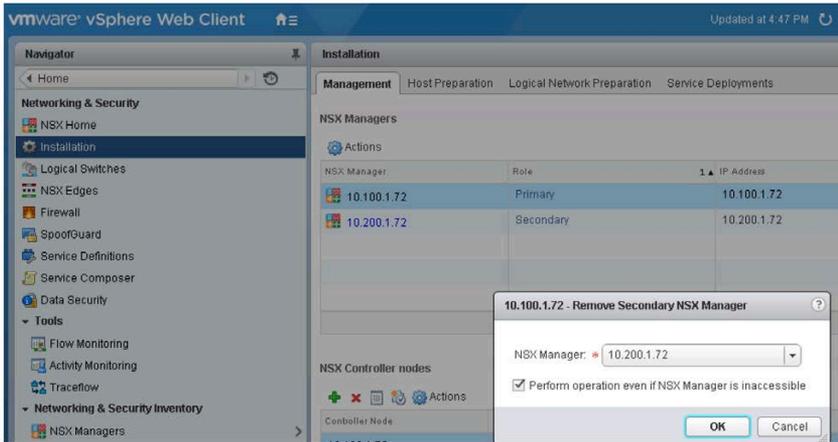


Figure 4.18 Confirming Removal of Secondary NSX Manager

Figure 4.19 shows the result once the secondary NSX Manager is successfully removed. The respective rows showing controller connectivity to the secondary NSX Manager/vCenter domain are automatically deleted once the role has been removed. At this point the secondary NSX Manager automatically goes into the transit role.

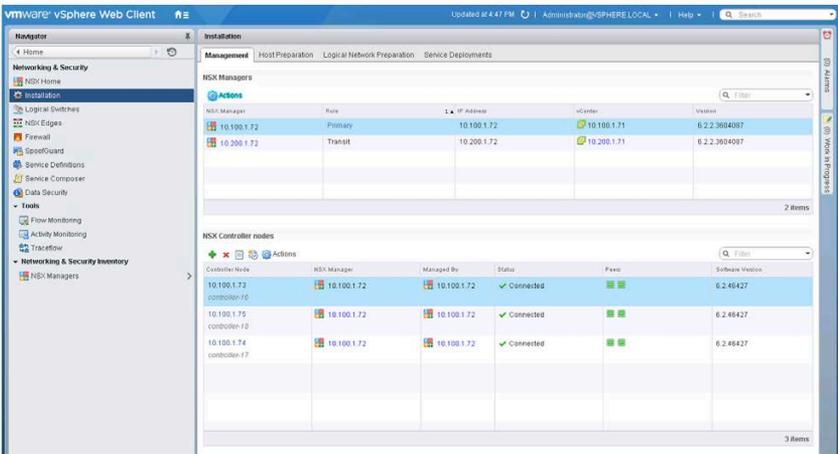


Figure 4.19 Secondary Role On Nsx Manager Successfully Removed



When moving an NSX Manager back to a standalone role, if universal objects exist, the NSX Manager will instead automatically be placed into a transit role. Universal objects must be manually deleted from the transit NSX Manager before the NSX Manager can be put into the standalone role.

## Universal Controller Cluster (UCC)

A maximum of three NSX Controllers are supported within a UCC. These NSX Controllers are used by all NSX Manager/vCenter domains participating in the cross-VC deployment. This UCC is associated with the primary NSX Manager, and must be deployed from the primary NSX Manager into the vCenter inventory of the vCenter that is linked to the primary NSX Manager. If the UCC already exists when secondary NSX Managers are registered, the respective controller information is imported to the secondary NSX Managers. The USS is leveraged to accomplish this.



Note that the UCC is used as a shared controller cluster for both universal and local objects for each NSX Manager.

Up to three controllers are supported. For resiliency and performance reasons, it is highly recommended that each controller reside on separate servers within the same cluster. Anti-affinity rules should be used to prevent controllers from residing on the same host. All controllers must be in the same vCenter domain linked to the primary NSX Manager.

The requirement to place the UCC at a single site controlled by the primary NSX Manager is to prevent split brain scenarios. A split brain scenario would cause controllers to behave independently at each site without knowledge of updates/changes at other sites. This would result in inconsistency and communication issues across sites.

Complete separation of control plane and data plane allows for continuous forwarding of packets using the existing data plane configuration even when connectivity issues to the control plane are experienced. Only new control plane operations such as dynamic routing updates to local DLR instances are affected upon a failure of the UCC. On the rare occasion of a primary site failure, the UCC can be deployed at a new primary site after promoting an NSX Manager at the new site to the primary role.

## Components Communication in the Cross-VC NSX Architecture

Figure 4.20 displays a diagram of NSX component communication in a Cross-VC NSX deployment. The same protocols and ports as a standalone NSX deployment are used. The only additional communication occurring stems from the USS synchronizing universal objects and configurations between the primary and secondary NSX Managers. This occurs over standard TLS/TCP (port 443) and leverages the NSX REST API.

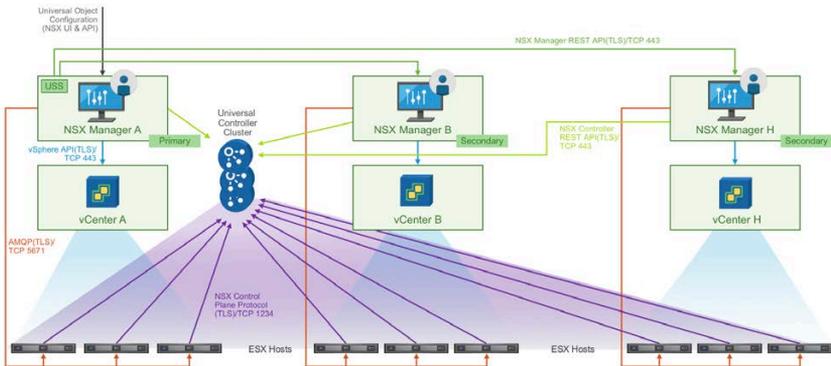


Figure 4.20 Cross-vCenter NSX Component Communication

While the USS only runs on the primary NSX Manager, every NSX Manager has a TLS-over-TCP connection (port 443) to its respective vCenter. Using vSphere API calls, NSX Managers are able to get the vCenter inventory, prepare hosts for NSX, and deploy NSX appliances as needed.

Each NSX Manager has a message bus connection using Advanced Message Queuing Protocol (AMQP) over TLS/TCP (port 5671) to ESXi hosts in its respective vCenter domain and is responsible for maintaining these connections. This message bus is used for tasks such as configuring Edge appliances and pushing DFW rules.

The UCC resides at the site of the vCenter managed by the primary NSX Manager and is also deployed via the primary NSX Manager. It has a secure TLS connection to each of the NSX Managers on port 443. When an NSX Manager registers as secondary - as shown in Figure 4.21 - a unique system user is created on the secondary NSX Manager with read-only privileges to universal objects. This system user is internal to the system and not accessible or viewable via GUI or REST API call.

When the UCC is deployed, the controller cluster configuration is synced to the secondary NSX Manager(s) via USS. The controller configuration is then pushed down to each host via the message bus connection each NSX Manager has with its respective ESXi hosts. Using this information, each ESXi host then uses its netcpa agent to make a control plane connection via an internal protocol to the UCC via TLS/TCP on port 1234. Over this connection, control plane information such as ARP, MAC, VTEP (Virtual Tunnel Endpoint), and forwarding table information is communicated.

In Figure 4.21, the secondary NSX Manager is added by selecting the primary NSX Manager and using the **Actions** button to select **Add Secondary NSX Manager**. Before it is added as a secondary NSX Manager, the NSX Manager has the default standalone role.

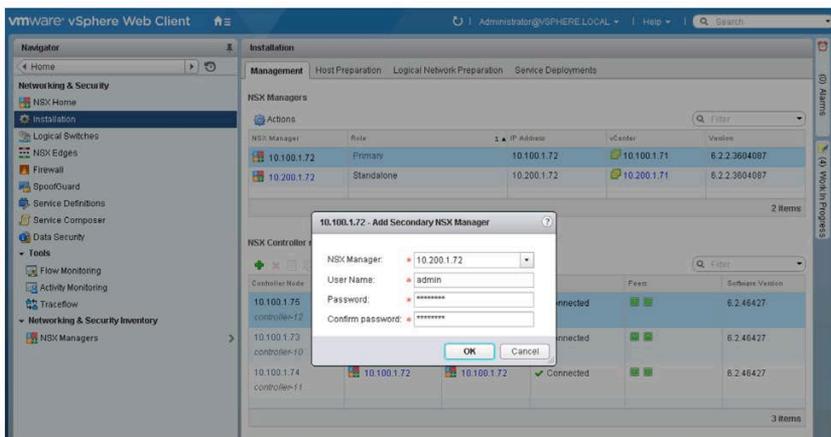


Figure 4.21 Adding a Secondary NSX Manager

## Universal Objects Architecture Considerations

### Architecture Notes

A few key points to remember in terms of universal objects:

- Only one Universal Transport Zone (UTZ) is supported; all universal objects will be part of this zone. The UTZ defines the span of the universal objects. Multiple local transport zones, if needed, can still be utilized in parallel
- All universal objects must be created and managed via the primary NSX Manager
- The USS only runs on the primary NSX Manager. It is always stopped on the secondary NSX Manager(s),

as shown in Figures 4.22 and 4.23. This service will synchronize universal object configuration to the secondary NSX Managers. This is how the secondary NSX Managers learn the UCC configuration

- A local LS cannot connect to a UDLR; a local LS must connect to local DLR. ESGs can connect to both local and universal logical switches, providing routing between local and universal networks

**NSX Manager Virtual Appliance**

DNS Name: -  
 IP Address: 10.100.1.72  
 Version: 6.2.2 Build 3604087  
 Uptime: 23 days, 4 hours, 16 minutes  
 Current Time: Monday, 16 May 2016 01:39:50 AM UTC

CPU Free: 2514 MHz  
 Used: 85 MHz Capacity: 2599 MHz

MEMORY Free: 9385 MB  
 Used: 6640 MB Capacity: 16025 MB

STORAGE Free: 65G  
 Used: 22G Capacity: 86G

Common components			
Name	Version	Status	
vPostgres		Running	Stop
RabbitMQ		Running	Stop

System-level components			
Name	Version	Status	
SSH Service		Running	Stop

NSX Management Components			
Name	Version	Status	
NSX Universal			
Synchronization Service	6.2.2 Build 3604088	Running	Stop
NSX Management Service	6.2.2 Build 3604087	Running	Stop

**Figure 4.22** USS Only Runs On Primary NSX Manager

**NSX Manager Virtual Appliance**

DNS Name: -  
 IP Address: 10.200.1.72  
 Version: 6.2.2 Build 3604087  
 Uptime: 40 days, 22 hours, 10 minutes  
 Current Time: Monday, 16 May 2016 02:15:53 AM UTC

CPU Free: 2434 MHz  
 Used: 165 MHz Capacity: 2599 MHz  
 MEMORY Free: 8027 MB  
 Used: 7998 MB Capacity: 16025 MB  
 STORAGE Free: 65G  
 Used: 21G Capacity: 86G

**Common components**

Name	Version	Status	
yPostgres		Running	Stop
RabbitMQ		Running	Stop

**System-level components**

Name	Version	Status	
SSH Service		Running	Stop

**NSX Management Components**

Name	Version	Status	
NSX Universal	6.2.2 Build	Stopped	
Synchronization Service	3604088	Stopped	Start
NSX Management Service	6.2.2 Build 3604087	Running	Stop

**Figure 4.23** USS is Always Stopped On Secondary NSX Manager(s)

Figure 4.24 shows both a local transport zone (scope of global) and a universal transport zone (scope of universal) being utilized. Universal objects always appear with a globe overlaid on top of the object icon. The control plane mode for each respective transport zone can be different; in this case both are using unicast mode. Control plane modes and handling of broadcast, unknown unicast, and multicast traffic will be discussed in more detail in the Handling Broadcast, Unknown Unicast, Multicast (BUM) Traffic section.

VMware vSphere Web Client

Administrator@VSPHERE LOCAL | Help | Search

Installation: Management | Host Preparation | Logical Network Preparation | Service Deployments

NSX Manager: 10.100.1.72 (Role: Primary)

VLAN Transport | Segment ID | Transport Zones

Name	Description	Scope	Control Plane Mode
Transport Zone		Global	Unicast
Universal Transport Zone		Universal	Unicast

**Figure 4.24** Local And Universal Transport Zone Being Utilized

The process of creating and replicating universal objects is described in the following two sections using examples of universal distributed firewall (UDFW) rule and universal logical switch (ULS) creation.

## UDFW Rule Creation Example

Figure 4.25 outlines the steps involved in the process of a UDFW rule creation:

1. UDFW rule is created on the primary NSX Manager,
2. The UDFW rule is stored locally in the NSX Manager database,
3. The UDFW rule is pushed down the message bus to the respective ESXi hosts in its vCenter domain. At the same time, the USS on the primary NSX Manager replicates the UDFW rule to the secondary NSX Managers,
4. Secondary NSX Managers store UDFW rule locally in the NSX Manager database.
5. The secondary NSX Managers push the UDFW rule down to their respective ESXi hosts via the message bus.
6. Similar to NSX standalone deployments, the UCC is not utilized in UDFW rule creation.

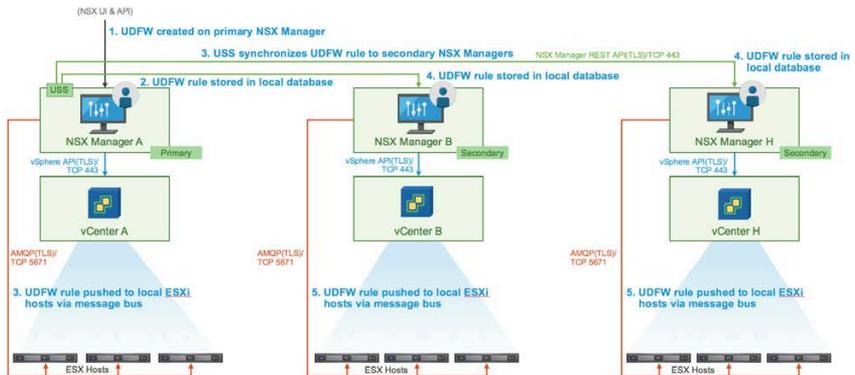


Figure 4.25 UDFW Creation Process

## ULS Creation Example

1. ULS is created on the primary NSX Manager.
2. The ULS configuration is stored locally in the NSX Manager database.
3. The ULS configuration is pushed to the UCC. At the same time, the USS on the primary NSX Manager replicates the ULS configuration to the secondary NSX Managers.
4. The secondary NSX Managers store the ULS configuration locally.

In Figure 4.26, the steps starting from the left represent the creation steps of a ULS. The steps starting from the right represent creation of a local logical switch. Universal objects can only be created from the primary NSX Manager, but local objects are created by the respective local NSX Manager. The LS configuration is also pushed to the UCC. The UCC is a shared object that stores both local objects for each NSX Manager domain and universal objects spanning all NSX Manager domains.

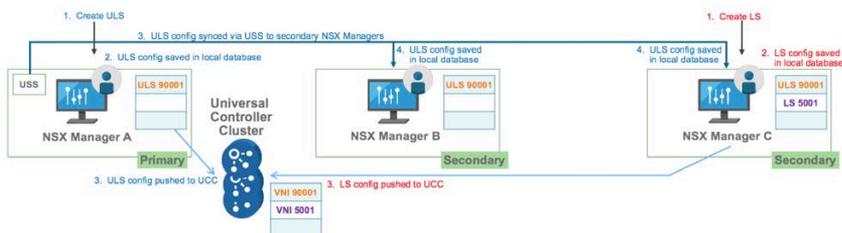


Figure 4.26 ULS Creation Process and LS Creation Process



When configuring the segment IDs (i.e., VNIs) used for local and universal logical switches, it is important to note that the local and universal segment IDs should not overlap. Since the UCC is a shared object managing both the universal and local objects for all vCenter domains, the local segment IDs for the different NSX Manager domains should also not overlap.



When deploying multiple NSX domains and considering Cross-VC NSX in the future, planning the segment IDs across NSX domains so there is no overlap can save considerable time and effort when eventually moving to a Cross-VC NSX setup.

For a Cross-VC NSX deployment in a brownfield environment where multiple NSX Manager domains already exist, the existing VNIs in each domain need to be changed/migrated to a unique range so there are no overlapping VNIs across different NSX Manager domains. See the VMware NSX-V: Multi-site Options and Cross-VC NSX Design Guide (<https://communities.vmware.com/docs/DOC-32552>) for additional details.

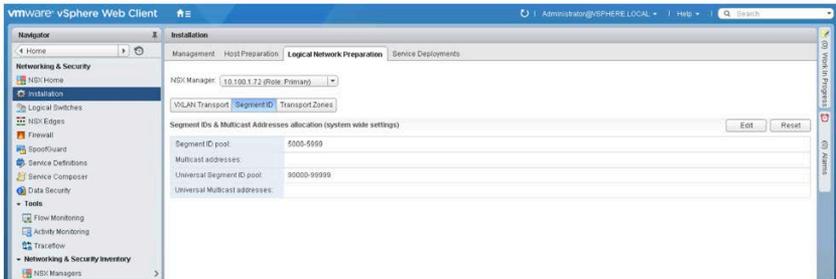


Figure 4.27 Configuring Segment IDs For Local And Universal Objects

## Controller Disconnected Operation (CDO) Mode

Controller Disconnected Operation (CDO) mode was introduced in NSX 6.3.2 and provides additional resiliency for the NSX control plane.

NSX already offers inherent resiliency for the control plane in several ways:

- Complete separation of control plane and data plane; even if the entire controller cluster is down, the data plane keeps forwarding
- A controller cluster of three nodes allows for the loss of a controller with no disruption to the NSX control plane
- vSphere HA provides additional resiliency by recovering the respective NSX Controllers on another node if the host it is running on fails

Although losing communication with the entire NSX controller cluster is unlikely, NSX versions from 6.3.2 onward further enhance control plane resiliency via CDO mode.

CDO mode targets specific scenarios where control plane connectivity might be lost but data forwarding must be maintained. Examples include a host losing control plane connectivity, loss of control plane connectivity to the controller cluster, or failure of NSX Controllers. CDO mode enhances control plane resiliency for both single site and multi-

site environments. Multi-site environments and typical multi-site solutions such as disaster recovery provide good use cases for CDO mode. The following examples look at the details of CDO mode operation, showing how it provides additional resiliency for specific scenarios.

Starting with NSX 6.3.2, CDO mode is enabled from the NSX Manager at the transport zone level. It can be enabled on a local transport zone or/and on a universal transport zone. When enabled on a universal transport zone, it must be enabled from the primary NSX Manager. Figure 4.28 shows CDO mode being enabled on a universal transport zone via the primary NSX Manager.

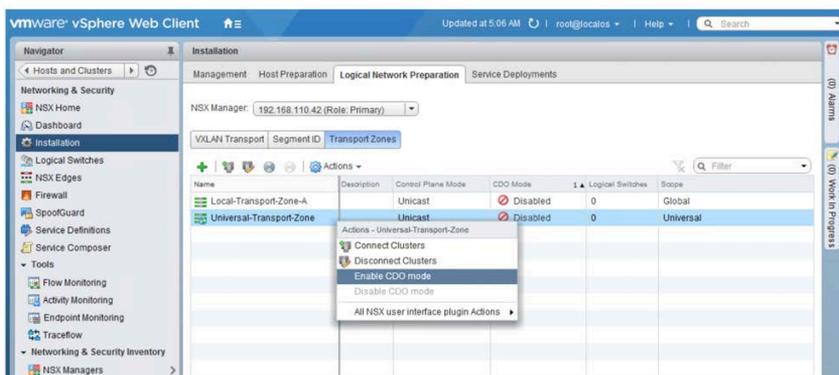


Figure 4.28 Enabling CDO Mode At Transport Zone Level

In NSX 6.4, CDO mode is configured at the NSX Manager level, allowing for enablement across multiple transport zones (**Networking & Security > Installation & Upgrade > Management tab > Select NSX Manager > Actions > Enable CDO mode**).

In the initial release of CDO mode in NSX 6.3.2, it could be enabled on multiple transport zones only when each transport zone was on a different VDS. If the VDS was shared by a universal transport zone and local transport zone, CDO could still be enabled on the universal transport zone, but not on both. This allowed for use of CDO mode on the universal transport zone where it would likely be needed for Cross-VC NSX multi-site use cases.

When CDO mode is enabled, the next available VNI is designated for the CDO logical switch. All hosts of the transport zone will join this LS. In the example of Figure 4.29, no universal logical networks have been created. In this case, the first available VNI from the Universal Segment

ID Pool that was configured when the Cross-VC NSX environment was set up is selected. For this example, the CDO logical switch VNI is 900000 since the configured Universal Segment ID Pool is 900000 – 909999.

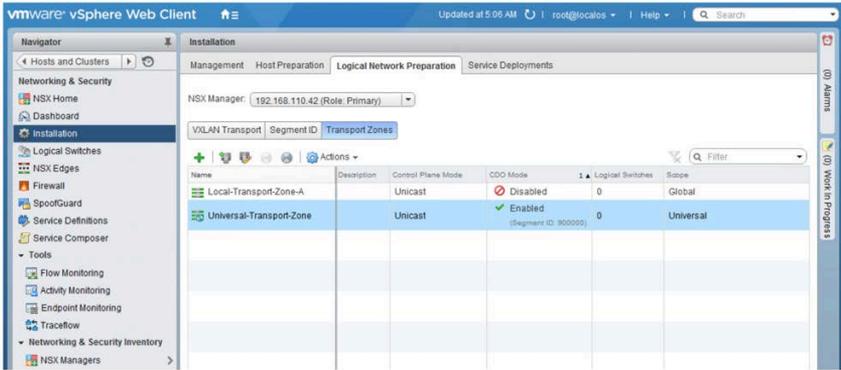


Figure 4.29 CDO Mode Enabled On Universal Transport Zone

Looking at the logical switches in the GUI, Figure 4.30 shows that there are no local or universal logical switches. Additionally, the CDO logical switch is not listed. Since the CDO logical switch is used only for control plane purposes, it is not visible under the **Logical Switches** tab.

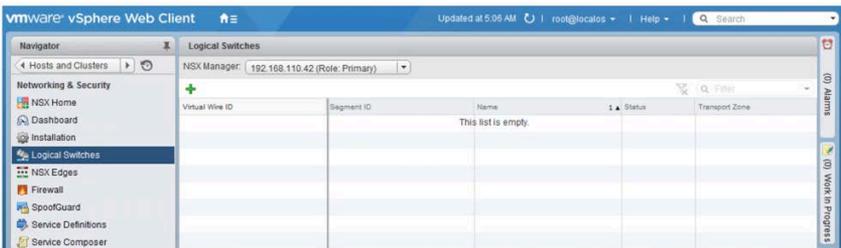
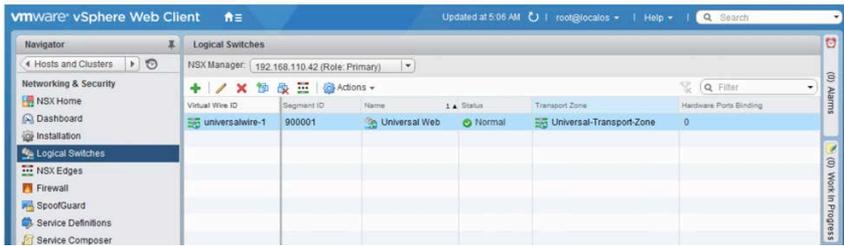


Figure 4.30 CDO Logical Switch not Visible Under 'Logical Switches' Tab

When a new logical switch is created in the universal transport zone, it will skip VNI 900000 and select VNI 900001 since VNI 900000 is being used by the CDO logical switch.



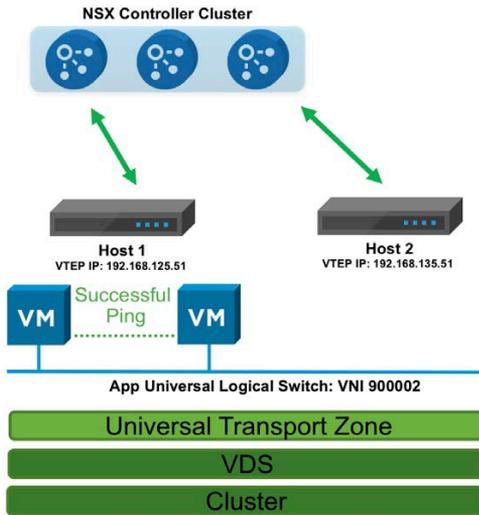
**Figure 4.31** New Universal Logical Switch Selects Next Available VNI 900001

When CDO mode is enabled on a transport zone, all hosts in the transport zone join the CDO logical switch. One controller in the cluster is given the responsibility to update all hosts in the transport zone with the VTEP information of every other host in the transport zone. Since all hosts are members of the CDO logical switch, this creates a global VTEP list that is initially populated when control plane connectivity is up. If control plane connectivity is later lost, this global VTEP list will be utilized to flood BUM messages to all hypervisors that are part of the transport zone.

If control plane connectivity was lost prior to the introduction of CDO mode, the data plane would continue to forward as expected due to the complete separation of control and data planes. However, if control plane connectivity or controllers were down and workloads on a logical switch moved to another host that was not already a member of that logical switch (e.g., VTEP of host not member of VNI), there would be data plane disruption to and from that workload. CDO mode targets this specific scenario to provide better control plane resiliency.

The following example steps through the behavior before and after CDO mode is enabled. It uses universal networks with one site for ease of demonstration/explanation.

In Figure 4.32, the NSX controller cluster is up and two VMs/workloads on host 1 are on the same universal logical switch (VNI 900002). Prior to NSX 6.3.2, or with CDO disabled, this will operate and communicate as expected. Figure 4.33 presents the VTEP table for VNI 900002 via CLI on the NSX Manager, showing that the controllers have been informed that host 1 (VTEP IP 192.168.125.51) is a member of the logical switch with VNI 900002. If other hosts had VMs/workloads on this same logical switch, the controllers would also have had their VTEP entries added to this list. This would ensure that the VTEP table for VNI 900002 was distributed to all hosts that are members of the logical switch.



**Figure 4.32** No CDO Mode, Controller Cluster Up, With Two VMs On The Same Host And Same Universal Logical Switch

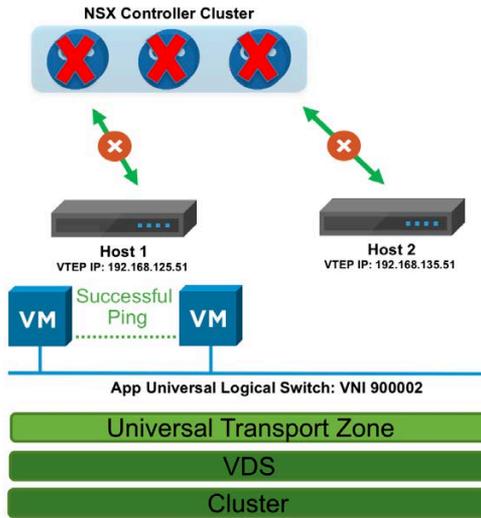
```

nexusmgr-1-01a> show logical-switch controller controller-1 vni 900002 vtep
VNI      IP           Segment      MAC           Connection-ID Is-Active
900002   192.168.125.51 192.168.125.0 00:50:56:6b:6d:41 4             YES

```

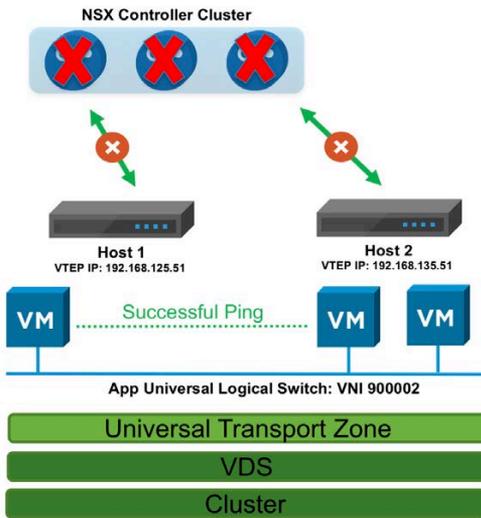
**Figure 4.33** NSX Manager Central CLI Displaying VTEP Table on Controller for VNI 900002

If control plane connectivity or the NSX Controller Cluster were to go down – as depicted in Figure 4.34 – communication between the VMs would continue. Additionally, communication with any other VMs on the same universal logical switch on other hosts would also continue to work because the NSX controllers would have already distributed the correct VTEP table information to their respective hosts.



**Figure 4.34** NSX Controller Cluster Down, Communication Between VMs On Universal Logical Switch VNI 900002 Continues To Work

Even in the scenario presented in Figure 4.35, where the two VMs communicating are on different hosts, communication would still continue to work if control plane connectivity or the NSX Controller Cluster were to go down. This works because the two hosts already had VMs on the same logical switch, so both hosts were already members of the logical switch/VNI.



**Figure 4.35** NSX Controller Cluster Down, Communication Between VMs On Universal Logical Switch VNI 900002 Continues To Work

Prior to shutting down the NSX Controllers for this example, the NSX Manager central CLI command of Figure 4.36 confirms that both host 1 (VTEP IP 192.168.125.51) and host 2 (VTEP IP 192.168.135.51) were members of VNI 900002.

```

nsxmgr-1-01a> show logical-switch controller controller-1 vni 900002 vtep
VNI      IP           Segment      MAC           Connection-ID Is-Active
900002   192.168.125.51 192.168.125.0 00:50:56:6b:6d:41 4             YES
900002   192.168.135.51 192.168.135.0 00:50:56:6f:9e:4d 7             YES

```

**Figure 4.36** NSX Manager Central CLI Displaying VTEP Table on Controller for VNI 900002

Even if the NSX Controllers are shut down, the VTEP information for the universal logical switch VNI 900002 has already been distributed to the respective ESXi hosts as shown in Figures 4.37 and 4.38. Due to this, there is no disruption to data plane communications when control plane connectivity is lost.

```
[root@esxcomp-01a:~] esxcli network vswitch dvs vmware vxlan network vtep list --vxlan-id=900002 --vds-name=ComputeA_VDS
-----
IP           Segment ID  Flags
-----
192.168.135.51 192.168.135.0 MTEP
```

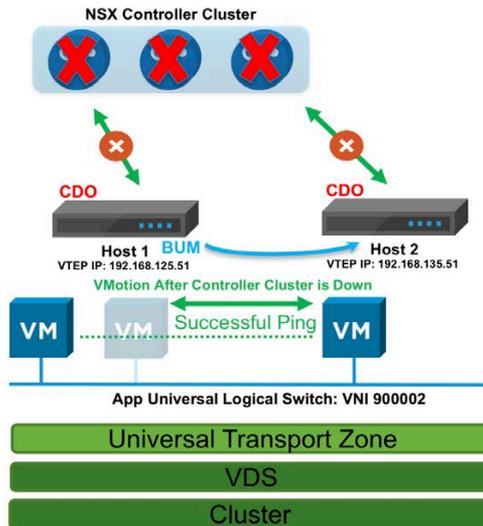
**Figure 4.37** 'Host 1' Has the VTEP Entry for 'Host 2' for VNI 900002

```
[root@esxcomp-02b:~] esxcli network vswitch dvs vmware vxlan network vtep list --vxlan-id=900002 --vds-name=ComputeB_VDS
-----
IP           Segment ID  Flags
-----
192.168.125.51 192.168.125.0 MTEP
```

**Figure 4.38** 'Host 2' has the vTEP Entry for 'Host 1' for 'VNI 900002

Similarly, if two VMs on the same universal logical switch on host 1 are communicating and control plane connectivity goes down, communication would continue even after one of the VMs moves to host 2 - either manually or automatically.

This works because another VM already exists on the same universal logical switch on host 2, so host 2 is already a member of the universal logical switch VNI 900002. Prior to loss of control plane connectivity, the host/VTEP membership information for the logical switch was distributed to all other hosts who have joint membership to that VNI. Similarly, if a new VM is added on host 2 during control plane downtime, it would also be able to communicate with the VM on host 1.

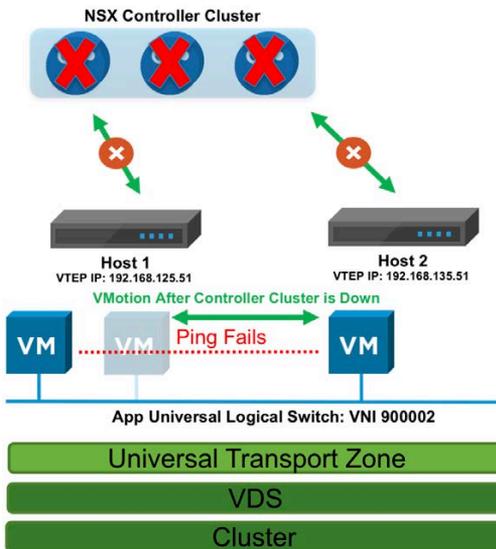


**Figure 4.39** NSX Controller Cluster Down and VM on Universal Logical Switch VNI 900002 on 'Host 1' vMotions to 'Host 2' with no Data Plane Disruption

The two specific scenarios that CDO mode targets are:

1. Provide resiliency for VM movement, either by manual intervention or automated methods (e.g., DRS), to another host that was never a member of the respective logical switch before control plane connectivity loss.
2. Allow VMs newly connected to a logical switch on a host which was not a member of the respective logical switch before control plane connectivity loss to communicate with other members of the logical switch.

In each scenario, a new host has become a member of a specific logical switch/VNI; however, since control plane connectivity is lost or controllers are unavailable, the NSX Controllers cannot be notified of the new member for the logical switch. Without CDO mode, the new logical switch membership information cannot be distributed to the other hosts. Figure 4.40 illustrates this issue.



**Figure 4.40** NSX Controller Cluster Down and VM on Universal Logical Switch VNI 900002 on 'Host 1' vMotions to 'Host 2' Causing Data Plane Disruption

With the CDO mode feature, both of these scenarios are handled by using the global VTEP list. All hosts in the CDO-enabled transport zone automatically become members of the CDO logical switch (e.g., next available VNI). When the host determines control plane connectivity is lost for the logical switch in question, the global VTEP list is leveraged and all BUM traffic is sent to all members of the transport zone.

In summary, CDO mode brings additional resiliency to the NSX control plane for specific scenarios, adding to the overall robustness of both single and multi-site solutions.



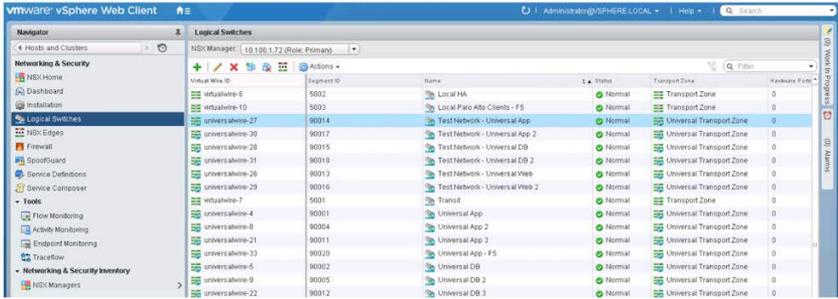
# Understanding VMware Cross-VC NSX Networking and Security

## Cross-VC NSX Switching and Routing

### **Universal Logical Switch (ULS)**

A logical switch is universal if it is deployed in the universal transport zone. Each ESXi host that is a member of the universal transport zone will have the respective ULS portgroups created on its DVS, and its VTEP(s) will be added to the VTEP table when a VM is deployed on the ULS on that host.

The VNI used for the ULS will be the next available VNI from the manually configured Universal Segment ID pool defined under the **Installation->Segment ID** tab within NSX Manager as shown in Figure 4.27. Figure 5.1 shows ULSs created in the universal transport zone and local LSs created in the local transport zone.



**Figure 5.1** ULSs Created In Universal Transport Zone And Local LSs Created In Local Transport Zone

## Universal Distributed Logical Router (UDLR)

The universal attribute of a UDLR must be selected at deployment time of a new DLR instance. Both local DLRs and universal DLRs can be deployed and co-exist within the same NSX environment. Every ESXi host that is a member of the universal transport zone and has respective universal logical switches will leverage the UDLR for routing between universal networks.

Figure 5.2 depicts a UDLR deployment. In this example it is deployed in HA mode, so that an active and standby UDLR control VM will be deployed for additional resiliency. The UDLR control VM is the control plane for the UDLR; it is what peers with the NSX ESGs at the border of the SDDC for dynamically learning and exchanging routing information. As with DLR control VMs, the UDLR control VM is a VM form factor. Different deployment models are discussed in the Cross-VC NSX Deployment section.

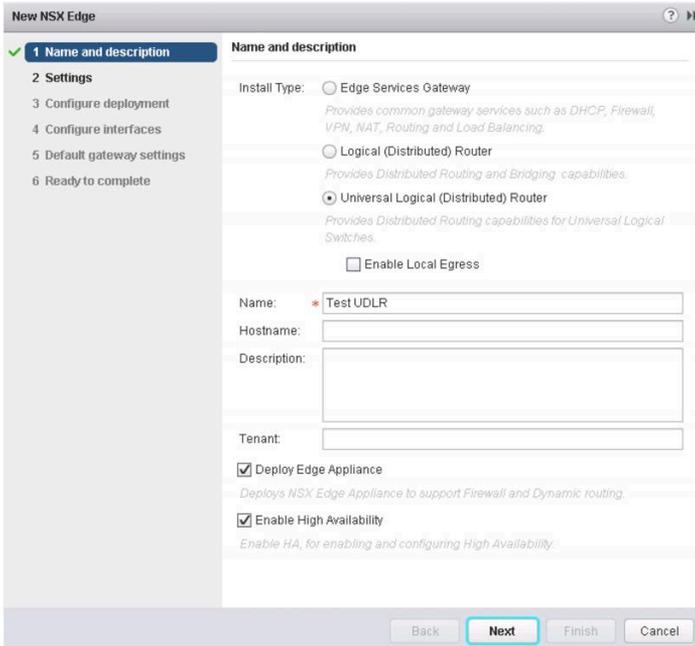


Figure 5.2 UDLR Being Deployed

ULSs can be connected to UDLRs. Select the UDLR and navigate to **Manage->Settings->Interfaces**, then create logical interfaces that connect the respective logical switches directly to the UDLR. The logical interface for the uplink connects to another transit network, typically another universal logical switch. The UDLR peers with the ESG(s) over this transit network using either BGP or OSPF. Figure 5.3 shows connected interfaces on a UDLR.

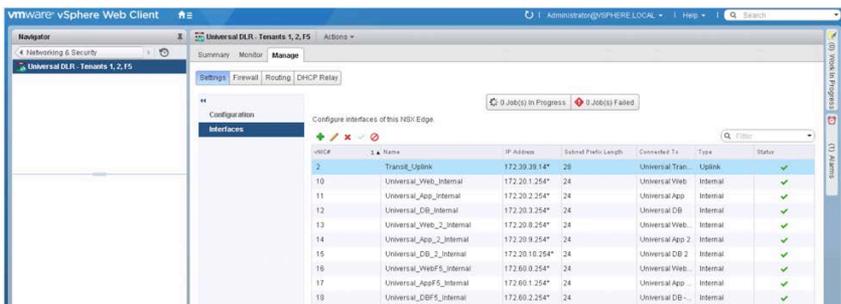


Figure 5.3 Logical Interfaces Created on The UDLR and Providing Connectivity to ULSs

Figure 5.4 shows Cross-VC NSX deployed with ULSs and a single UDLR. In this deployment, Cross-VC NSX is leveraged for a DR solution where active workloads run at primary site 1 and place-holder/standby VMs are at recovery site 2.

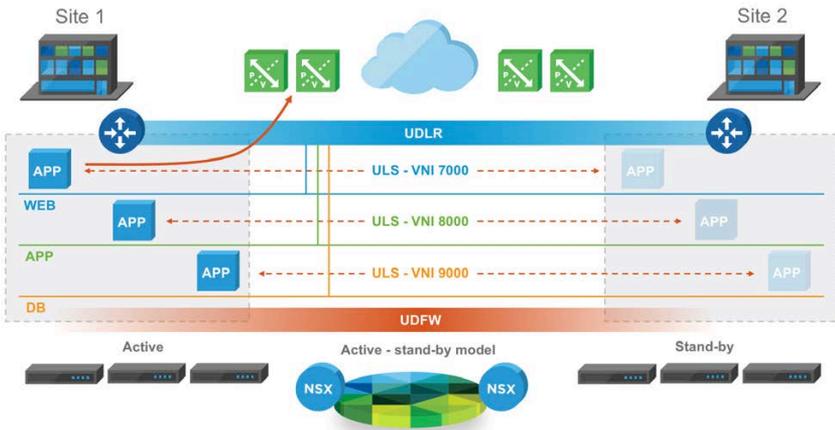


Figure 5.4 Cross-VC NSX Deployment for Disaster Recovery Solution

## Edge Services Gateway (ESG)

The NSX ESGs are the same appliances that are used by non-Cross-VC NSX deployments. There is no concept of universal ESG. ESGs peer via BGP or OSPF with the UDLR control VM and upstream physical L3 gateways. Upon learning new routes from the ESGs, the UDLR control VM forwards the best routes to the UCC, which then communicates the information to all ESXi hosts in the environment. The ESG can be deployed in HA mode with stateful services, or deployed in ECMP mode to provide multiple ingress/egress paths for both path resiliency and north/south throughput performance considerations. Up to eight-way ECMP per tenant is supported. Stateful services including Edge firewall, load balancing, and NAT are supported in HA mode only, though Edge HA can still be configured on ECMP edges. Deployment models are discussed in more detail in the [Cross-VC NSX Deployment](#) section. Figure 5.5 demonstrates a Cross-VC NSX deployment with ESGs in HA mode.

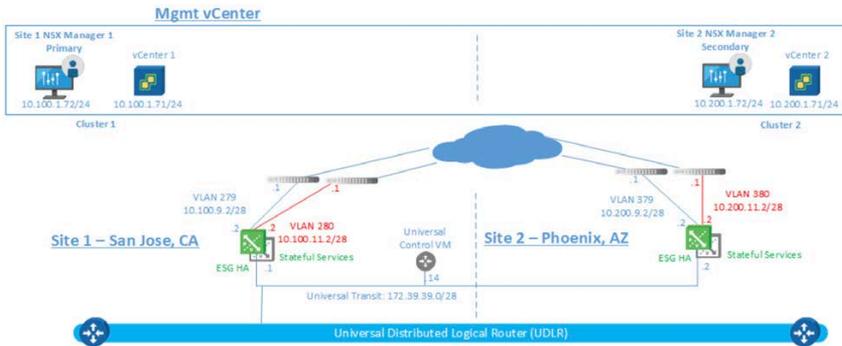


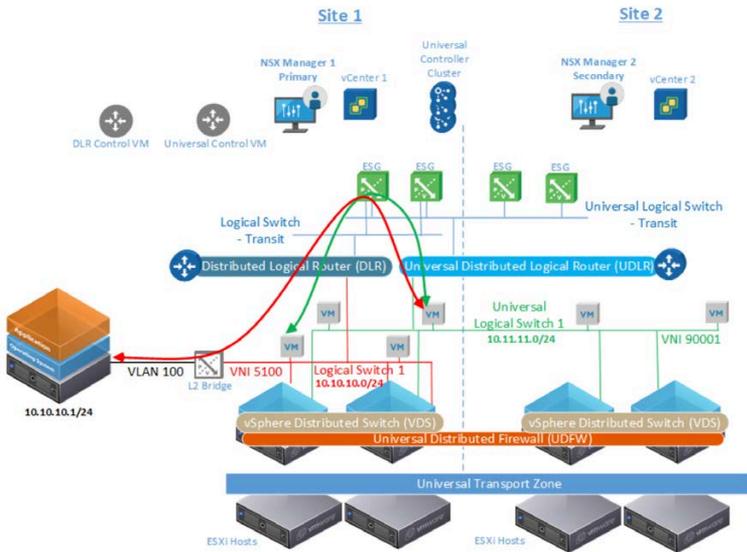
Figure 5.5 Cross-VC NSX Deployment with ESGs in HA Mode

## Cross-VC NSX L2 Bridging Between Logical and Physical Network

In addition to routing, bridging to a local logical switch can be used for communication between workloads on physical and logical networks. As of NSX 6.4, support for L2 bridging between logical switches and physical VLANs is not available with universal logical switches. Figure 5.6 shows how a physical workload that needs to bridge with virtual workloads would do so on a local DLR bridge. Routing from the logical switches to workloads on other universal logical switches is then implemented by connecting the local and universal DLRs via an ESG.



This design leverages the fact that ESGs can connect to both local and universal logical switches.



**Figure 5.6** NSX L2 Bridge Used For Communications Between Physical and Virtual Workloads

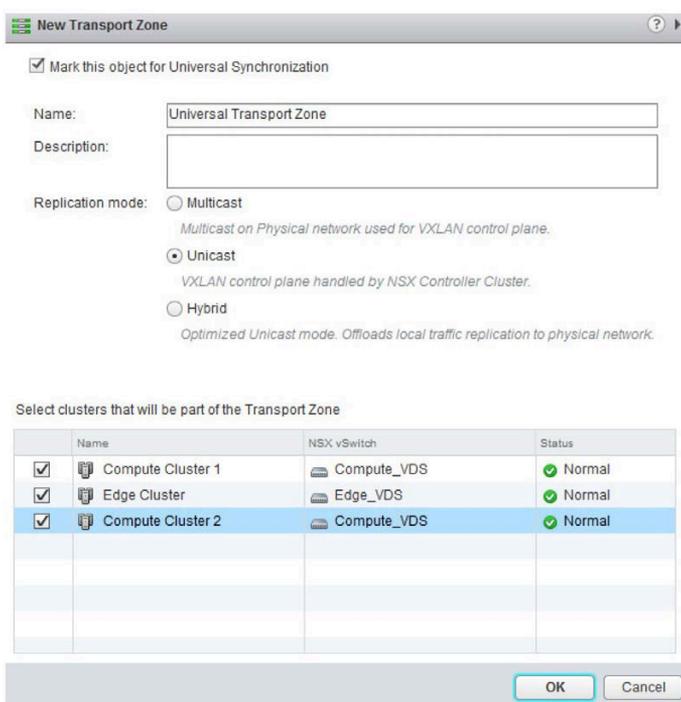
## Handling Broadcast, Unknown Unicast, Multicast (BUM) traffic

There is no difference in the handling of multi-destination traffic (e.g., BUM traffic) in a Cross-VC NSX deployment compared to a non-Cross-VC deployment. In both cases, there are three BUM replication modes – unicast, hybrid, and multicast – which can be set at either the transport zone level or the logical switch level. Setting BUM at the LS level allows override of transport zone-level replication settings on a per-LS basis.

Figure 5.7 shows the option to mark the transport zone as universal and select the desired replication mode. The replication options are the same as when creating a local transport zone. Once a universal transport zone is created, future transport zone creation will not display the checkbox to make it universal as only one universal transport zone is permitted. Multiple local transport zones can be created for local objects that can co-exist with the universal transport zone.

In a Cross-VC NSX deployment, the UCC informs the ESXi hosts across all vCenter domains of the respective VTEPs via the control plane. This control plane connection consists of a user world agent (netcpa) on each ESXi host, which connects to the UCC with TLS over TCP. Further details are described in the Architecture and Key Concepts section. The UCC is a shared object used by both local and universal objects. It only exists within the vCenter domain linked to the primary NSX Manager.

The three different replication modes will only be described briefly; a more detailed explanation is provided in the VMware NSX Design Guide (<https://communities.vmware.com/docs/DOC-27683>).



**Figure 5.7** Creating A Universal Transport Zone And Selecting The Replication Mode

Similarly, to local logical switches in a local transport zone, universal logical switches are deployed in a universal transport zone. The same replication mode options exist at the universal logical switch level, as shown in Figure 5.8. The transport zone's replication mode is selected by default; this can be overridden at the universal logical switch level.



Best practices recommend keeping the default replication mode inherited by the universal transport zone. It is possible to create logical switches leveraging hybrid or multicast replication modes while continuing to use the default unicast replication mode for all other logical switches. This may be desired for applications on specific subnets with large volumes of broadcast or multicast traffic.

**New Logical Switch**

Name: \*

Description:

Transport Zone: \*  [Change](#) [Remove](#)

Replication mode:  Multicast  
*Multicast on Physical network used for VXLAN control plane.*

Unicast  
*VXLAN control plane handled by NSX Controller Cluster.*

Hybrid  
*Optimized Unicast mode. Offloads local traffic replication to physical network.*

Enable IP Discovery

Enable MAC Learning

**OK** **Cancel**

**Figure 5.8** Deploying a Universal Logical Switch

## ARP Suppression

Most BUM traffic on a network consists of ARP traffic. As ARP traffic is handled efficiently by the NSX Controllers, most BUM traffic is eliminated in the data center. An NSX Controller populates its local ARP table as VMs power-on through the following process. ARP requests sent by VMs are intercepted by the local ESXi host. The controller cluster is queried for the MAC address, and if present (e.g., destination is a VM which MAC is known to the controllers) the information is sent back to the originating ESXi host, then forwarded to the VM as if the destination VM had responded directly. This ARP suppression mechanism reduces the need to flood ARP broadcast requests across the fabric and L2 broadcast domains.



Due to this ARP suppression functionality, the majority of BUM traffic is removed entirely. The BUM replication modes described in the following sections are utilized mostly for non-ARP traffic.

## BUM Replication Modes

### Unicast

Unicast mode leverages the UCC to allow BUM replication without the need for specific configuration on core switches to be implemented, as is the case with multicast or hybrid modes. Logical and physical networks are completely decoupled in unicast mode as there is no reliance on the physical switches to assist with replicating BUM traffic

In unicast mode, the source ESXi host replicates traffic locally and unicasts all packets to participating VTEPs according to information stored in the VTEP table for that VNI segment.

For VTEPs on different segments, a Unicast Tunnel End Point (UTEP) is selected by the source ESXi host. Unicast traffic is sent to the remote segment's UTEP with the **REPLICATE\_LOCALLY** bit set in the VXLAN header. The remote segment's UTEP then replicates the BUM traffic locally on its L2 segment via unicast.

Each UTEP will replicate traffic only to ESXi hosts that have at least one VM actively connected to the logical network. This optimizes performance by sending BUM traffic only to hosts that have active workloads for the respective logical network. It further implies that traffic will only be sent by the source ESXi node to the remote UTEPs if there is at least one active VM connected to an ESXi host in that remote segment.

Unicast mode is suitable for most deployments due to its scalability and avoidance of physical network dependencies. If there is a large amount of broadcast or multicast traffic generated by applications, or if L2 segments for the transport network are large, hybrid or multicast mode can be used to avoid excessive CPU overhead imposed by the headend replication of BUM packets.

## Hybrid

Hybrid mode is preferred over multicast mode since it only requires L2 multicast configuration on physical switches as opposed to configuration for both L2 and L3. With hybrid, both unicast and multicast traffic are leveraged. ARP suppression is not available in multicast/hybrid modes.

Hybrid mode leverages L2 multicast for BUM replication that is local to the L2 transport segment and unicast across VTEP subnets. This requires L2 multicast configuration on the ToR switches. Each logical switch must also be associated with a multicast group within NSX, with the multicast address consumed from the range defined during NSX configuration. **IGMP snooping** is required on the ToR switches. **IGMP querier** is highly recommended to optimize the handling of multicast traffic on the switches and trigger periodic IGMP report messages. In hybrid mode, the ESXi hosts send an IGMP join when there are local VMs interested in receiving multi-destination traffic.

For VTEPs on different segments, a multicast tunnel endpoint (MTEP) is selected by the source ESXi node. Unicast traffic is sent to the MTEP with the **REPLICATE\_LOCALLY** bit set in the VXLAN header. The MTEP replicates the BUM traffic locally on its L2 segment via L2 multicast.

Hybrid mode is recommended for deployments where applications generate a large amount of broadcast or multicast traffic, or where the L2 segments for the transport network are large.

## Multicast

With multicast replication mode, NSX offloads the replication of BUM traffic to the physical switches. L2 multicast is used for local transport segments while L3 multicast (via PIM) is used for VTEPs on different segments.

Multicast mode is the process for handling BUM traffic as specified by IETF RFC 7348 for VXLAN. It does not leverage the NSX Controllers and the decoupling of logical and physical networking is not a factor since communication in the logical space is predicated on the multicast configuration required in the physical network infrastructure. In this mode, as in hybrid mode, multicast IP addresses are associated to each logical switch.

To ensure multicast traffic is delivered to VTEPs in a different subnet from the source VTEP, the network administrator must configure PIM and enable L3 multicast routing.



Similarly to not overlapping local segment IDs and universal segment IDs, there should not be any overlap between the multicast addresses used for universal logical switches and those used for local logical switches. This would create unnecessary BUM traffic across local and universal transports and would be inefficient in a multi-vCenter deployment.

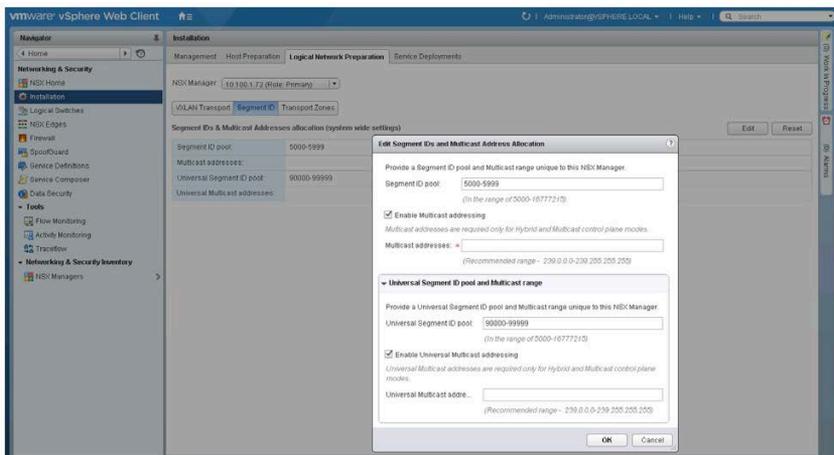


Figure 5.9 Enabling Multicast Addressing

Consideration must be given on how to perform the mapping between VXLAN segments and multicast groups. The first option is to perform a 1:1 mapping. This has the advantage of delivering multicast traffic in a granular manner; a given ESXi host would receive traffic for a multicast group only if at least one local VM is connected to the corresponding multicast group. Conversely, this option may significantly increase the amount of multicast state required in physical network devices. Understanding the maximum number of groups those platforms can support is critical.

The other choice involves leveraging a single multicast group for all the defined VXLAN segments. This dramatically reduces the amount of multicast state information the transport infrastructure has to keep track of, but may cause unnecessary traffic to be received by the ESXi hosts.

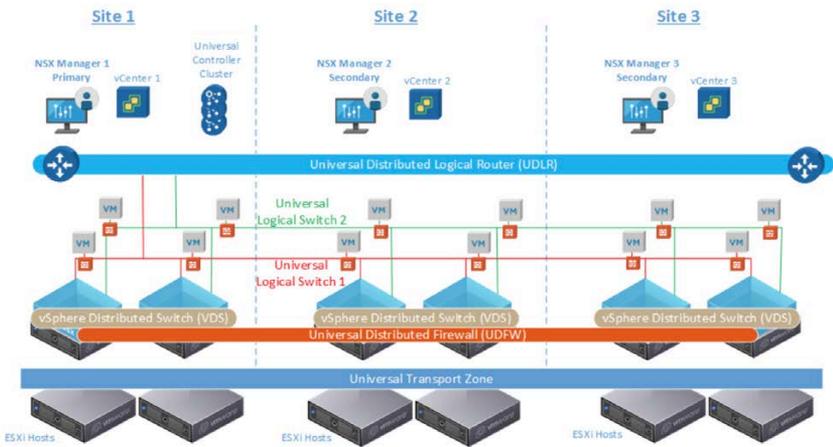
The most common strategy involves the decision to deploy a m:n mapping ratio as a trade-off between these two options. With this configuration, every time a new VXLAN segment is instantiated, up to

the maximum specified “m” value, it will be mapped to a multicast group part of the specified range in a round robin fashion. VXLAN segment “1” and “n+1” will be using the same group. Segment “2” and “2+n” will be using the same multicast group, the groups used by segment “1” and “2” being different. Care must still be taken to ensure multicast addresses are not mixed between local and universal segments.

Multicast mode is an available legacy option as defined by IETF RFC 7348 for VXLAN. It is typically not required or desired due to complete dependency on complex physical switch configuration for L3 multicast and the lack of optimization brought by the NSX Controllers.

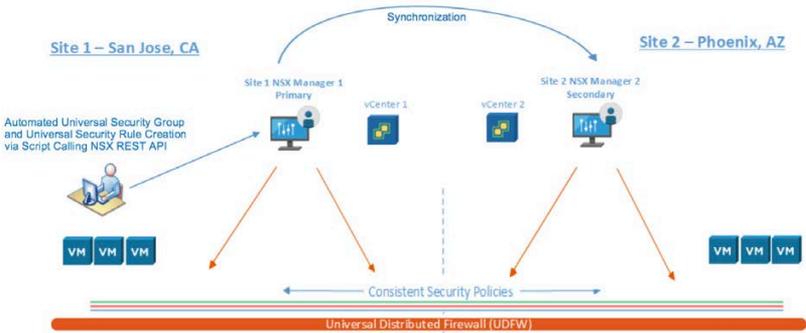
## Cross-VC NSX Security

With Cross-VC NSX it is also possible to create universal security rules that span multiple vCenter domains and/or sites. UDFW rules are configured from the primary NSX Manager, similar to any other universal object. As with networking, an end user can configure consistent security policies across multiple vCenter domains/sites from a single centralized point.



**Figure 5.10** Consistent Security Policies Across vCenter Domains with Cross-VC NSX

Furthermore, the NSX REST API offers a single point of configuration for a custom script/orchestration platform to leverage, enabling the automation of consistent security policies across vCenter boundaries as shown in Figure 5.11.



**Figure 5.11** Leveraging NSX REST API at Primary Site to Get Consistent Security Across Sites

This specific scenario of automating security policies through NSX is discussed in more detail in the VMware NSX Network Virtualization Blog post: Automating Security Group and Policy Creation with NSX REST API (<https://blogs.vmware.com/networkvirtualization/2016/04/nsx-automating-security-group.html/>).

Consistent security across vCenter domains/sites is critical for solutions such as disaster recovery and active/active architectures where flows can traverse sites. Figure 5.12 demonstrates how UDFW is leveraged to provide micro-segmentation across vCenter domains when a workload in a 3-tier app is moved from site 1 to site 2, maintaining the enforcement of a block rule for app tier to web tier traffic.



**Figure 5.12** UDFW Providing Micro-segmentation Across vCenter Boundaries

## DFW Sections

Cross-VC NSX creates a **Universal Section** for the universal rules within the DFW section of NSX. Starting from NSX 6.3, multiple universal sections can be created and coexist with the local sections. The universal sections contain universal distributed firewall rules whereas the local sections contain local firewall rules.

Multiple universal sections provide efficiency in terms of the following:

- Rules can easily be organized per tenant or application.
- If rules are modified within a universal section, only UDFW rules for that section are synced to the secondary NSX Managers and only rule updates for that section are published to the hosts.

Figure 5.13 displays both universal and local sections within the DFW configuration section on the primary NSX Manager.

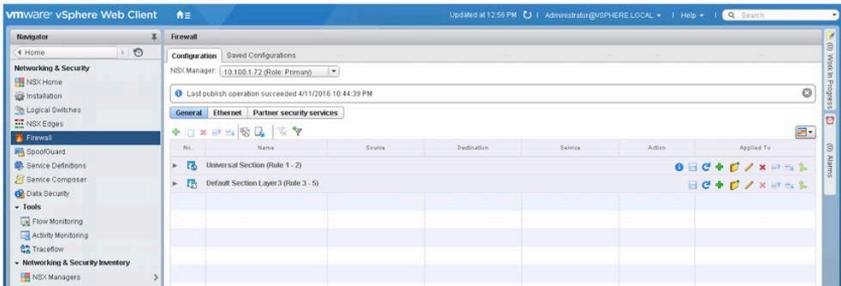


Figure 5.13 Universal Section within DFW Configuration



By default, the universal sections are on top of any local sections, so the universal firewall rules will always be applied before any local rules. It's also important to note that the universal section firewall rules span NSX Manager domains.



On secondary NSX Managers, the universal sections will also always remain on top of all local sections. This is important to keep in mind during the design/configuration stage because the universal sections will always have their security policies applied first. The security policies within sections are applied in sequential order from top to bottom, so if traffic is denied by a DFW rule in a universal section, it will never hit the DFW rules of the local sections.

Figure 5.14 shows a new DFW section being added with a right click and selection of **Add section**.

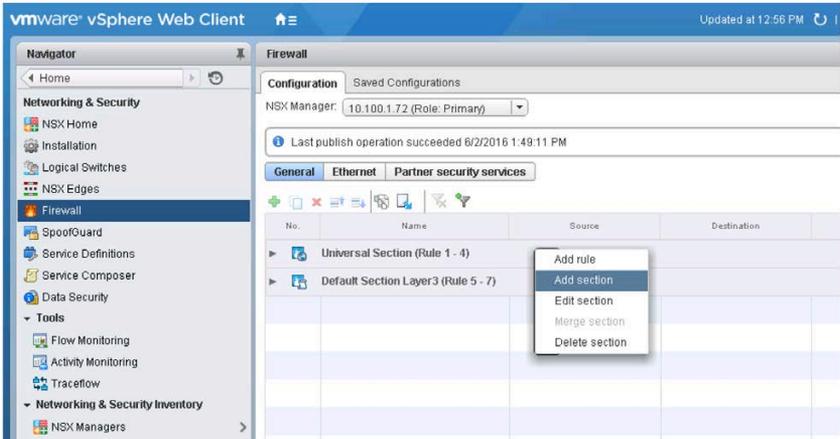


Figure 5.14 Adding new UDFW section

Figure 5.15 shows the selection of the secondary NSX Manager from the NSX Manager drop down box to confirm that there is no change on the secondary NSX Manager's DFW rule configuration. This example also shows that the universal section remains on top.

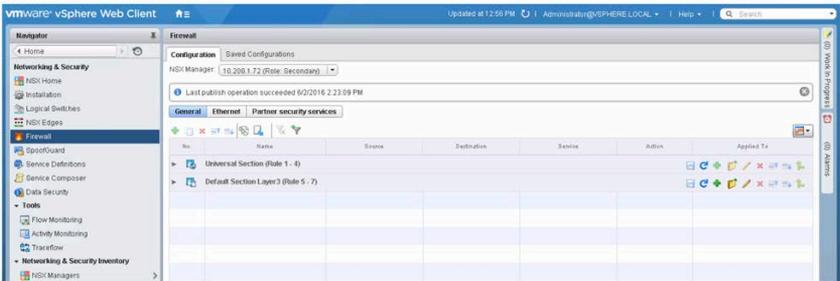


Figure 5.15 Universal Section Always On Top On Secondary NSX Managers

## UDFW Rule Objects

The following grouping objects can be used within the UDFW security policies:

- Universal IP Sets
- Universal MAC Sets
- Universal Security Groups containing universal IP Sets or universal MAC Sets
- Universal Service
- Universal Service Groups
- Universal Security Groups based on Universal Security Tags for active/standby scenarios
- Universal Security Groups based on VM name match for active/standby scenarios



Most vCenter objects cannot be used in universal firewall rules (e.g., ESXi hosts, vSphere cluster), because each vCenter has a 1:1 relationship with an NSX Manager and its local objects are not identifiable by other NSX Managers, they cannot be used as source or destinations in universal distributed firewall rules.

Universal security tags (USTs) and matching based on VM name were introduced in NSX 6.3. They can be used as membership criteria for universal security groups in active/standby scenarios only; this matching criteria is not supported for policies on applications which span sites due to the vCenter/NSX Manager relationship. For active/active deployments, UDFW rules should use universal MAC and IP Sets as well as universal security groups referencing universal MAC and IP Sets.

Figure 5.16 shows a deployment where USTs are leveraged in UDFW rules. One application is active only at site 1 and the other application is active only at site 2.

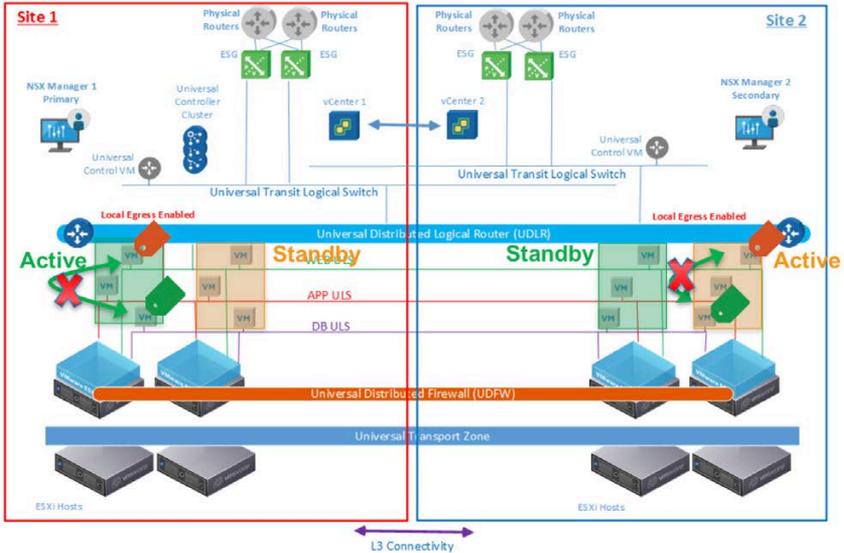


Figure 5.16 Application Must be Entirely at Site 1 or Site 2 to Leverage New Matching Criteria

Other vCenter objects and NSX objects can be leveraged in local firewall rules, but for Cross-VC domain traffic flows, the UDFW only support the aforementioned grouping objects.

Creation of universal grouping objects is performed in the **NSX Manager > Manage > Grouping Objects** section.

Figure 5.17 shows the creation panel for a universal IP Set. Ensure the **Mark this object for Universal Synchronization** checkbox is selected. This would apply for any other grouping object being created, such as a universal security group. If security groups are created through Service Composer, the checkbox to make the object universal will be missing, as universal objects cannot be created through the Service Composer. Similarly, the checkbox to make the object universal will also be missing if creating any grouping objects via secondary NSX Managers, as universal objects can only be created from the primary NSX Manager.

**New IP Set** [?] [X]

Name: \* Web IP Set

Description:

IP Addresses: \* 172.20.1.1-172.20.1.253

eg:192.168.200.1,192.168.200.1/24,  
192.168.200.1-192.168.200.24

Enable inheritance to allow visibility at underlying scopes

Mark this object for Universal Synchronization

[OK] [Cancel]

**Figure 5.17** Creating A Universal IP Set

Figure 5.18 displays three universal IP Sets created on the primary NSX Manager.

Security Group	Name	Details	Scope	Inheritance
IP Sets	App IP Set	172.20.1-172.20.2.253	Universal	[X]
MAC Sets	CD IP Set	172.20.1-172.20.2.253	Universal	[X]
Service	Web IP Set	172.20.1-172.20.1.253	Universal	[X]
Service Groups				
IP Pools				

**Figure 5.18** Three Universal IP Sets Created on The Primary NSX Manager

Since these IP Sets were marked as universal, they are synchronized by the USS to the secondary NSX Managers upon creation. Figure 5.19 displays the same IP Sets on the secondary NSX Manager.

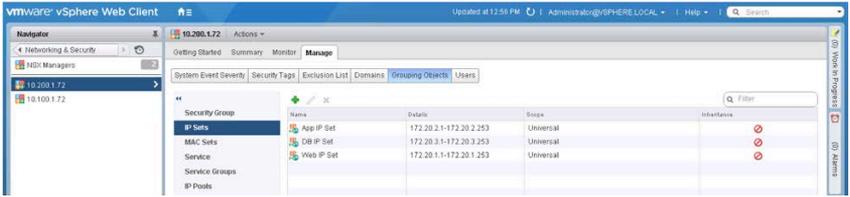


Figure 5.19 Universal IP Sets Synced by USS To Secondary NSX Managers

These universal IP Sets can now be included as part of a universal security group. Figure 5.20 displays the universal **Web IP Set** added as a member of the universal **Web Security Group**.

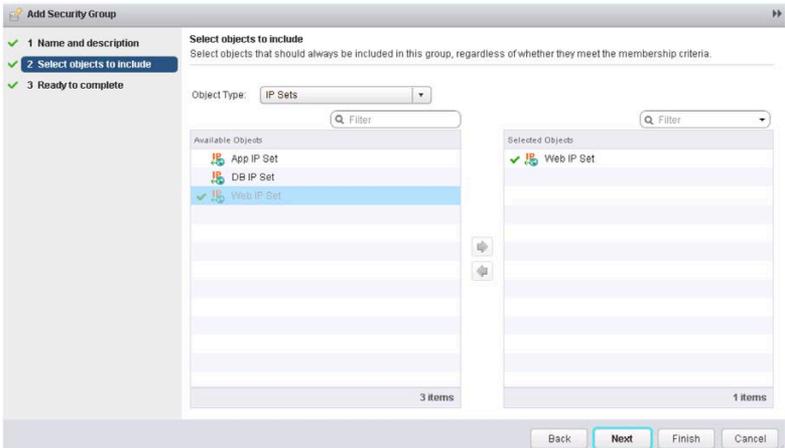


Figure 5.20 Including Universal 'Web IP Set' as Part of Universal 'Web Security Group'

Figure 5.21 displays three universal security groups created with respective universal IP Sets as members.

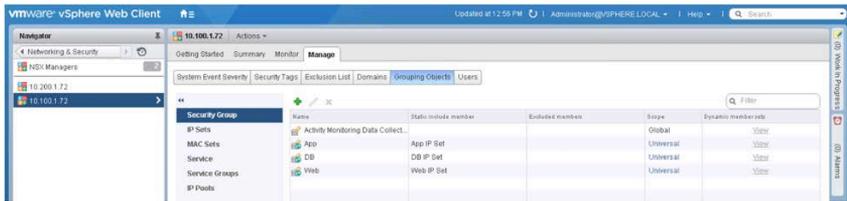
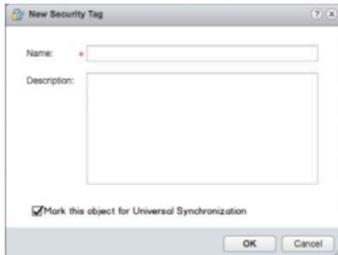


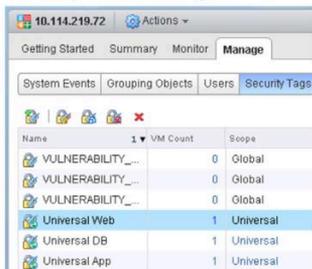
Figure 5.21 Three Universal Security Groups Created For Respective Universal IP Sets

Starting with NSX-V 6.3, Cross-VC NSX supports universal security tags. When a universal security tag is created on the primary NSX Manager, it is automatically synchronized to the secondary NSX Manager(s) as shown in Figure 5.22.

Create Universal Security Tag on Primary NSX Manager



Primary NSX Manager



Secondary NSX Manager

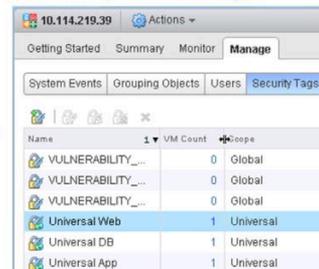


Figure 5.22 Creation and Sync of Universal Security Tags

Matching by USTs or VM name is more applicable to active/standby use cases such as disaster recovery. When a VM is vMotioned or recovered at a secondary site, the secondary NSX Manager needs a method of attaching the correct security tag to the correct VM. The **Unique ID Selection Criteria** set under **Installation->Management-**

>**Actions** on the primary NSX Manager is used to achieve this, as shown in Figure 5.23.

For the **Unique ID Selection Criteria**, **Virtual Machine Instance UUID** is typically sufficient for cases such as vMotion or DR with SRM. Virtual Machine Instance UUID is guaranteed to be unique for each VM within a specific vCenter. Although it is not guaranteed to be unique across different vCenters, it is rare for these to overlap since each Virtual Machine Instance UUID is also based on a unique vCenter ID. vSphere replication and vMotion will both maintain the Virtual Machine Instance UUID of the VM at the recovery/secondary site. Different or multiple selection criteria can also be used based on environmental specifics. If each VM in the environment has a unique VM name, that attribute can also be used for the unique selection criteria to ensure the correct security tags are attached to the correct VMs.

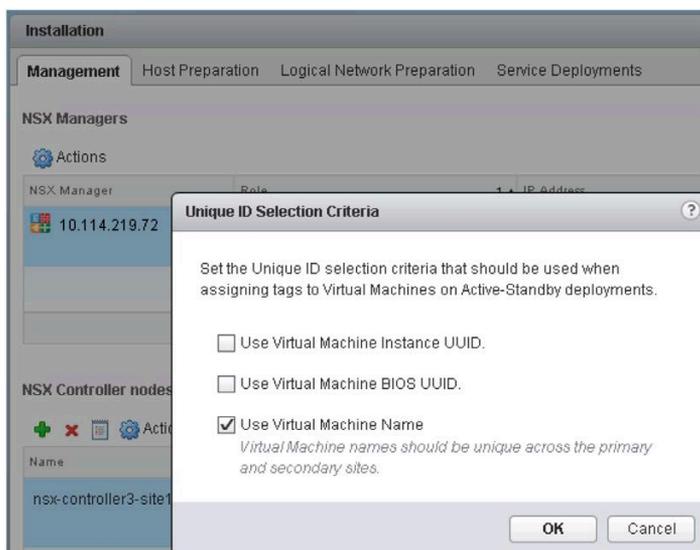


Figure 5.23 Setting Unique ID Selection Criteria on Primary NSX Manager

Figure 5.24 shows the flow for utilizing universal security tags in active/standby deployments.

1.) Set Unique Selection Criteria on the Primary NSX Manager



2.) Create Universal Security Tag on Primary NSX Manager and Attach to Respective VM



4.) Local NSX Manager Automatically Attaches Universal Security Tag to correct VM at Secondary Site Based on Unique Selection Criteria (Upon DR event or vMotion)



3.) Automatic Synchronization of Universal Security Tags between Primary/Secondary NSX Managers

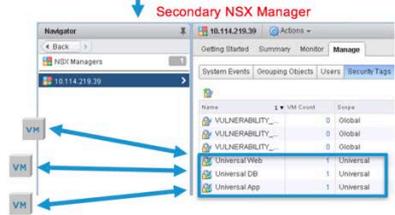


Figure 5.24 Flow for Utilizing Universal Security Tags in Active/Standby Deployments

When creating a security group, two checkboxes, **Mark this object for Universal Synchronization** and **Use for active standby deployments**, must be marked to use the matching criteria of USTs and VM name within universal security groups. These will then be used in universal security policies when creating the security group. This is shown in Figure 5.25.

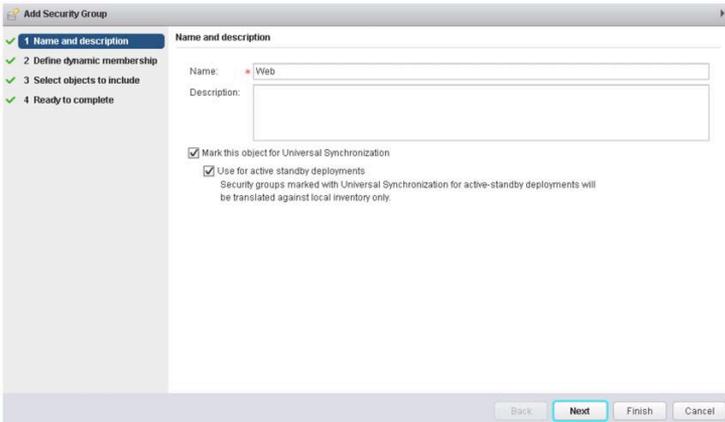


Figure 5.25 Creating Universal Security Group Which Can Match on UST and VM Name

Figure 5.26 and 5.27 show how one can choose members based on VM name using the dynamic membership selection criteria while static membership criteria are used for selecting membership based on Security Tag, Mac Set, and IP Set.

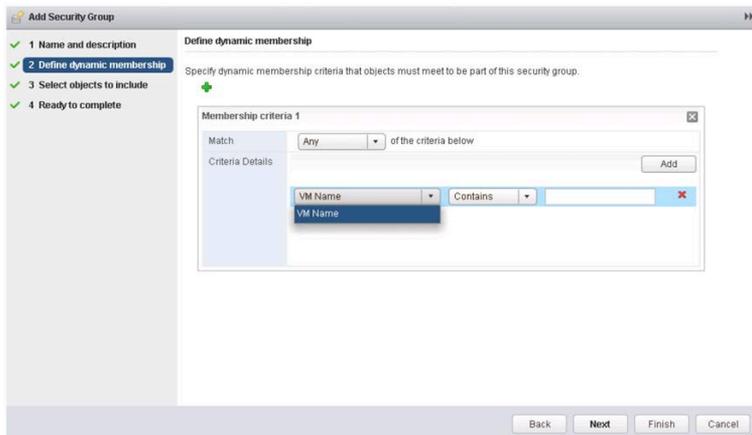


Figure 5.26 Within USG, Selecting VM Name for Matching Criteria

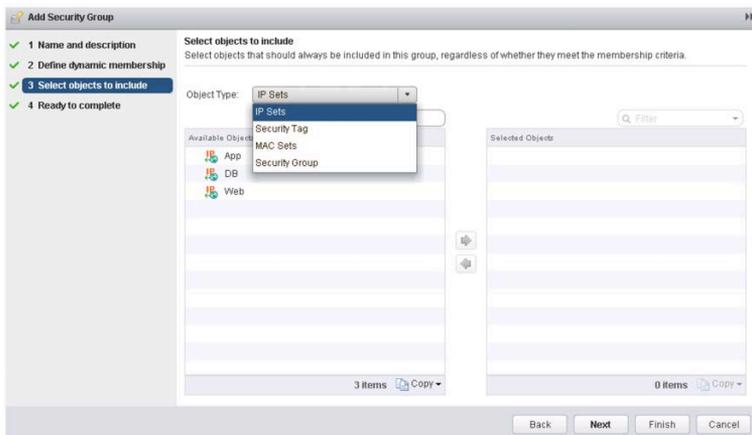
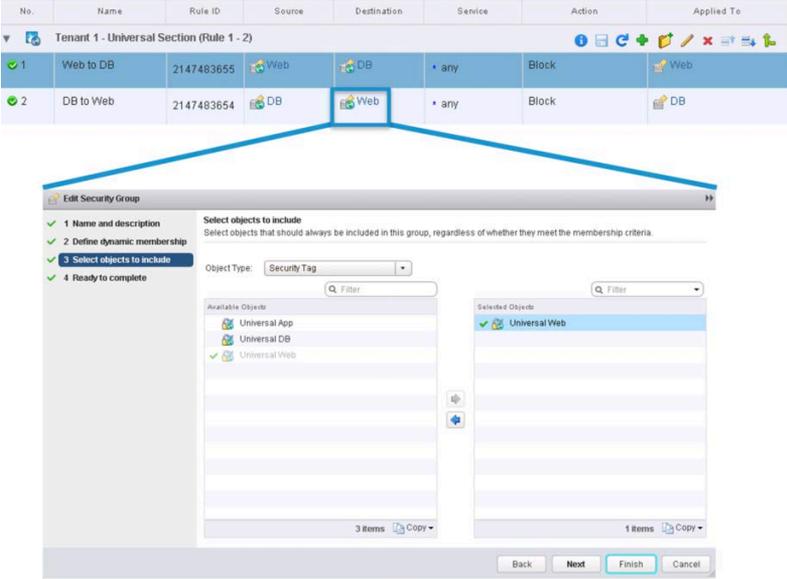


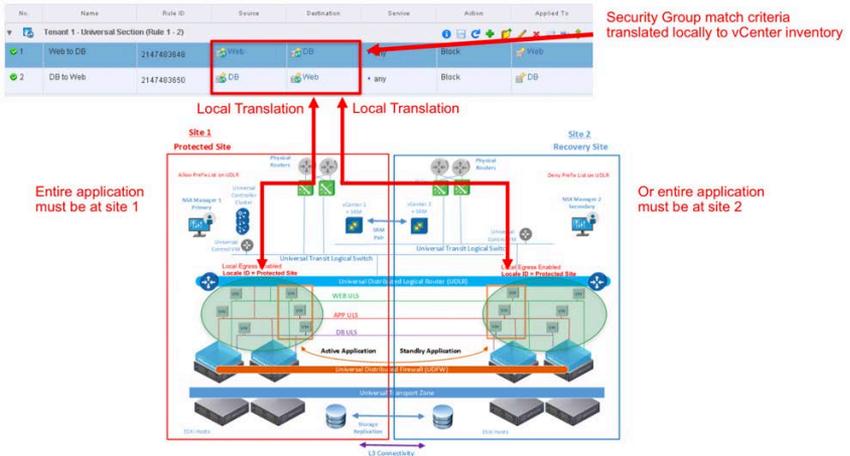
Figure 5.27 Within USG, Selecting Universal Security Tag for Matching Criteria

Figure 5.28 shows the USG with matching criteria of UST used in a universal security policy.



**Figure 5.28** USG with Matching Criteria of UST Being Used in a Universal Security Policy

Enhancements of matching criteria of UST and VM name are applicable primarily for active/standby use cases such as DR. Since the local NSX Manager does not have visibility into the inventory of the other NSX Managers' vCenters, only local VMs/workloads will be found as members of the security groups when matching universal security tags or VM names. When leveraging universal security groups with supported matching criteria, the entire application must be at the same site as shown in Figure 5.29. If the application spans sites and there are Cross-VC traffic flows, the security policy for the application will not provide the desired results.



**Figure 5.29** Entire Application Must be at Same Site When Using UDFW Rules Leveraging USGs



For management purposes and avoidance of rule sprawl, it is recommended to leverage objects such as USGs to group multiple objects that will have the same security policies. A universal security group can contain multiple universal IP Sets, universal MAC Sets, universal security tags, etc. Additionally, synchronizing a change in rule membership has lower overhead than synchronizing rule changes.

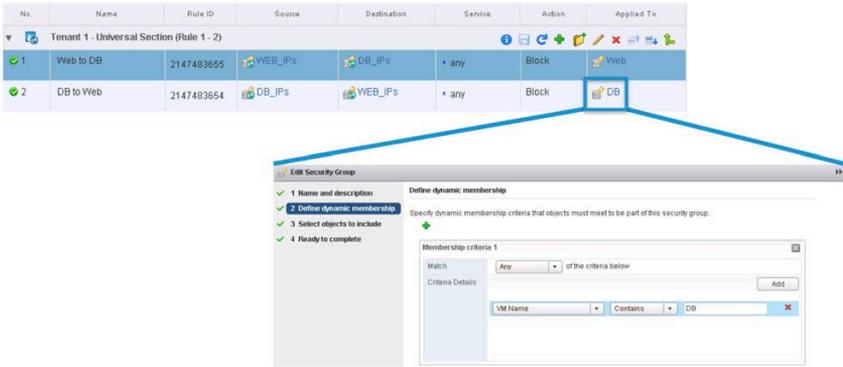
## Apply To

**ApplyTo** is a critical feature used to efficiently apply security policies to only the VMs/workloads that require the policy, aiding performance and scalability. ApplyTo also works with UDFW; however, prior to NSX-V 6.3, UDFW could only use universal logical switch with ApplyTo. This behavior had two limitations:

The most granular level a universal security policy could be applied to is at the universal logical switch level. All VMs/workloads on the universal logical switch would inherit the security policy applied to its vNIC.

If a user wanted to deploy security/micro-segmentation without deploying network virtualization, they were not able to leverage ApplyTo since it only worked with the universal logical switch. With NSX-V 6.3, ApplyTo can leverage USGs with the dynamic matching criteria of VM name or static matching criteria of universal security tag. This allows for more granular application of universal security policies at the VM level. Additionally, users can leverage

ApplyTo even in situations where network virtualization is not utilized (i.e., an NSX deployment only for the benefits of Cross-VC NSX multi-site security). Since ApplyTo is always applied locally, even when using USGs with UST or VM name, it can work in both active/standby and active/active deployments where applications are spanning multiple sites.



**Figure 5.30** Leveraging Universal Security Group with ApplyTo in UDFW

ApplyTo will work across vCenter domains with overlapping IPs. Each NSX Manager knows which VNI is tied to which respective backend local port-group and can apply the security policy to VMs on the respective logical switch.



For multi-tenant scenarios with overlapping IPs, ApplyTo can be utilized to apply certain security policies to specific tenants. In an example where tenant 1 workloads are on a set of universal logical switches (e.g. Universal\_Web, Universal\_App, and Universal\_DB), and tenant 2 workloads are on their own set of logical switches but use IPs that overlap with those of tenant 1, the ApplyTo feature allows applying security policies to the respective tenant workloads regardless of the overlapping IPs. If there are no overlapping IPs, ApplyTo is not required but still recommended.



ApplyTo is highly recommended as it efficiently applies security policies to only those VMs that require the policy instead of applying the security policy to all VMs on all hosts.

## SpoofGuard

SpoofGuard can be utilized in a Cross-VC NSX environment, though SpoofGuard information will not migrate with the VM since it is stored via local NSX Manager. New SpoofGuard information will be associated with the VM at the new NSX Manager domain. The associated SpoofGuard information is only relevant within the respective local NSX Manager domain.

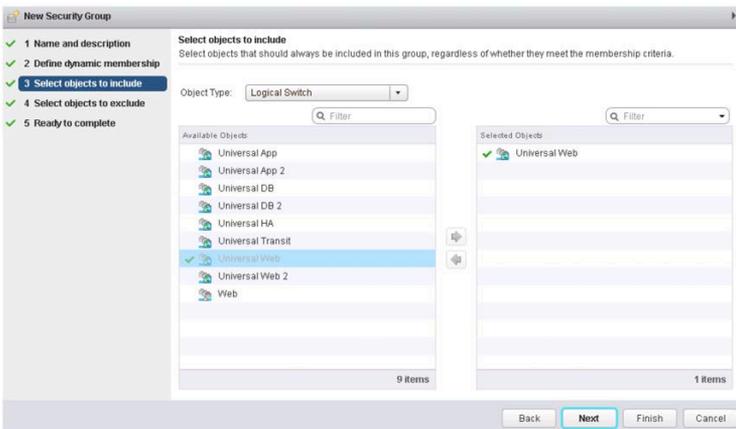
## Service Composer

Service Composer cannot be used for creating universal security groups or policies for Cross-VC traffic flows.



Service Composer can be used to create local security groups to dynamically identify local workloads. The same can also be done in the DFW configuration. This allows local security policies to be applied to workloads on universal networks. The configuration must be manually created on both the primary and secondary sites since the scope is local.

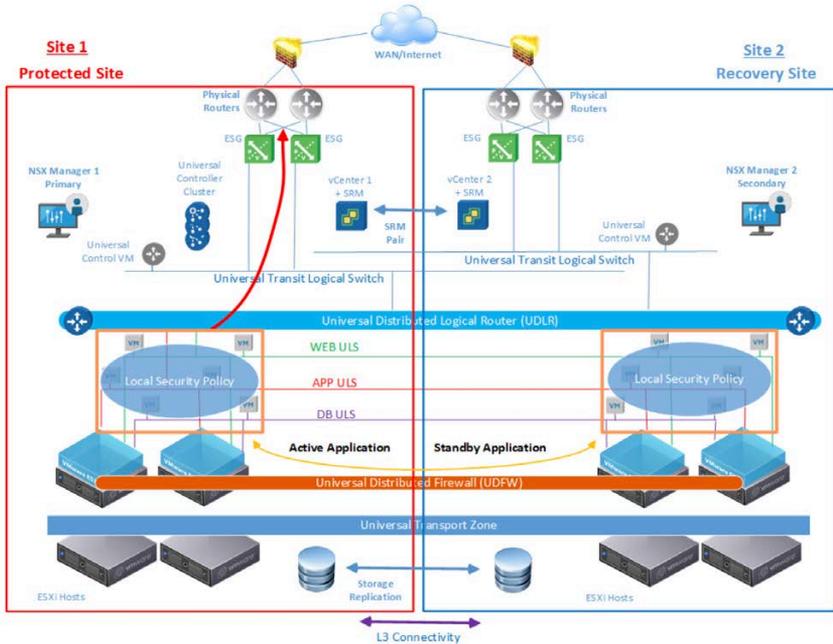
Creating local security groups that contain ULS objects is also supported, as shown in Figure 5.31.



**Figure 5.31** Including ULS in Local Security Group



This method of using local security groups to identify workloads on universal networks will not work with Cross vCenter domain flows; for this, the UDFW must be used. This approach could be adequate for the use cases where there are active/passive flows between vCenter domains/sites. Figure 5.32 shows an example of a DR solution with no Cross-VC traffic flows.



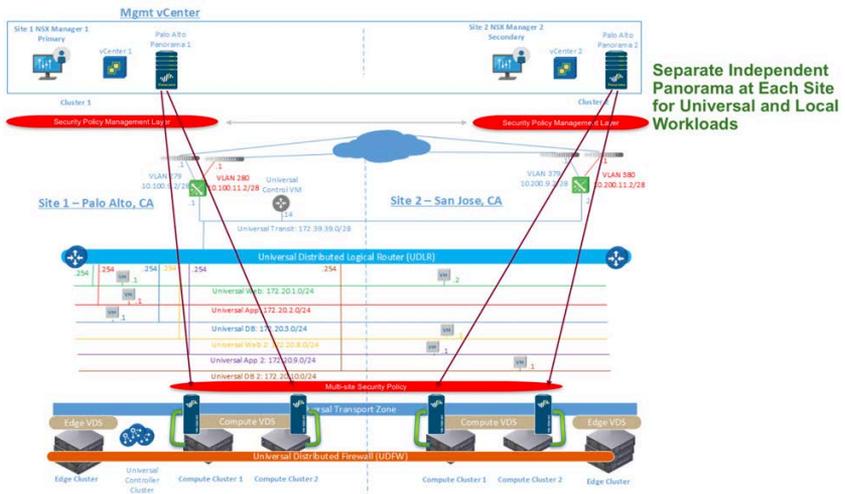
**Figure 5.32** DR Solution Leveraging Cross-VC NSX with Local Security Policies

Such a solution allows the use of vCenter and NSX objects local to each site within local security policies. This method is also useful for firewall/security policies on north/south traffic flows where the destination is identified via an IP address.

For these scenarios, the VMware vRealize® Orchestrator™ (vRO) NSX plugin can be leveraged to create workflows that replicate security policies across sites.

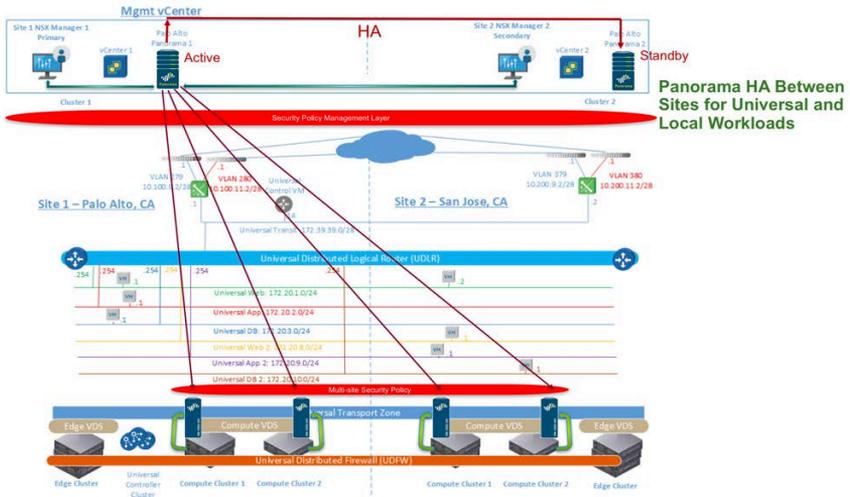
## Third Party Security Services

Third party security services are supported within a Cross-VC NSX deployment. There are multiple deployment options possible; a complete discussion of this is beyond the scope of this book. Figure 5.33 gives an example with a Palo Alto Networks security appliance. In this deployment model, a separate Panorama instance is deployed at each site. From the local Panorama, service VMs (SVMs) are installed. From each NSX Manager, redirection rules must be manually created. In addition, security policies on Panorama at each site have to be manually created as they are not automatically synced. Security policies are then pushed down to the local SVMs from each respective Panorama instance. In this example, NSX UDFW handles up to L4 security and Palo Alto Networks handles up to L7 (e.g., application-level) security.



**Figure 5.33** Cross-VC NSX Deployment Using Palo Alto Networks Security with Separate Panoramas at each Site

Another deployment model with Palo Alto Networks leverages the enhancements in PAN 8.0 with the NSX Manager 2.0 plugin. In this deployment model, a single Panorama can talk to multiple NSX Managers, allowing for the deployment model shown in Figure 5.34. In this example, the active Panorama pushes security policies to all hosts across both sites.



Panorama HA Between Sites for Universal and Local Workloads

**Figure 5.34** Cross-VC NSX Deployment HA Using Palo Alto Networks Security with Separate Panoramas at each Site



Similarly to leveraging Service Composer locally, it is possible to create local security groups that statically contain ULS objects or dynamically identify workloads. These can then be redirected to third party security services/SVMs such as Palo Alto Networks or Checkpoint.



Starting from NSX 6.2.3, universal security groups and universal IP Sets can be used to redirect traffic to third party security services.





# Cross-VC NSX Implementation & Deployment Considerations

With a good understanding of Cross-VC NSX concepts, architecture, components, and implementation now in place, this chapter will walk through a Cross-VC NSX deployment example while discussing different deployment models and options.

# Physical Underlay Network Consideration

VMware NSX can run on any IP network – whether L2, L3, or a combination of both. The requirements for NSX do not change with Cross-VC NSX.



When deploying a Cross-VC NSX solution across sites for both networking and security services, the requirements for interconnectivity between sites are:

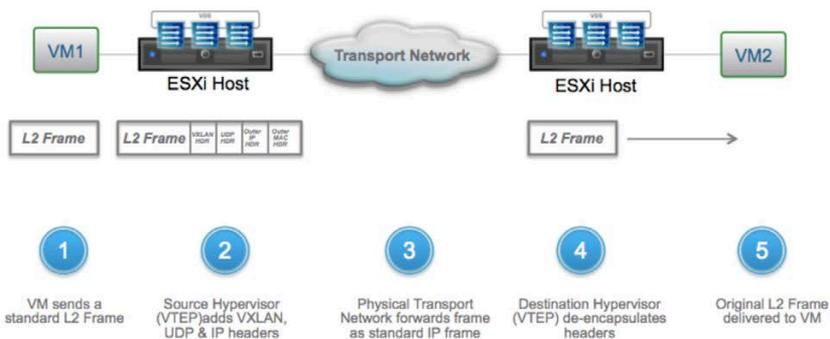
1. IP Connectivity
2. 1600 MTU for VXLAN
3. Sub-150 ms RTT latency

In addition, since logical networking and security span multiple vCenter domains, there must be a common administrative domain across the numerous vCenter deployments participating in the cross-VC NSX deployment. This means the same organization or entity must be managing the different domains/sites.

The requirements for the physical underlay fabric as outlined in the VMware NSX Design Guide (<https://communities.vmware.com/docs/DOC-27683>) remain the same with Cross-VC NSX deployments. As VXLAN encapsulation adds 50 bytes to each frame, the recommendation is to increase the MTU size to at least 1600 bytes.

Since clusters and storage are not stretched, the long distance vMotion 150 ms latency requirement is the limiting factor that dictates the network latency requirement for a multi-site deployment. The control plane latency requirement for Cross-VC NSX aligns with that of long distance vMotion. In a vMSc solution, storage replication calls for a much more aggressive latency requirement (e.g., 10ms).

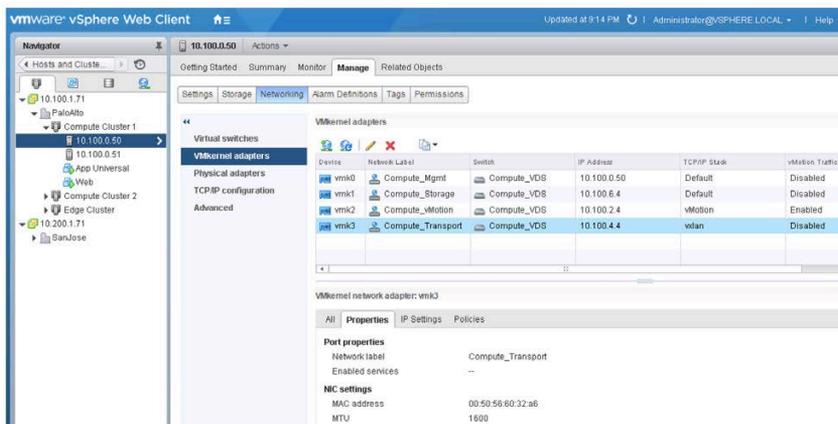
The physical network can be any L2/L3 fabric supporting 1600 bytes MTU. The physical network is the underlay transport for logical networking and forwards packets across VTEP endpoints. Overlay network details (e.g., VM IPs, MAC addresses) are transparent to the physical environment because of the VXLAN encapsulation as shown in Figure 6.1. Encapsulation/decapsulation of the VXLAN header is done by the VTEPs on ESXi hosts, and the physical network must support the 1600 MTU to be able to transport the VXLAN encapsulated frames.



**Figure 6.1** Physical Network Becomes Underlay Transport for Logical Networking

The MTU for the VDS uplinks of the ESXi hosts performing VXLAN encapsulation is automatically increased to that of the MTU specified in NSX during cluster VXLAN configuration.

Figure 6.2 shows the vmkernel port created on the VDS along with its VXLAN settings. The MTU on the vmkernel port is set to the MTU chosen in NSX during the cluster's logical networking preparation - 1600 in this example.



**Figure 6.2** VMkernel interface for Transport Network



VXLAN does not support fragmentation and has the do not fragment (DF) bit set within the frame. This is done to avoid fragmentation-related performance issues. Since the requirements for Cross-VC NSX with VXLAN are IP

connectivity with 1600 byte MTU and latency below 150ms RTT, it is typically not possible to use Cross-VC NSX over a VPN connection like IPSEC over the Internet. The MTU across the entire path from source to destination must be considered and a 1500 MTU is commonly used over the Internet.

Some IPsec implementations can set the TCP maximum segment size (MSS) and perform fragmentation before the packets are encrypted. Even with this effective override of the DF bit setting, the latency must still be less than 150ms RTT as required by Cross-VC NSX control plane and vMotion requirements. While technically possible, this configuration has not been validated and is not supported.

L2/L3 over dedicated fiber or a shared medium like MPLS service from an ISP is typically used for connectivity between sites. L3 connectivity is preferred for scalability and avoidance of common L2 issues such as propagation of broadcast traffic over the DCI link or STP convergence issues.

## Cross-VC NSX Deployment Example and Options

When deploying Cross-VC NSX, there are several design considerations to take into account, including cluster design, placement of management and Edge components, north/south egress, and security requirements.

Multi-site data center solutions with NSX is an advanced topic, sharing the foundational design guidelines discussed in the VMware NSX Design Guide (<https://communities.vmware.com/docs/DOC-27683>). The same fundamental principles for cluster and VDS design should be followed such as using separate management, Edge, and compute clusters and VDS instances for large, high-scale environments. The following sections expand upon these fundamental principles.

The Cross-VC NSX deployment shown in Figure 6.3 stretches across two sites, with separate NSX Manager and vCenter pairs at each site. Site 1 (Palo Alto) is the primary site and site 2 (San Jose) is secondary. Each site has dedicated management, Edge, and compute clusters and VDS instances.

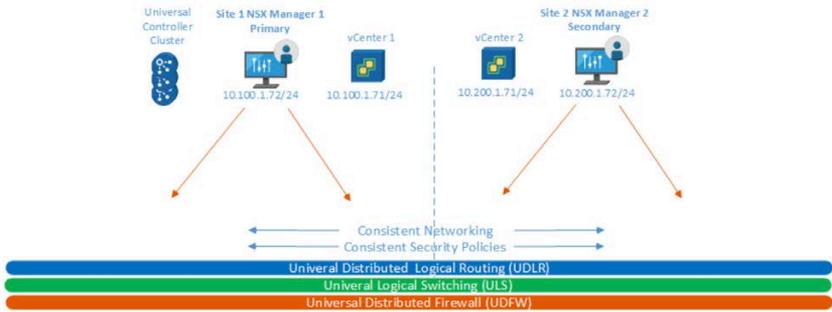


Figure 6.3 Example Cross-VC NSX setup

### Platform Services Controller (PSC)

Deployed, but not shown in these examples, is an external Platform Services Controller (PSC). Introduced in vSphere 6.0, the PSC decouples infrastructure services such as single sign-on (SSO) from vCenter and allows multiple VC to be connected in enhanced linked mode. Note, in vCenter 6.7 (and 6.5U2), enhanced linked mode can also be configured with embedded PSCs. This is not possible in versions prior to vSphere 6.5U2. In prior versions, enhanced linked mode is only supported with external PSC.

The PSC is installed as a virtual appliance from the same VMware vCenter Server® Appliance™ ISO as vCenter. After the PSC is installed, multiple vCenters can be installed and pointed to the PSC for shared infrastructure services. The PSC should be installed on the management network, but can also be routed. vCenter simply needs connectivity to the PSC.

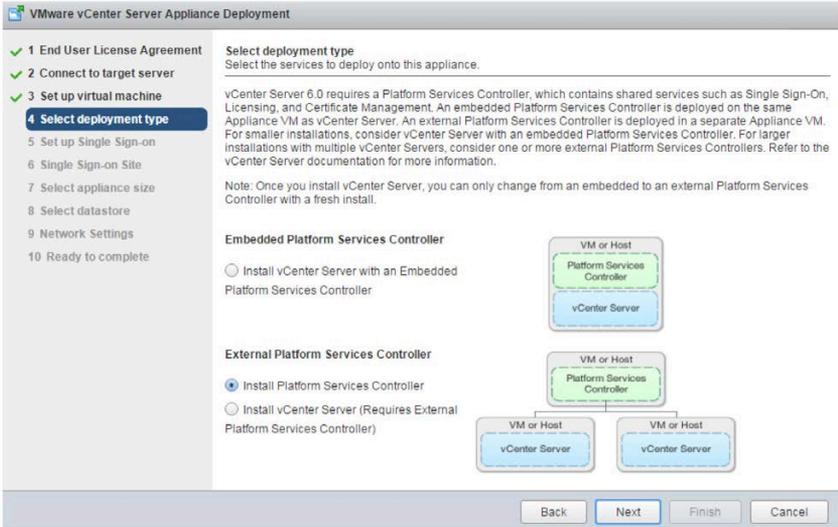


Figure 6.4 Installing PSC From the vCenter Server Appliance ISO

Once the PSC is installed on the management network, the same vCenter Server Appliance ISO can be used to install the respective vCenter instances and link to the PSC as shown in Figures 6.5 and 6.6.

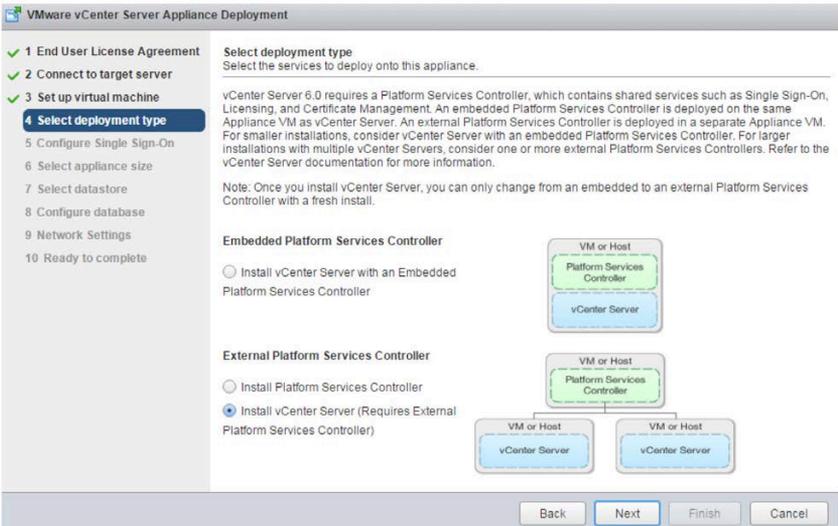
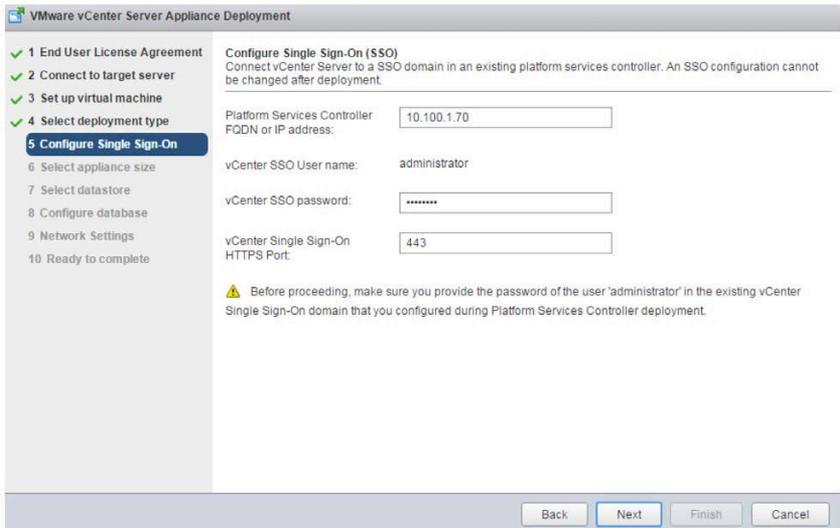


Figure 6.5 Installing vCenter From the vCenter Server Appliance ISO



**Figure 6.6** Connecting vCenter to PSC

Both vCenters from the example connect to and leverage the external PSC, which also enables Enhanced Linked Mode. ELM enables the user to manage multiple vCenters from one GUI, as shown in Figure 6.7. It also allows performing vMotion operations via the GUI from one vCenter domain to the other.



For resiliency reasons, the PSC should be deployed across sites so PSCs can replicate across sites.

With vSphere 6.7 (and 6.5 U2), embedded PSC deployment also supports ELM. In versions prior to 6.5U2, embedded PSC did not support ELM and thus vMotion could not be performed via GUI, but only CLI.

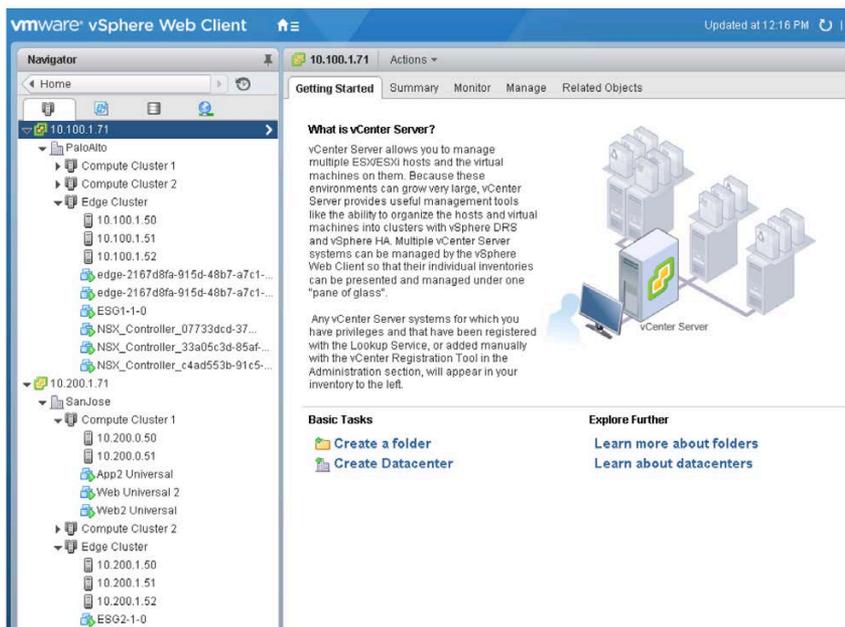


Figure 6.7 Enhanced Link Mode Allows for Central Management of Multiple vCenter Domains

## Management and UCC Component Deployment and Placement

### Primary NSX Manager

Once the NSX Manager at site 1 is deployed via standard NSX Manager installation procedures (e.g., an OVF file), it can be promoted to the primary role as shown in Figure 6.8. For detailed installation instructions, refer to the Cross-vCenter NSX Installation Guide. The primary NSX Manager is deployed within the management vCenter domain as described in the next section.

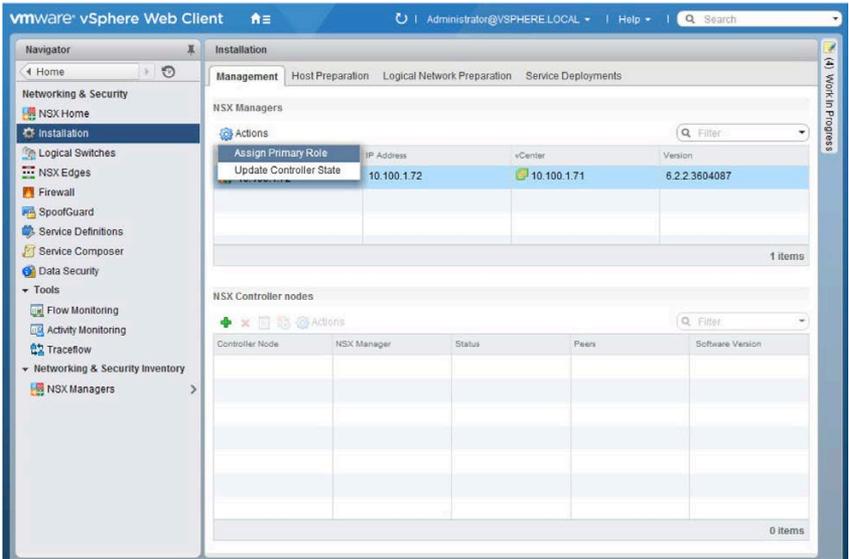


Figure 6.8 Assigning Primary Role to NSX Manager

Once the NSX Manager is made primary, the USS will automatically start running on the NSX Manager. It is stopped by default, and the status will always be **Stopped** on secondary NSX Managers.

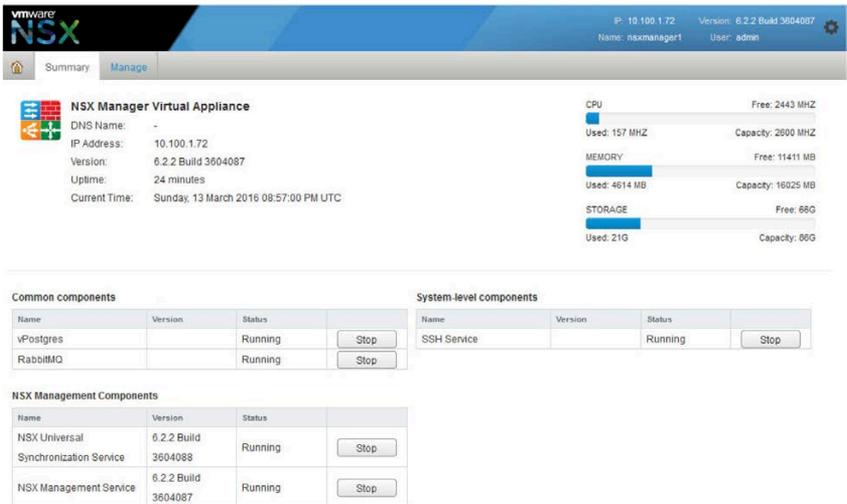


Figure 6.9 NSX USS Running on Primary NSX Manager

Once the primary NSX Manager is configured, the UCC can be deployed from the primary NSX Manager. For detailed installation instructions, reference the Cross-vCenter NSX Installation Guide.

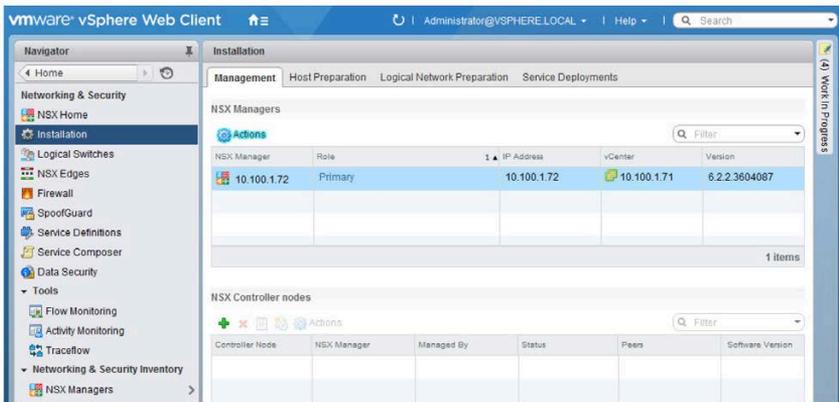


Figure 6.10 NSX Manager Promoted to Primary Role

## Universal Controller Cluster (UCC)

Clicking the green '+' symbol under the NSX Controller nodes section will deploy an NSX Controller. It is recommended to wait until one controller is fully deployed before proceeding to deploying another. Three controllers, the maximum number supported, must be deployed for a proper implementation; three controllers are sufficient to support all vCenter domains and respective logical networks.



Similar to standard resiliency design recommendations, the NSX Controllers should be deployed on separate physical hosts. Leverage anti-affinity rules to ensure that multiple NSX Controllers do not end up on the same physical host. If NSX Controllers are deployed on the same host, resiliency is lost since a physical host failure can bring down more than one controller, possibly the entire cluster if all controllers are on the same host. An IP pool is recommended for NSX Controller configuration to provide simplicity and consistency in IP addressing.

The controllers distribute forwarding information as well as VTEP, MAC, and IP tables to the ESXi hosts while maintaining complete separation from the data plane. If one controller is lost, UCC will keep functioning normally. If two controllers are lost, the remaining controller will go into read-only mode; no new control plane information will be learned but data forwarding will continue.

If the entire controller cluster is lost, the data plane will continue functioning using the last known state. Forwarding path information on the ESXi hosts does not expire; however, no new information can be learned until at least two controllers are recovered.

**Add Controller** ?

NSX Manager: \* 10.200.1.72

Datacenter: \* SanJose

Cluster/Resource Pool: \* Edge Cluster

Datastore: \* EMC\_VNX\_1-1

Host: 10.200.1.50

Folder:

Connected To: \* Edge\_Mgmt Change Remove

IP Pool: \* NSX Controllers Select

Password: \* \*\*\*\*\*

Confirm password: \* \*\*\*\*\*

OK Cancel

**Figure 6.11** Deploying A Controller

As shown in Figures 6.7 and 6.11, the UCC is deployed into the Edge cluster of site 1 because a separate vCenter domain is used to manage the other site. This design deploys a vCenter on an ESXi host it is not managing.

Since the management components do not run over VXLAN, they can be managed by a separate management vCenter. This vCenter may even run on an older version of ESXi since it has no NSX/VXLAN compatibility requirements.

## Mgmt vCenter



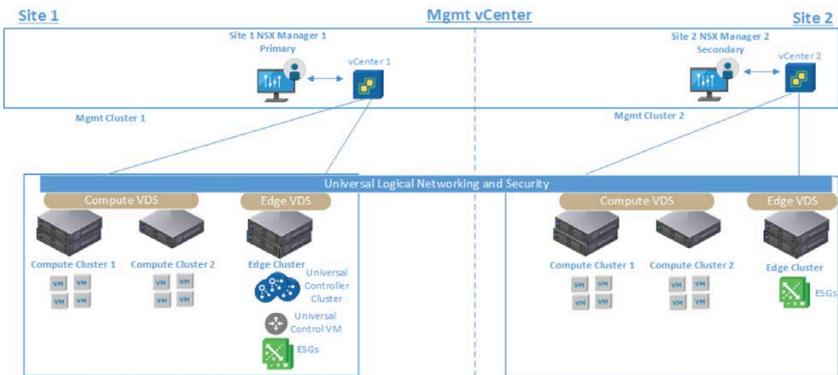
**Figure 6.12** Management vCenter Managing all vCenters and Respective NSX Managers



Having a separate management vCenter is not a requirement, but is useful for the operational reasons mentioned above. It also provides flexibility to install the vCenter and NSX management components on hosts that are on older versions of ESXi.

Managing the management components from a separate vCenter also allows for operational efficiency. Upgrading NSX can be simplified as management is decoupled from the compute/Edge vCenter domain; however, this model requires an additional vCenter license for the management vCenter.

Figure 6.13 shows the separate management vCenter model. In this example, the UCC is deployed on the Edge cluster at the site associated with the primary NSX Manager.



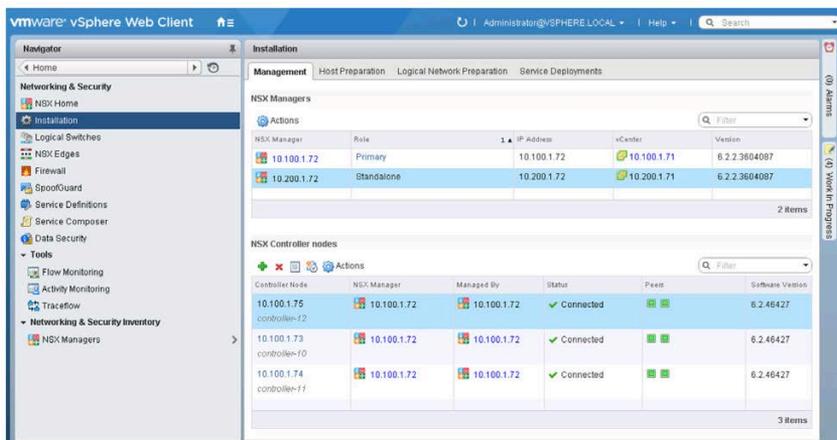
**Figure 6.13** Separate Management vCenter For Multi-site Deployment with Cross-VC NSX

Figure 6.14 shows a model where each site is managed by a dedicated vCenter instance. Both the vCenter and NSX Manager appliances reside in that vCenter's management cluster. Instead of running on an externally controlled management cluster, a separate management cluster within the same vCenter is used, similar to the compute and Edge clusters. This model does not require an additional vSphere license, and no VXLAN configuration is needed on the management cluster.



**Figure 6.14** vCenter and NSX Manager installed on Mgmt Cluster It's Managing

The screenshot in Figure 6.15 shows that all three universal controllers have been deployed at site 1. The UCC must be deployed within the vCenter domain where the primary NSX Manager is linked and cannot be distributed across two vCenter domains. As mentioned in the Cross-VC NSX overview chapter, the UCC should remain within a single site because of the constant communication between controllers in the cluster. The maximum control plane latency supported aligns with the maximum vMotion network latency supported – 150ms RTT.



**Figure 6.15** Universal Controller Cluster Deployed at Site 1

## Secondary NSX Manager

Figure 6.15 shows that another NSX Manager has been deployed and associated with a different vCenter at site 2, designated with the IP address 10.200.1.71. The NSX Manager at site 2 is still in the default standalone mode; it must be registered with the primary NSX Manager as a secondary NSX Manager before universal logical constructs can be deployed across sites.

The primary NSX Manager is selected, and using the **Actions** button, a secondary NSX Manager is registered with the primary as shown in Figure 6.16.

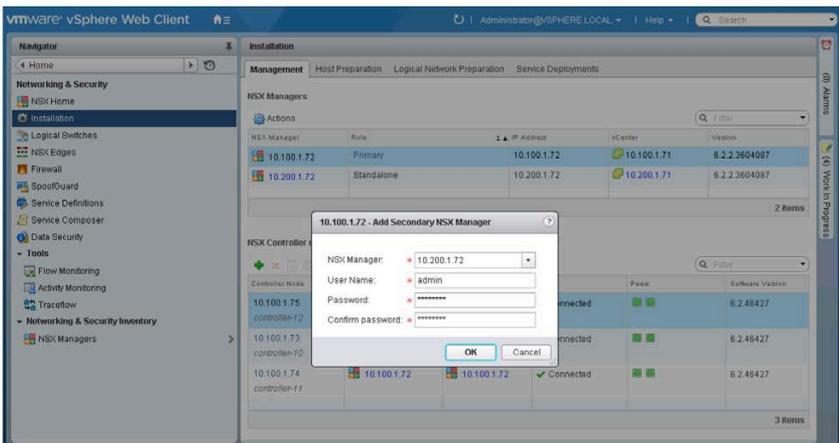


Figure 6.16 Secondary NSX Manager Registered with the Primary NSX Manager

Figure 6.17 displays the secondary NSX Manager once it has successfully registered with the primary NSX Manager.



Three additional rows are shown under NSX Controller nodes. The IP addresses of these controller nodes confirm they are the same controllers initially deployed. If another secondary NSX Manager is registered, three more rows would appear.

These rows signify the connectivity from the respective NSX Manager to the controllers. Each NSX Manager maintains a connection to each of the controllers. The NSX Manager connects to the UCC to push relevant logical networking configuration to the controllers. A periodic keepalive is sent to monitor the state of the controller cluster and measure disk latency alerts; this column displays that status.

The initial status for the recently registered secondary NSX Manager is **Disconnected**, but will quickly change to **Connected**. The **NSX Manager** column confirms the controllers are linked to different NSX Managers. The rows that have green status box icons under the peers column signify that controller nodes are successfully connected to the primary NSX Manager. Only three universal controllers exist at the primary site. These manage all universal and local objects for all vCenter domains within the Cross-VC NSX deployment.

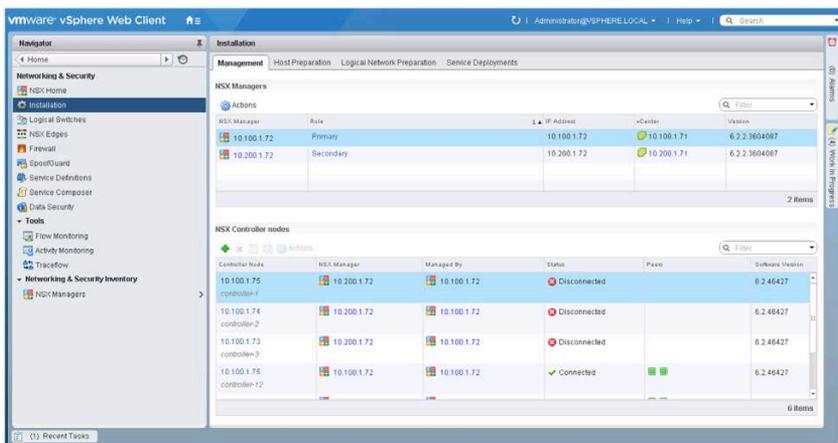
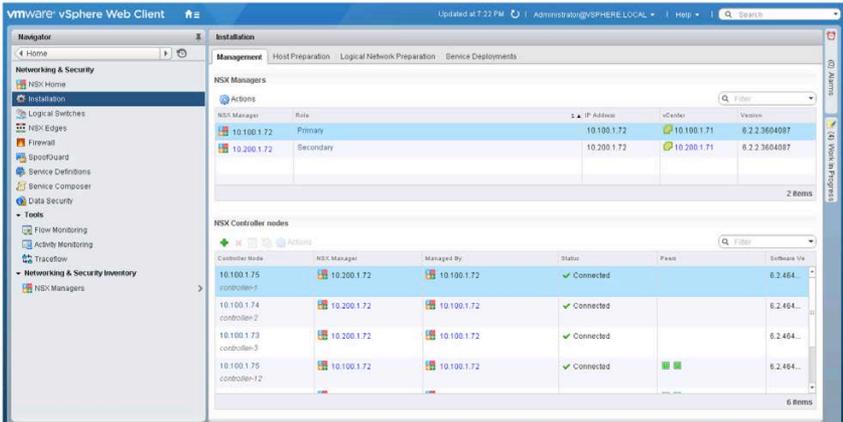


Figure 6.17 Secondary NSX Manager Successfully Registered with Primary NSX Manager

The new rows under the NSX Controller nodes section which have status of **Disconnected** change to **Connected** as shown in Figure 6.18.



**Figure 6.18** Secondary NSX Manager with Successful Connectivity to UCC

The following process occurs once the secondary NSX Manager successfully registers with the primary NSX Manager:

1. USS service on the primary NSX Manager copies the UCC configuration to the secondary NSX Manager.
2. The secondary NSX Manager connects via TLS over TCP port 443 to all controllers of the UCC.
3. The secondary NSX Manager pushes the UCC configuration down to its vCenter domain ESXi hosts via TLS over TCP port 5671.
4. Each ESXi host uses the UCC configuration data to establish control plane connections using TLS over TCP (port 1234) to one or more controllers via its user space control plane agent (netcpa).

## UDLR Universal Control VM Deployment and Placement

The universal DLR control VM is a control plane proxy for the UDLR instances. Similar to the DLR control VM in non-Cross-VC NSX deployments, the universal control VM is typically deployed on the Edge cluster and will peer with the NSX ESG appliances.

Since universal control VMs are local to the vCenter inventory, NSX control VM HA does not occur across vCenter domains. If deployed in HA mode, the active and standby control VM must be deployed within the same vCenter domain.

There is no failover or vMotion of universal control VMs to another vCenter domain. The control VMs are local to the respective vCenter domain.

The number of universal control VMs deployed in a Cross-VC deployment depends on the type of deployment. There are several deployment models for Cross-VC NSX that are discussed further in the Cross-VC NSX Deployment Models section.



A deployment that does not have local egress enabled will have only one universal control VM for the UDLR. If there are multiple NSX Manager domain/sites, the control VM will sit at only the primary site and peer with all ESGs across all sites.

Upon site failure, the control VM would need to be manually redeployed at a new primary site, though this process can be automated.



A multi-site multi-vCenter deployment with local egress enabled will have multiple universal control VMs for a UDLR – one for each respective NSX Manager domain - to enable site-specific north/south egress. Each control VM will connect to a different transit logical network to peer with the ESGs local to its site. Upon site failure, no control VM needs to be manually redeployed at a new primary site as each site already has a control VM deployed.

When looking at the UDLR under NSX Edges within the NSX Manager, a status of **Deployed** means that a UDLR control VM exists on that NSX Manager. A status of **Active** means that although the UDLR instance is enabled on the ESXi hosts, there is no control VM on the respective NSX Manager. The only time both the primary and secondary NSX Managers have a UDLR status of deployed is when local egress is enabled. Local egress is discussed in more detail in the Local Egress section.

NSX Edges				
NSX Manager: 10.100.1.72 (Role: Primary) ▼				
<span style="color: green;">+</span> <span style="float: right;">0 Installing <span style="color: red;">❗</span> 0 Failed</span>				
Id	Name	Type	Version	Status
edge-2167d8fa-915...	Universal DLR	Universal Distributed Router	6.2.2	Deployed

**Figure 6.19** UDLR ON Primary NSX Manager with Universal Control VM Deployed

## Edge Component Deployment and Placement

Since ESGs are local to the vCenter inventory, NSX ESG HA does not occur across vCenter domains. If deployed in HA mode, the active and standby ESG must be deployed within the same vCenter domain.

There is no failover or vMotion of ESGs to another vCenter domain. The ESG and any respective L3-L7 services are local to the respective vCenter domain.



Since ESGs support up to eight-way ECMP, the support for north/south egress resiliency across sites upon Edge/site failure can be achieved by design. Figure 6.20 displays a multi-site deployment where ESGs are deployed across sites 1 and 2. Site 1 ESGs are preferred on egress via higher BGP weight / lower OSPF cost set on the UDLR and ESGs. In the event of a site 1 connectivity failure upstream of the ESGs, the universal control VM will send the universal controller cluster routes pointing to site 2 ESGs for egress traffic. This enables ESG resiliency across sites without ESG failover/recovery.

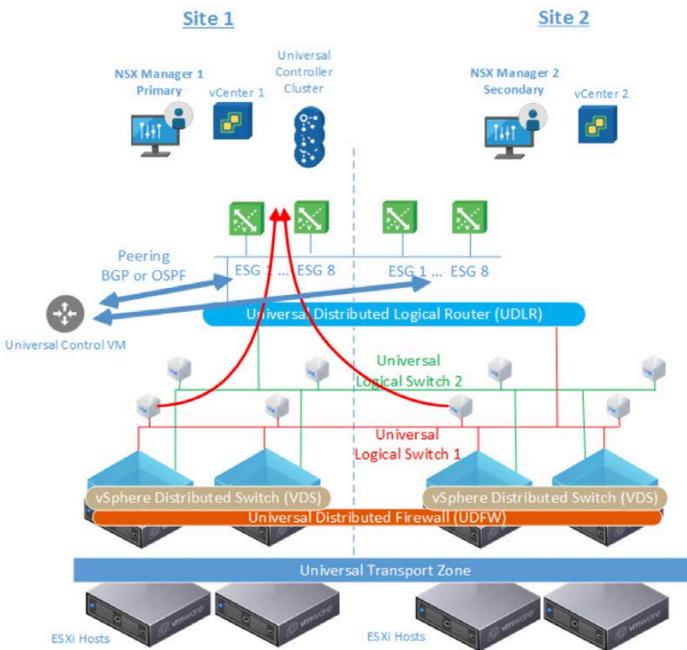
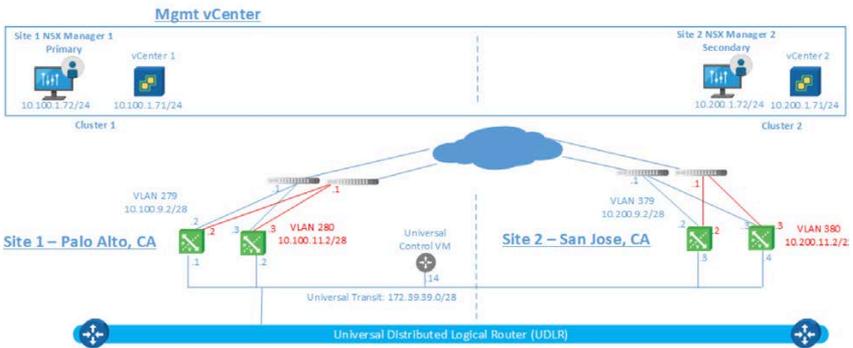


Figure 6.20 ESG Deployed Across Two Sites for North South Traffic Resiliency

## ESG Stateful Services

NSX Edge HA is recommended when the ESG is running stateful services like firewall and load balancer in an active/passive north/south deployment model across two sites. HA should not be stretched across the sites, and the active and standby ESGs must be deployed at the same site.

The ESG stateful services can be deployed on the perimeter ESGs or on a second tier of ESGs that either run inline or in one-arm mode. If ECMP is utilized on the perimeter across multiple ESGs, stateful services on these ESGs cannot be used because of asymmetric routing; flows would be dropped when traversing an ESG that did not initiate the stateful connection.



**Figure 6.21** Multiple ESGs in ECMP Mode Deployed Across Two Sites



Stateful services can be deployed on ESGs running in the following environments:

- Perimeter ESGs if NSX Edge HA mode is utilized instead of ECMP across ESGs
- Second tier of ESGs below the ECMP ESGs
- An ESG in one-arm mode

In each of these cases, the stateful service configurations need to be manually replicated at each site. This can be automated via vRO or custom scripts leveraging the NSX REST API. The network services are local to each site within both the active/passive north/south and the active/active north/south egress models.

In Figure 6.22, a pair of ESGs in HA mode has been deployed with stateful services in one-arm mode at each site. Only the site 1 ESGs have their vNIC connected to the network; the overlapping IP addresses require the site 2 ESG interfaces to be manually disconnected until failover occurs. Upon active ESG failure at site 1, the standby ESG at site 1 will take over. Upon full site failure at site 1, the ESGs at site 2 must have their vNICs connected to the network. ESGs at site 1 do not fail over to ESGs at site 2, and vice versa, as the ESGs are virtual appliances local to the respective vCenter inventory.



**Figure 6.22** Multi-site, Multi-vCenter Deployment With Active/Passive N/S and One-Arm Load Balancer

## Graceful Restart



Similarly to non-Cross VC NSX deployments, when deploying ESGs in ECMP mode in a multi-site environment, graceful restart on the universal control VM and ESGs should be disabled. When enabling dynamic routing protocols on the ESG, graceful restart is enabled by default. If it is left to the default with aggressive timers for routing protocols (e.g., BGP), traffic to the ESG will be black-holed upon failure of an ESG since the forwarding state is preserved on the DLR instances by graceful restart extended hold timers. Even if BGP timers are set to 1:3 seconds for keepalive/hold, the failover will take longer because the graceful restart hold timers take precedence over the BGP timers. The only scenario where graceful restart may be desired on ECMP ESGs is when an ESG needs to act as graceful restart helper for a physical ToR that is graceful restart capable (e.g., hold routes pointing to TOR in case of TOR failure). Graceful restart is more often utilized in a chassis with dual route processor modules than on a ToR switch.

## Local Egress

Local egress is a feature introduced in NSX 6.2 which allows for active/active site-specific north/south egress. This feature allows control of which forwarding information is provided to the ESXi hosts based on a unique identifier for an NSX Manager domain/site. The UCC learns the routes and associated unique identifier from the respective universal DLR control VMs located at each site. ESXi hosts then receive filtered forwarding information from the UCC based on this unique identifier. Hosts for each environment will receive only site-local routes by default which enables localized north/south egress. If local egress is not enabled, locale ID is ignored and all ESXi hosts connected to the universal distributed logical router receive the same forwarding information. Local egress can be enabled only when creating a UDLR.

Local egress and locale ID are discussed in detail under the Active/Active Site Egress (Local Egress Utilized) section under Cross-VC NSX Deployment Models.



# Cross-VC NSX Deployment Models

The examples in this chapter build off of the initial two-site deployment model of UCC along with primary and secondary NSX Managers discussed in the prior section.

There is more than one way to design a Cross-VC NSX topology. The deployment models discussed in detail implement multiple vCenters and multiple sites. These scenarios could also apply to multiple vCenter domains within a single site. Additionally, multi-site topologies leveraging a single vCenter domain are possible.

The deployment models covered in this chapter are:

1. Multi-site with multiple vCenters
  - a. Active/passive site egress - routing metric or local egress utilized
  - b. Active/active site egress - local egress utilized
2. Multi-site with single vCenter
  - c. Active/passive site egress - routing metric or local egress utilized
  - d. Active/active site egress - local egress utilized

Other common deployment scenarios for Cross-VC NSX include single-site/multi-VC deployments. Such deployment models introduce no additional challenges and complexities outside of those covered by the previously discussed deployment models. As this book focuses on multi-site solutions, the single site/multi-VC model will not be covered.

The concepts discussed in the multi-site multi-VC use case overlap with those covered in the multi-site single VC use cases as they share the same subset of functionalities. Specific differences and caveats will be outlined in their respective sections.

It is also possible to deploy a multi-tenant environment where one tenant uses a UDLR instance configured with active/passive north/south egress, while another tenant uses a separate UDLR instance with active/active north/south egress leveraging the local egress feature.

If the tenants have no overlapping IP addresses, they can leverage the same ESGs. If overlapping IP addresses are used, separate ESGs must be deployed. Thus, within a Cross-VC NSX domain, some UDLRs can be deployed with active/passive north/south egress while other UDLRs can be deployed with active/active north/south egress via the local egress feature.

Figure 7.1 shows a scenario leveraging both active/standby and active/active (with local egress) north/south topologies in a single NSX deployment. In this deployment, tenant 1 and 2 use different ULSs but the same UDLR and ESGs. They leverage an active/passive site ingress/egress model where all traffic ingresses/egresses through the primary data center. Tenant 3 uses different ULSs and a different UDLR enabled with local egress. This tenant has a requirement for active/active site egress.

This demonstrates NSX's flexibility, highlighting how different deployment models can easily be deployed within the same NSX environment.

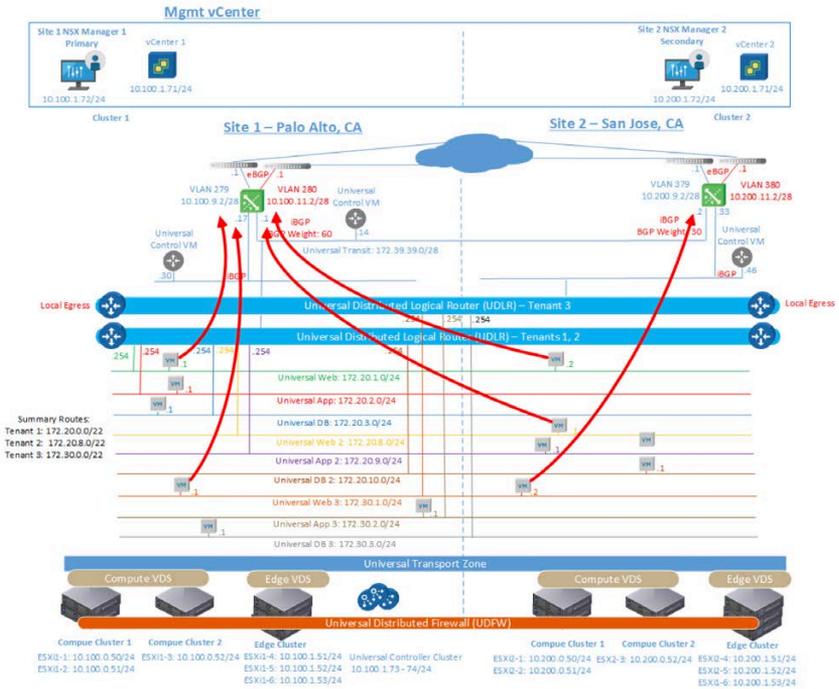


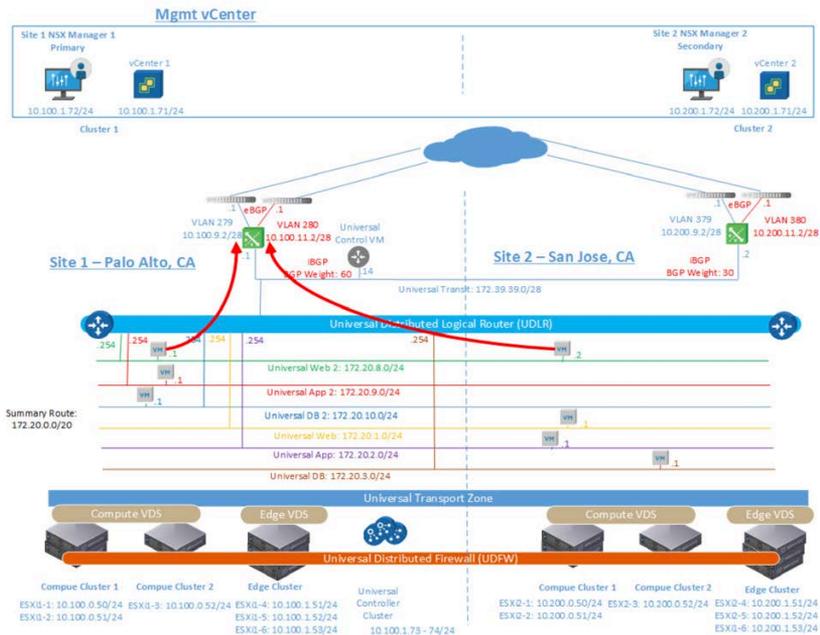
Figure 7.1 Cross-VC NSX Providing Different Deployment Models for Multiple Tenants

# 1. Multi-site with Multiple vCenters

## a. Active/Passive North/South with Routing Metric or Local Egress

### Topology Details

Figure 7.2 displays a diagram of a multi-site multi-vCenter deployment where north/south egress is in active/passive mode.

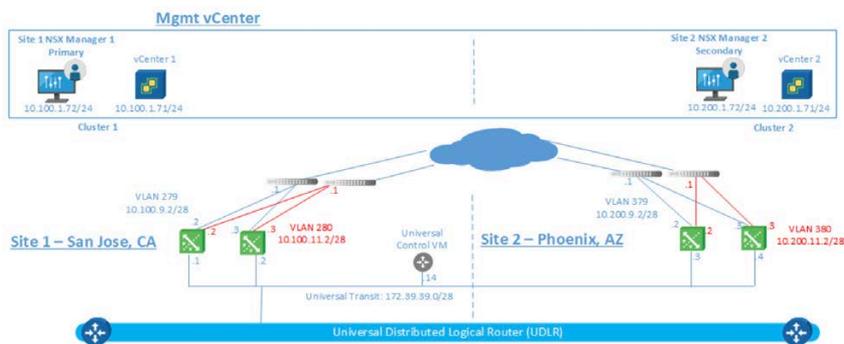


**Figure 7.2** Multi-site with Multiple vCenters & Active/Passive Site Egress

In this example, NSX is deployed across two sites leveraging Cross-VC NSX. A separate vCenter domain is used to host and manage the management components at both sites (e.g., vCenter, NSX Manager). This separate management vCenter should be deployed so that each site's management components are located in their respective region – NSX Manager1/vCenter1 in Palo Alto and NSX Manager2/vCenter2 in San Jose. The management vCenter components are installed at site 1 and manage clusters at both sites 1 and 2. The network between sites is completely routed; the management vCenter only needs connectivity to hosts/clusters at both sites. Each site has its own vCenter deployed locally, so even if the management vCenter at site 1 is lost, the vCenter at site 2 will still be accessible.

In this design, a single universal DLR Control VM is deployed at site 1 and workloads from both sites use site 1 for north/south ingress/egress. The UDLR status at the primary site is **Deployed**, while the UDLR status across all secondary sites is **Active**.

For simplicity of demonstration, only one ESG per site is used with both ESGs doing ECMP northbound. The DLR does not need to run ECMP as the routes advertised by ESG1 at site 1 have precedence over those advertised by ESG2 at site 2. In a production environment, multiple ESGs should be deployed at each site for additional resiliency as discussed in the Edge Component Deployment and Placement section. ECMP should be enabled on the DLR to take advantage of multipath forwarding provided by the multiple ESGs. Such connectivity between the UDLR and ESG would look similar to that shown in Figure 7.3.



**Figure 7.3** Multiple ESGs in ECMP Mode Deployed Across Two Sites

Routes advertised by the set of ECMP ESGs at site 2 would still have different routing cost metrics than those at site 1 in order for site 1 ECMP ESGs to be the primary point of ingress/egress.



In an active/passive north/south deployment model across two sites, it is possible to deploy an ESG in NSX Edge HA mode within one site to implement stateful services like firewall and load balancer. If deploying ESGs in HA mode, both the active and standby ESGs must be deployed at the same site as shown in Figure 7.4.



In this diagram, a pair of ESGs in NSX Edge HA mode have been deployed with stateful services at each site, but only the ESGs at site 1 have their uplinks and downlinks connected to the network. Disconnecting the site 2 ESG interfaces from the network is needed as this ESG has overlapping IP addresses with the ESG at site 1. Upon active ESG failure at site 1, the standby ESG at the same site will take over. ESGs at site 1 do not fail over to ESGs at site 2, and vice versa, as the ESGs are virtual appliances local to the respective vCenter inventory. Upon full site 1 failure, the ESGs at site 2

must have their vNICs connected to the network to take over the north/south connectivity responsibilities; this can be automated via vRealize Orchestrator or custom scripts leveraging the NSX REST API.

The stateful services for the ESG need to be manually replicated across sites. This can also be automated via vRO or custom scripts leveraging NSX REST API. The network services are local to each site within this active/passive north/south egress model.

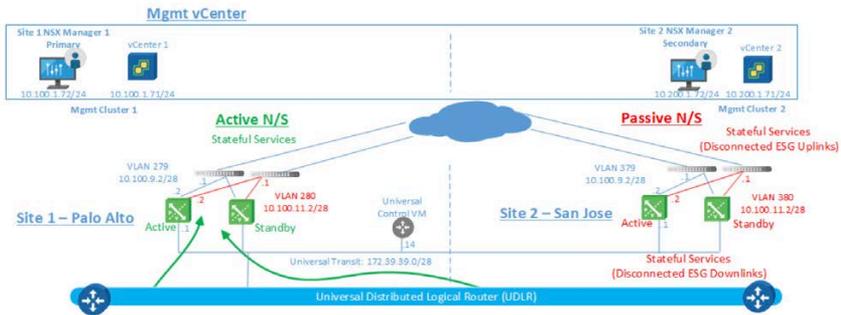


Figure 7.4 HA ESGs with Stateful Services Deployed At Two Sites

With Edge HA, the active and standby ESGs exchange keepalives every second to monitor connectivity and health status. The keepalives are L2 probes sent over an internal port-group which can be a VXLAN-backed port-group. The first vNIC deployed on the NSX Edge is utilized for this heartbeat mechanism unless a different interface is explicitly selected.



If the active Edge fails, the standby takes over the active duties at the expiration of a **Declare Dead Time** timer on the ESG. The default value for this timer is 15 seconds. The recovery time can be improved by reducing the detection time through the UI or API. The minimum supported value is 6 seconds.

Since syncing of L4-L7 services across the ESGs for an active/standby pair also occurs over the same internal port-group, be careful not to set the timer too low where failover occurs prior to all state being successfully synced. A timer setting of at least 9 seconds is recommended.

## Routing Details

BGP is a preferred protocol in modern data center spine-leaf architectures due to its scalability, granular control through many BGP attributes, and ease of deployment for multi-tenant environments. NSX supports both eBGP and iBGP.

For multi-site deployments, use of iBGP and a private ASN between the ESG and UDLR is recommended for simplicity of implementation. As BGP is used for connectivity to the Internet, consistency in routing protocol from the virtual environment to the Edge provides additional simplicity.

If desired, OSPF can also be used on the UDLR and ESG; however, for consistency and ease of deployment, the same routing protocol used between the UDLR and ESGs should be used between the ESGs and the physical network.



As of release 6.3, NSX supports 4-byte autonomous system numbering (ASN). Use of private BGP AS numbers for NSX tenants is recommended. IETF RFC 6996 describes a contiguous block of 1023 autonomous system numbers (e.g., 64512-65534) that have been reserved for private use. This should provide sufficient ASNs for most implementations. As described in IETF RFC 7300, ASN 65535 is reserved; it is not a private AS, and should not be used as such. Any attempt to configure it will be rejected by UI/API.

iBGP is used between the universal control VM and ESG, while eBGP is used between the ESG and ToR switches/routers. As shown in Figure 7.5, the same private ASNs are used for the universal control VM and ESGs. Since the NSX ESG does not support BGP **next-hop-self**, the ESG interfaces connecting to the TOR must be redistributed into BGP.

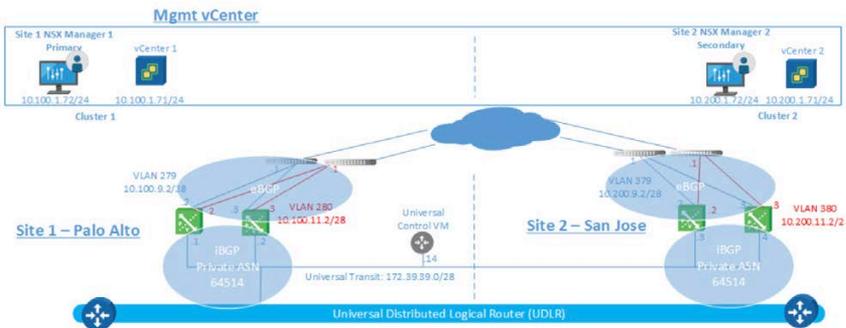
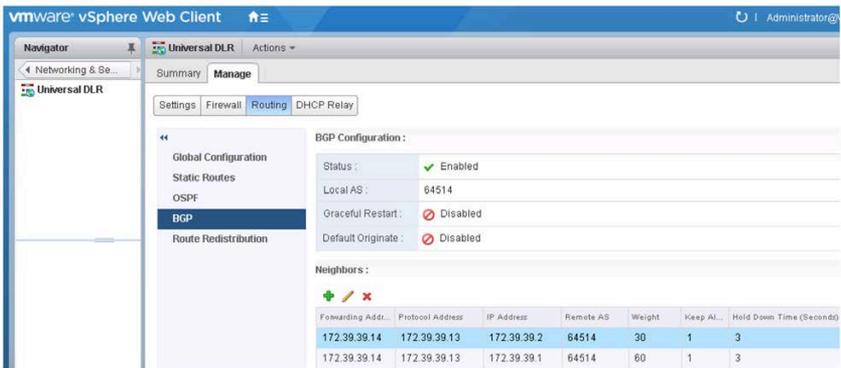


Figure 7.5 NSX Multi-Site Cross-VC Setup With BGP

A private ASN is used for iBGP between the universal control VM and ESGs. eBGP is typically used when exchanging routes with physical networks while iBGP is typically used between UDLR and ESG within the NSX domain. Setting the routing metric on the UDLR allows control of egress traffic. Setting BGP weight on the UDLR influences which route workload traffic will take. Since weight is a non-transitive BGP property – meaning it is local to the router – it must be set on the UDLR which has visibility to all the routes across ESGs at both sites.

Figure 7.6 shows the BGP weight attribute set higher on the UDLR for the neighbor relationship with the site 1 ESG, causing it to be the preferred route.



**Figure 7.6** BGP Weight Attribute Used to Prefer Routes to ESG 1 at Site 1

Figure 7.7 shows the result of a traceroute from a VM on the web logical switch at site 1 to a physical workload on a VLAN sitting on the physical network. The correct ESG at site 1, 172.39.39.1, is used due to the BGP weight setting set for the respective neighbors.

```

Administrator: Command Prompt
C:\Users\Administrator>
C:\Users\Administrator>tracert 10.114.223.70

Tracing route to DNS [10.114.223.70]
over a maximum of 30 hops:

  0  <1 ms    <1 ms    <1 ms    172.20.1.254
  1  <1 ms    <1 ms    <1 ms    172.39.39.1
  2  1 ms     <1 ms    <1 ms    10.100.11.1
  3  1 ms     1 ms     <1 ms    DNS [10.114.223.70]

Trace complete.
C:\Users\Administrator>

```

**Figure 7.7** Traceroute From VM on Web Universal Logical Switch at Site 1 To Physical Workload Confirms ESG At Site 1 is Being Used for Egress

Controlling ingress traffic is important to avoid asymmetric traffic flows. This is especially critical if there are any stateful services in the core that packets traverse at each site. Asymmetric routing is a situation which occurs when egress traffic from a specific workload takes one route and ingress traffic for the same workload returns through another. Since different paths are used for ingress compared to egress, issues can arise if there are any stateful services in between, such as firewalls or load balancers. Even without stateful services, inefficient traffic flows can occur. Due to these concerns, care must also be taken to control ingress traffic.



The NSX ESG does not support BGP attributes to dynamically control ingress such as **AS Path Prepend**. Three options exist to control ingress:

1. Configure the physical network so ingress traffic is routed to the correct ESGs at site 1 (e.g., BGP attributes **AS Path Prepend** or **MED**). AS path is a transitive BGP property. Unlike non-transitive BGP properties, changes to the AS path are propagated down the network, thus traffic engineering is possible by leveraging this behavior.

AS Path adds an additional AS number to the AS Path. The AS path is one attribute the BGP routing algorithm uses to determine the path to take. AS path prepending is most often leveraged for BGP updates sent towards a transit ISP to influence incoming traffic. Best practice is to use the organization’s own AS for the prepending to avoid any transit issues further downstream.

This method has the advantage of requiring no changes to route filtering or manual redistribution of routes at secondary sites when the primary site fails. A disadvantage is that not all physical devices support these BGP attributes. Additionally, a change on the physical network is required, which may require coordination with other teams within an organization.

2. Filter routes on site 2 ESG(s) so respective NSX logical networks are only advertised out site 1 ESG(s). This method requires no manual configuration on the physical network and all ingress/ egress control can be done within the NSX domain. This may appeal to some organizational structures where changing anything on the physical network would require additional approval and coordination with other teams. A disadvantage is that failover will require intervention to change the filtering configuration.
3. On site 2 ESG(s), do not redistribute the logical networks into eBGP; the respective NSX logical networks will be learned only by eBGP at site 1. This method prevents the routing entries from being learned on the TOR since there is no redistribution. Redistribution only occurs upon site 1 failover – through manual or automated steps – so routing entries for the logical networks are only learned from site 2 ESGs when needed. This method will have a slightly higher convergence time as eBGP must learn the logical networks when redistribution occurs.
4. An advantage of this method is that there is no manual configuration required on the physical network and all ingress/ egress control can be done within the NSX domain. This may appeal to some organizational structures where changing anything on the physical network would require additional approval/ collaboration with other teams. A disadvantage is that failover will require manual intervention to change the filtering configuration.

Figure 7.8 shows a traceroute done from a physical workload on a VLAN sitting on the physical network to the VM on the web logical switch at site 1. It shows that the correct ESG at site 1, 10.100.9.2, is used.

```

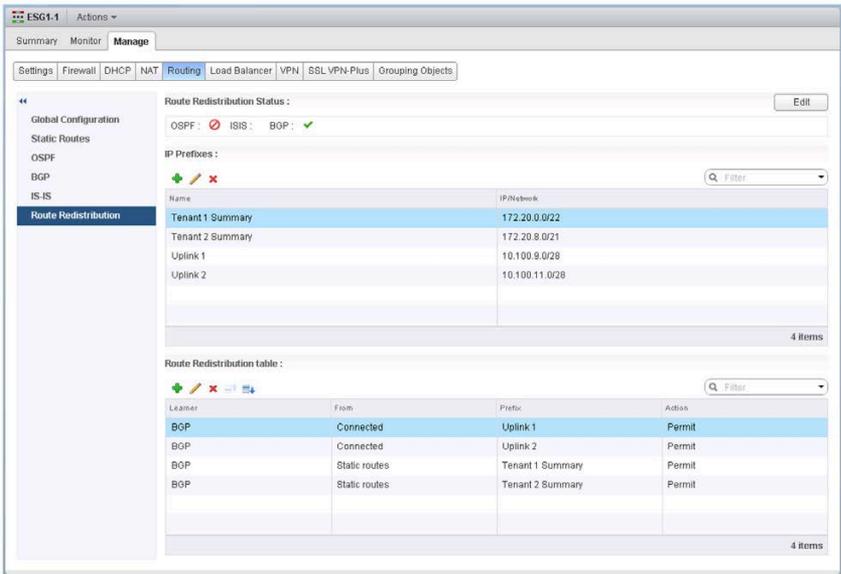
Tracing route to WIN-PN7SD16F52K [172.20.1.1]
over a maximum of 30 hops:
  1  <1 ms  <1 ms  <1 ms  10.114.223.66
  2  <1 ms  <1 ms  <1 ms  10.100.9.2
  3  <1 ms  <1 ms  <1 ms  172.39.39.14
  4  1 ms   <1 ms  <1 ms  WIN-PN7SD16F52K [172.20.1.1]

Trace complete.

```

**Figure 7.8** Traceroute From Physical Workload to VM on Web Universal Logical Switch at Site 1 shows ESG 2 Is Being Used

Figures 7.9 and 7.10 show how NSX logical networks are advertised from only site 1 since the NSX logical networks are not redistributed into iBGP to control ingress traffic. Upon site 1 Edge or complete site failover, the redistribution of respective logical networks at site 2 would need to be permitted. Alternatively, BGP filters (e.g., prefix lists) as well as configuration on the physical network can be used to control ingress traffic.



**Figure 7.9** NSX Logical Network Summary Routes Redistributed at Site 1 ESG

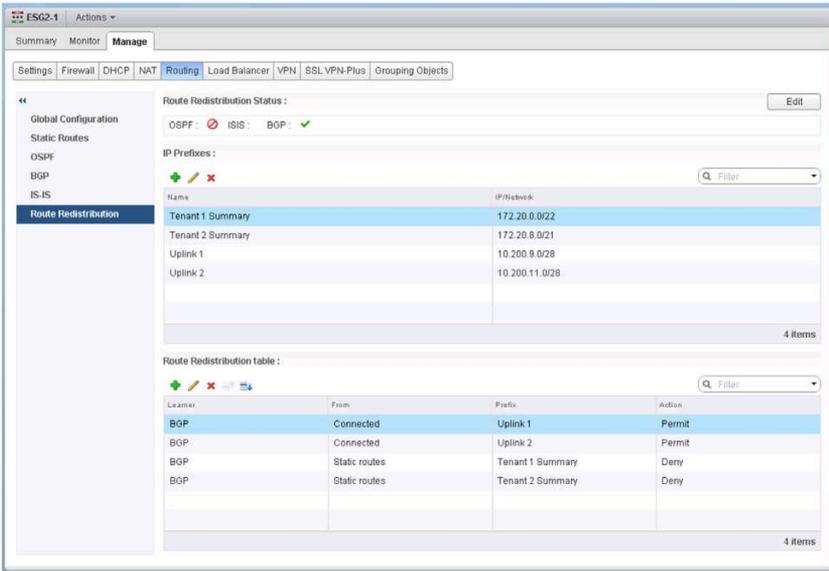


Figure 7.10 NSX Logical Network Summary Routes Not Redistributed at Site 2 ESG



Figures 7.9 and 7.10 show ESGs advertising summarized routes of NSX logical networks. This is recommended as the virtualized environment acts as a stub network and the physical environment needs only reachability details for the NSX logical network topology. Similarly, the virtual environment does not need to know all the routes and specifics about the physical environment; it only needs to learn the default route for traffic to exit the NSX domain. The summary routes and directly connected uplinks on the Edge are redistributed into BGP, and a default route is advertised into BGP at the TOR – either through a **default-originate** command or explicit static route.

As shown in Figure 7.5, every ESG from each site has two external interfaces. Each interface connects to a different site-local ToR switch. Each interface on an ESG will map to a different VLAN-backed port group and peer with the corresponding SVI interface configured on the respective ToR switches.

The next example shows how BGP filtering via prefix lists is used to advertise NSX logical networks out of a single site to control ingress traffic. At site 1, as shown in Figures 7.11 and 7.12, the summary routes for the NSX logical networks are permitted in the prefix list. The prefix list is applied to both interfaces on the ESG.

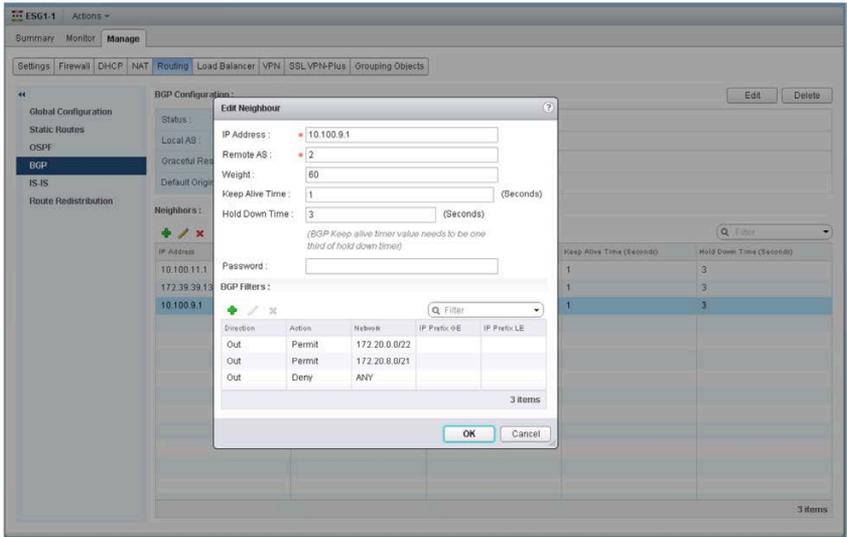


Figure 7.11 Summary routes for NSX permitted in the prefix list at Site 1 ESG Interface 1

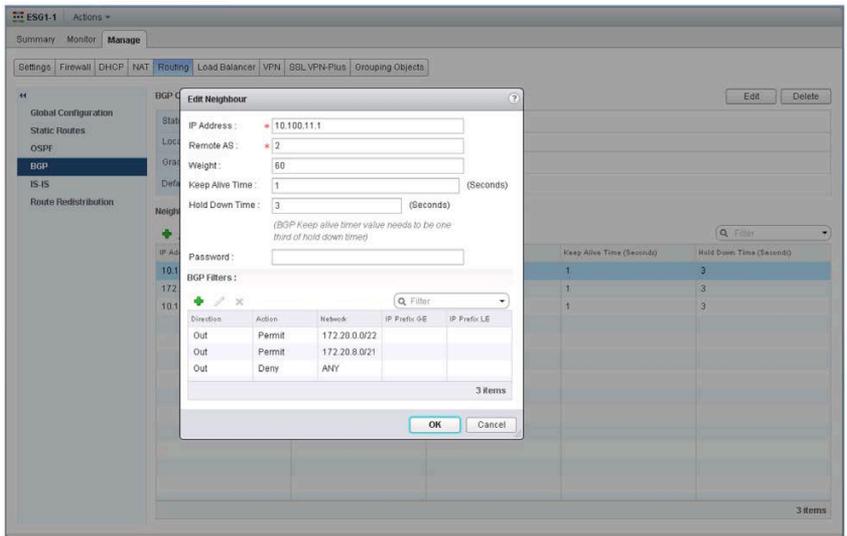


Figure 7.12 Summary routes for NSX permitted in the prefix list at Site 1 ESG Interface 2

At site 2 ESGs, the summary routes for the NSX logical networks are denied in the prefix list (Figures 7.13 and 7.14). The prefix list is applied to both interfaces on the ESG. If a site failure occurs at site 1, the prefix list at site 2 would be modified to permit.

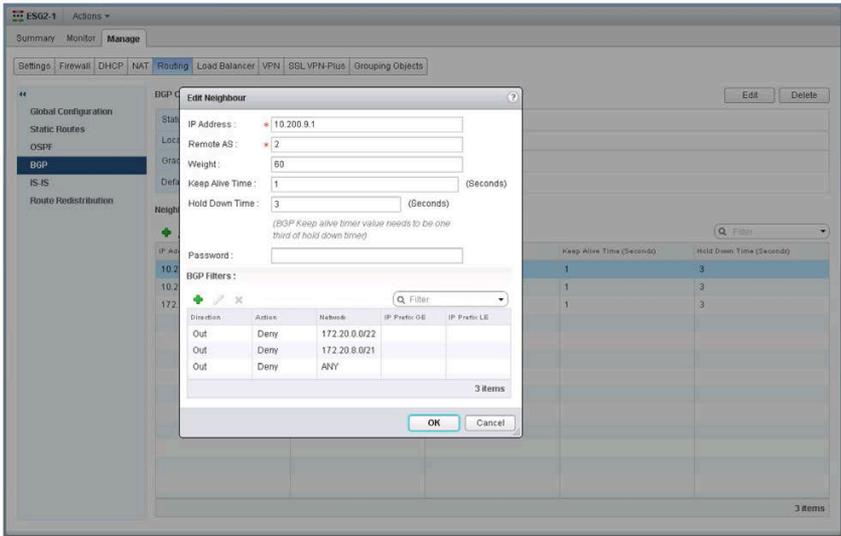


Figure 7.13 Summary routes for NSX denied in the prefix list at Site 2 ESG Interface 1

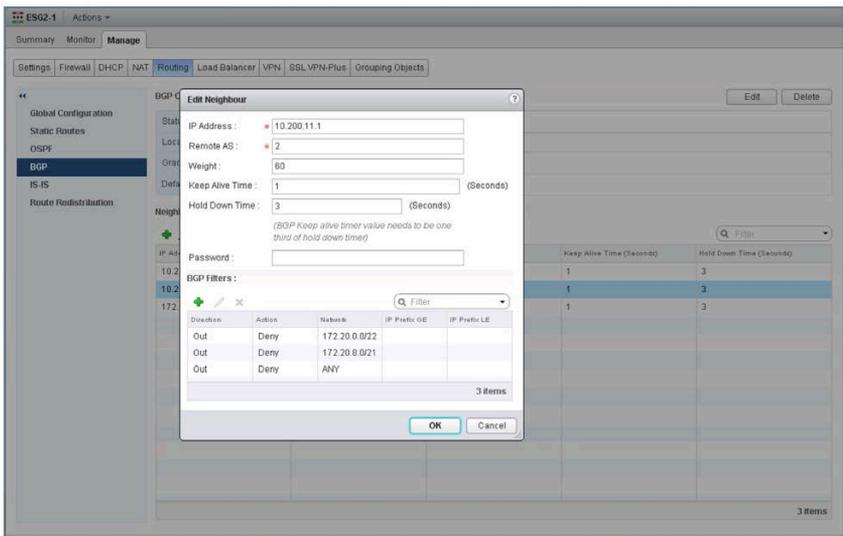


Figure 7.14 Summary routes for NSX denied in the prefix list at Site 2 ESG Interface 2



The BGP keepalive and hold-down timers have been set to 1 and 3 seconds, down from their defaults of 60 and 180 seconds in Figures 7.11–7.14. The keepalive is set to 1/3<sup>rd</sup> of the dead timer as a standard best practice. NSX supports aggressive tuning of the routing hello/hold timers to minimize traffic loss and expedite traffic failover upon Edge failure. If setting aggressive timers, graceful restart on the universal control VM and ESGs should typically be disabled. See the Graceful Restart section for additional information.



In an ECMP ESG topology, the routing protocol hold-time is the main factor influencing the failover timing to another ESG. When the ESG fails, the physical router and the DLR continue to send traffic to the failed ESG until the expiration of the hold-time timer. Once the hold-time expires, the adjacency is brought down and the forwarding table is updated to point to a new ECMP ESG. A manual, graceful ESG shutdown is an exception; in this case, the TCP session for the BGP protocol is terminated and failover will be instant.



Stateful services are not used across ESGs in an ECMP design since state is not synchronized between ESGs in ECMP and flows taking asymmetric paths would be dropped by devices running stateful services. Local egress, which also affects routing, is discussed in the next section.

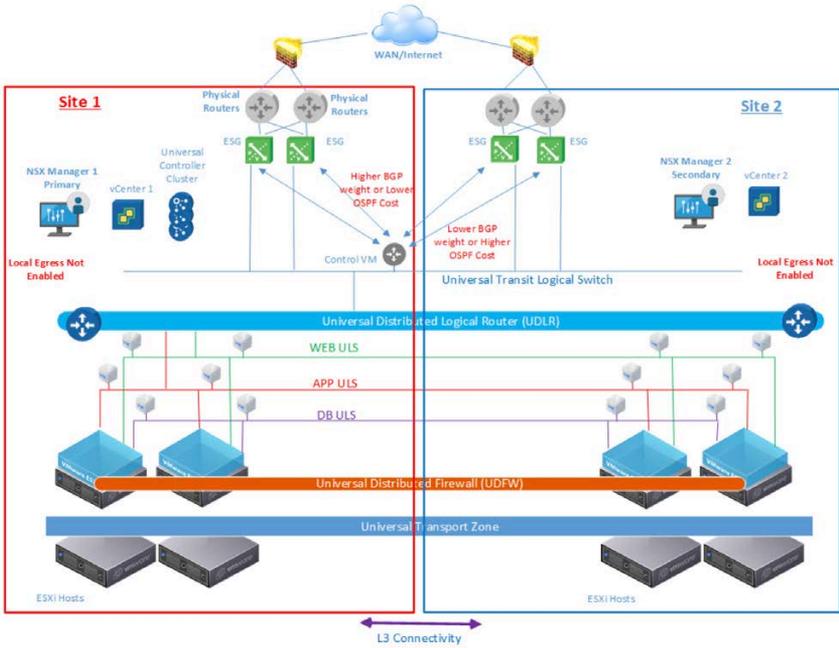


If a private ASN is used with NSX ESG, default behavior is set to strip all private AS paths from the BGP advertisements upon sending routes to eBGP peers. In this case, the prefixes should be filtered outbound on the ESG or inbound on the ToR switch/router so the NSX domain does not become the transit AS.

In NSX 6.4, the default behavior with respect to private ASNs remains the same, but users have the option to prevent the ESG from stripping the private ASN.

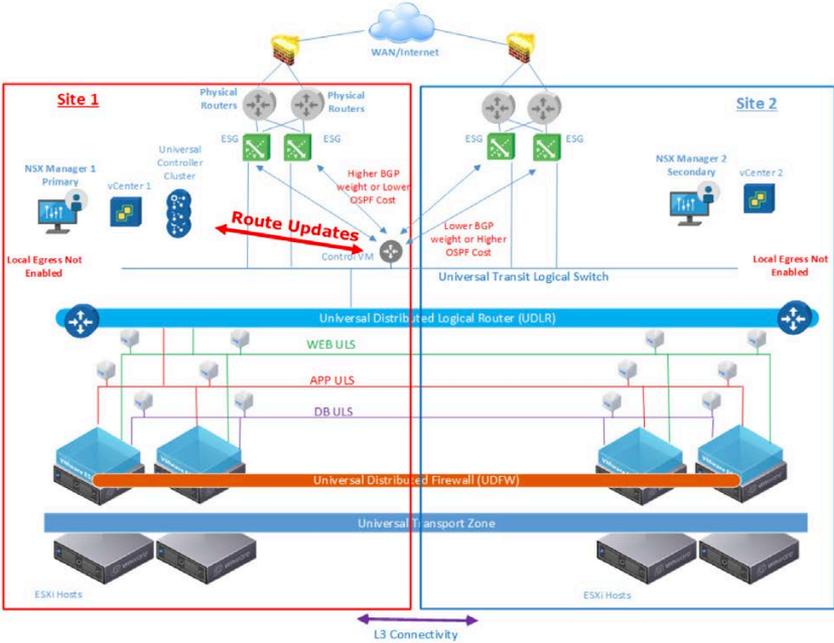
## Traffic Flow

Figure 7.15 illustrates a Cross-VC NSX setup across two sites with active/passive north/south egress. The scenario also consists of active/active workloads across both sites. ECMP ESGs are deployed across both sites and a single DLR control VM is used. Routing metrics are utilized to ensure all traffic for a given workload, regardless of site locality, will egress via site 1 ESGs. This is controlled via routing cost metrics – either higher BGP weight / lower OSPF cost for site 1 routes on the UDLR and ESGs. No local egress feature is used in this example. One universal control VM exists at the primary site which learns routes from both site 1 ESGs and site 2 ESGs.



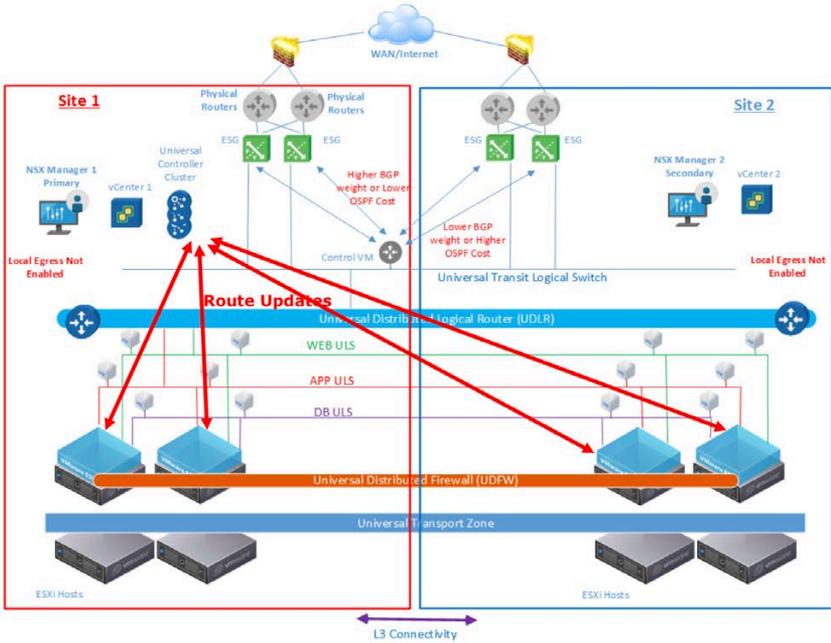
**Figure 7.15** Cross-VC NSX setup – Multi-vCenter, Multi-Site, Active/Passive Egress

Figure 7.16 shows the universal DLR control VM communicating the best forwarding paths for north/south egress to the UCC. In this example, the forwarding paths to ESGs at site 1 have the better metric/cost. This communication is done through the netcpa-UCC control plane connection of the ESXi server hosting the DLR control VM.



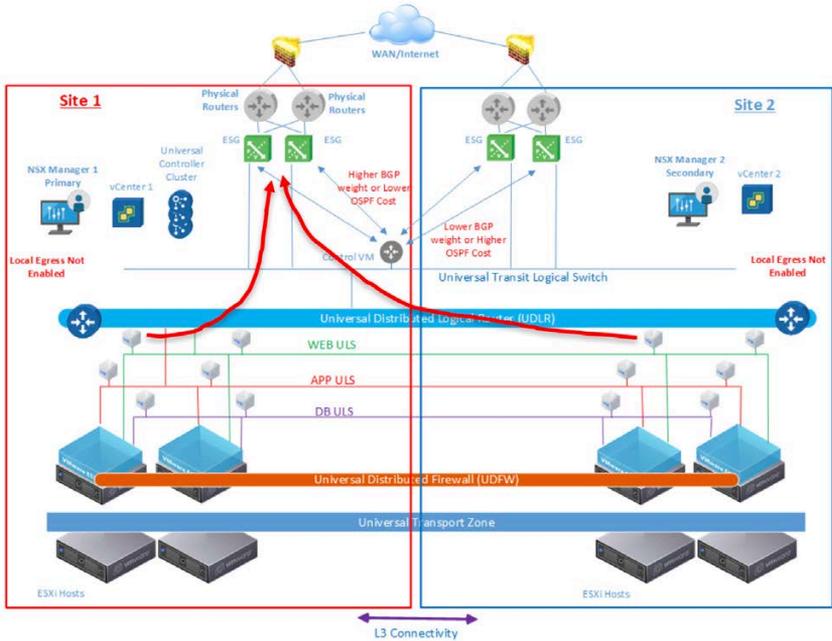
**Figure 7.16** Universal Control VM Informs UCC of Best Forwarding Paths

The UCC distributes the forwarding information learned from the universal control VM to all the ESXi hosts at all sites, as shown in Figure 7.17. In this model, the forwarding table is the same across all ESXi hosts in the multi-site environment.



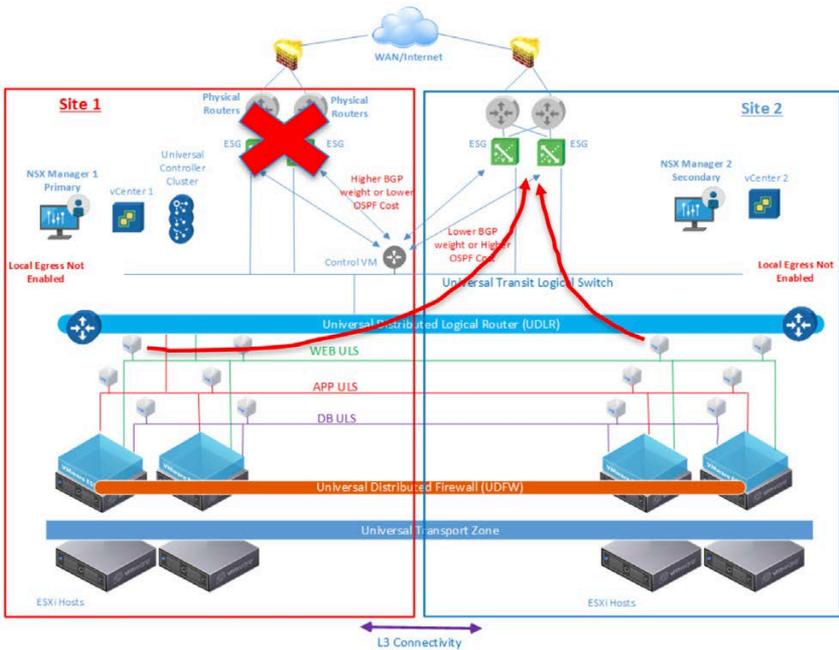
**Figure 7.17** UCC Distributing Forwarding Information to ESXi hosts across All Sites

Since the best forwarding path communicated to the UCC and ESXi hosts are the routes out of site 1 ESGs, all north/south egress utilizes site 1 egress points as shown in Figure 7.18.



**Figure 7.18** ESXi Hosts Across All Sites Use Site 1 Egress

In the event of a failure of site 1 ESGs or of upstream connectivity, the universal control VM communicates to the UCC that the new forwarding path out of the NSX environment is through site 2 ESGs. The UCC then distributes this information to the ESXi hosts across all sites, and north/south egress for both sites switches to utilize the site 2 ESGs. This scenario is pictured in Figure 7.19.



**Figure 7.19** Site 2 ESGs Used For Site 2 Egress Upon Site 1 ESG/Upstream Connectivity Failure

In this type of deployment, NSX components can be protected locally via standard methods of HA. In the rare event of a complete primary site failure, a secondary NSX Manager must be promoted to a primary role and the UCC and universal control VM redeployed at the new primary site.

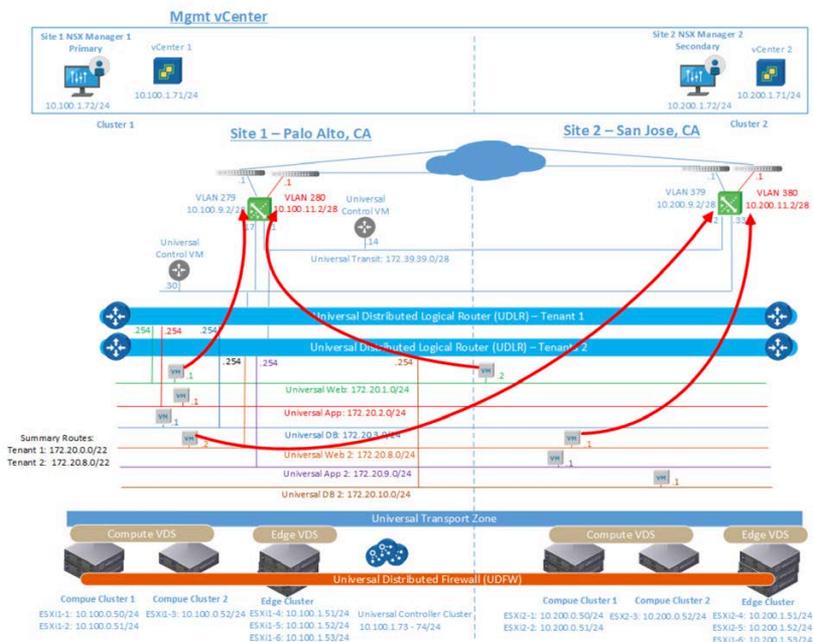


When a failure occurs, data forwarding will continue using the last known state. The control plane functionalities provided by the NSX Controllers, such as route publications or ARP suppression, will be lost. Furthermore, no new universal or local objects can be created until at least two controllers are brought up in the new UCC at new primary site.



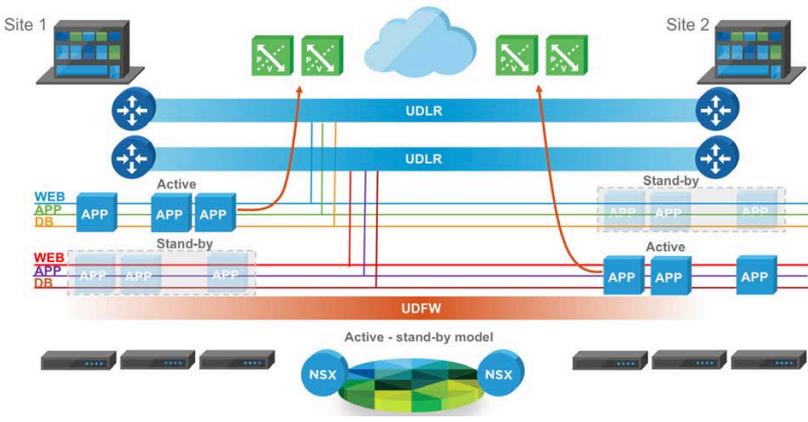
If UCC is down, handling of BUM traffic falls back to data plane flooding. If a VM on a logical switch does not know how to reach a destination VM, it will broadcast an ARP request on the logical network. **CDO mode** can ensure control plane functionality continues in the case of a complete controller cluster outage, however no new routes can be learned.

Using this deployment model, it is possible to deploy multiple UDLRs for different tenants/workloads. One UDLR can be set to select egress routes out of site 1 ESGs, while the other UDLR can prefer egress through site 2 ESGs. This deployment model is still active/passive ingress/egress for respective tenants/workloads, but some tenants/workloads will utilize site 1 ESGs while workloads from the other tenant will utilize site 2 ESGs. An example of such a deployment is shown in Figure 7.20.



**Figure 7.20** Cross-VC NSX Deployment with Multiple UDLRs – ESGs at both Sites Being Utilized

A common use case for such a deployment is bi-directional DR, where each site provides mutual protection for each other (e.g., active/standby and standby/active). Some applications are active at site 1 and standby at site 2, while other applications are active at site 2 and standby at site 1. Figure 7.21 presents an example of this model.



**Figure 7.21** Cross-VC NSX Deployment for Bi-directional DR



Active/passive north/south traffic flow can also be accomplished with the local egress feature. By default, local egress uses a unique identifier called the locale ID from the local NSX Manager to filter routes. This ensures workloads at site 1 always use site 1 egress, and workloads at site 2 always use site 2 egress. By changing the locale ID for workloads at site 2 to the locale ID at site 1, all workloads will egress site 1. Upon Edge or site failure, the locale ID can be switched to the locale ID of site 2 - either manually or through automation - to provide active/passive north/south egress.



Although using local egress removes the need to redeploy the universal control VM at site 2 upon site 1 failure, the likelihood of such an event is rare; an Edge or upstream failure is more likely. In that case, dynamic routing will recover automatically based on standard routing protocol and metrics, whereas a deployment using local egress would require manual intervention or development of an automated process to switch the locale ID to site 2. For this reason, the dynamic routing model with one universal control VM can be advantageous for active/passive north/south egress. Local egress and its uses are discussed in more detail in the [next section](#), which highlights its most common usage scenario - active/active site egress.

## b. Active/Active North/South with Local Egress

This deployment model is a multi-site multi-vCenter design with active/active north/south utilizing local egress. The local egress feature is described in detail before proceeding with an example of using it to achieve active/active north/south egress.

Active/active site egress deployment is less common than the active/passive north/south model with dynamic routing discussed in the prior section. Customers deploying a multi-site solution typically already have a reliable, high-bandwidth connection between sites. With this in place, the ease of deployment in leveraging one site for north/south traffic - typically about 15-20% traffic within a data center - offsets the complexity and overhead involved in configuring and maintaining an active/active site egress deployment.

Local egress is a feature that allows for localized north/south egress in a multi-vCenter, multi-site environment. Local egress must be enabled when the UDLR is first created as shown in Figure 7.22. Local egress is only available and only applies in a UDLR context. Local egress does not apply to a local DLR context.

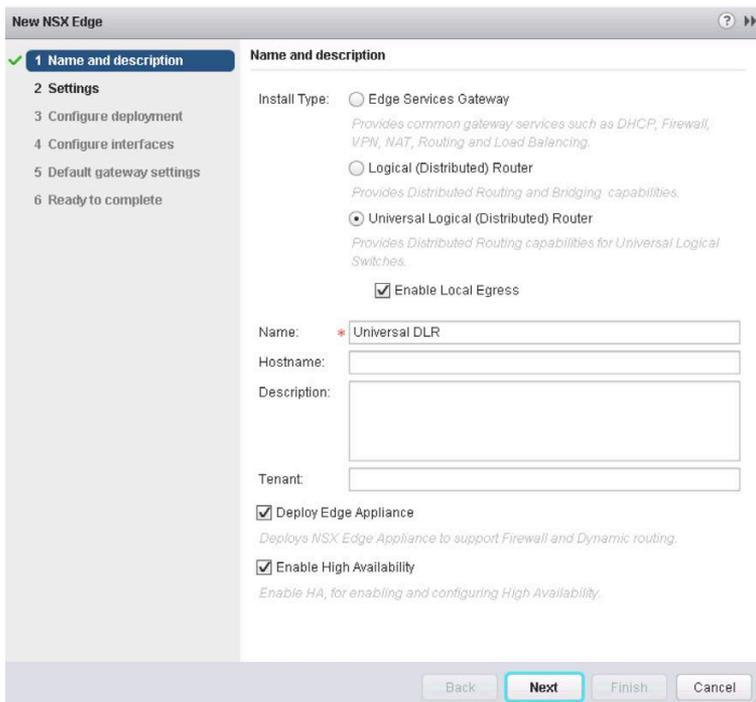


Figure 7.22 Enabling Local Egress Upon UDLR Creation

When local egress is enabled on the UDLR, the NSX Controller receives routes from the respective universal control VM. These routes will also be associated with a locale ID.



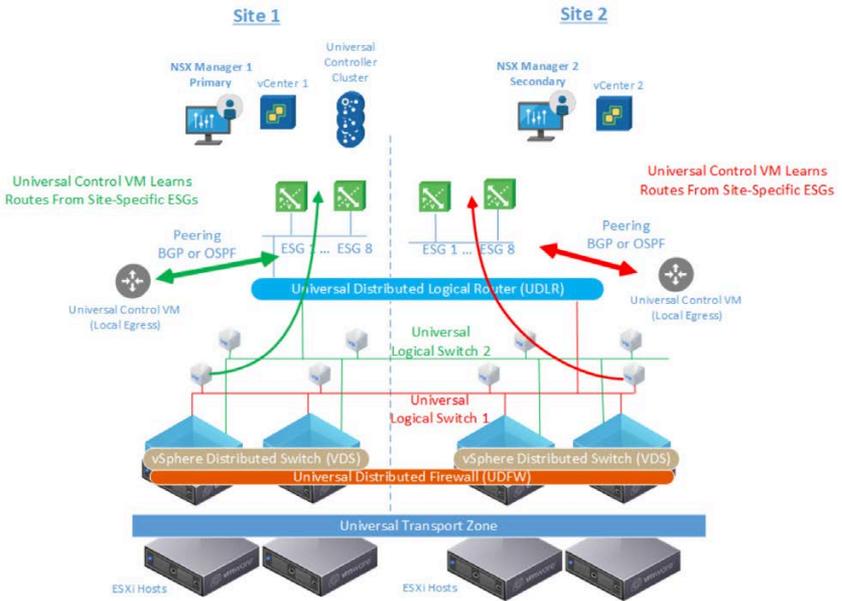
NSX 6.2 introduced locale IDs for enhancing multi-site deployments. By default, the locale ID in each NSX domain is set to match the local NSX Manager's universal unique identifier (UUID). All hosts inherit this locale ID, and it is used to determine which routes are propagated to which hosts when the controllers learn routes from DLR control VMs of different NSX domains/sites. By default, locale ID is set at a site level, but can also be set on a per cluster, host, logical router, or static route basis. The locale ID can be seen as a filtering mechanism for UCC to distribute site-specific routes to the correct hosts, enabling site-specific or local north/south egress.



Once the universal controller cluster learns the best forwarding paths, it sends the forwarding information only to ESXi hosts with a matching locale ID. This enhancement enables site-specific local routing configuration, but requires a universal control VM to be deployed at both sites. One UDLR will have a universal control VM at site 1 in the site 1 vCenter domain, and another universal control VM at site 2 in the site 2 vCenter domain. It is also possible to utilize local egress in a single vCenter/multisite deployment, though no universal control VM exists and only static routes are supported via NSX Manager configuration.

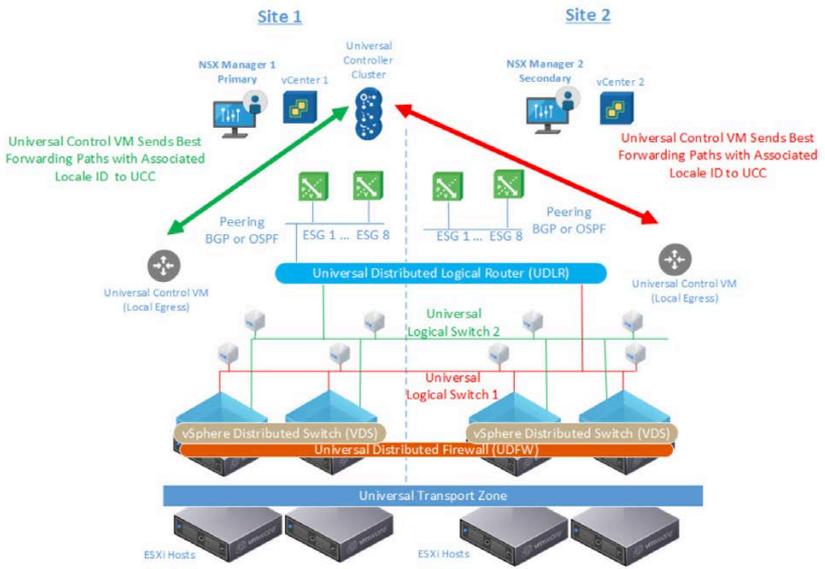
If local egress is not enabled on the UDLR, the locale ID value is ignored. Instead, all hosts in the NSX environment will receive the same routing information, which is the default behavior.

With local egress in a multi-site, multi-vCenter environment, a universal control VM is deployed within every NSX Manager domain at each site. The universal control VM at the secondary sites must be deployed manually. The respective universal control VMs peer via BGP or OSPF with the site-specific ESGs and learn the routing information as shown in Figure 7.23.



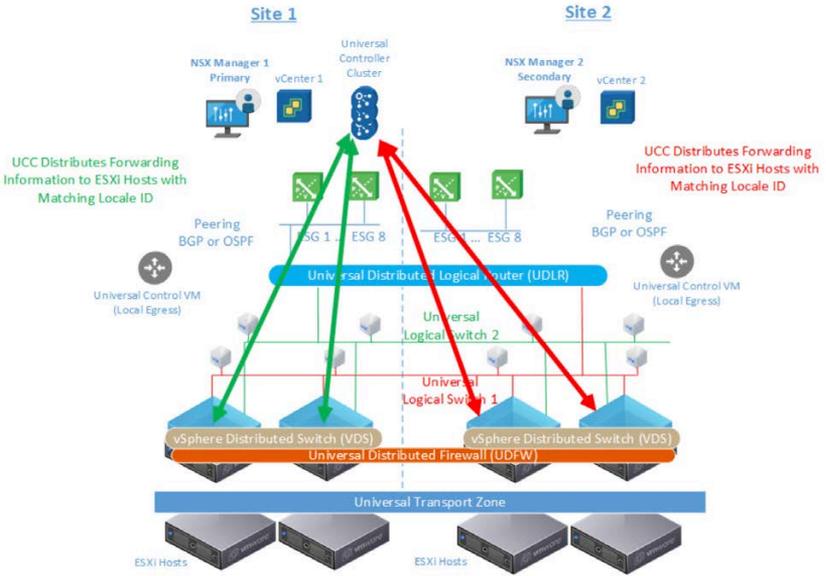
**Figure 7.23** Universal Control VMs Learn Routes From Site-Specific ESGs

Once the universal control VMs learn the routing information, the best forwarding paths are pushed to the UCC as depicted in Figure 7.24.



**Figure 7.24** Universal Control VMs Sends Best Forwarding Paths with Associated locale ID to UCC

Once the UCC has the best forwarding paths, it uses the locale ID to push the appropriate forwarding information to ESXi hosts across all sites with matching locale IDs (Figure 7.25).



**Figure 7.25** UCC Uses locale ID to Distribute Forwarding Information to ESXi Hosts with Matching locale ID

ESXi hosts at the respective sites use the forwarding information distributed by the UCC, enabling site-specific active/active north/south egress as shown in Figure 7.26.

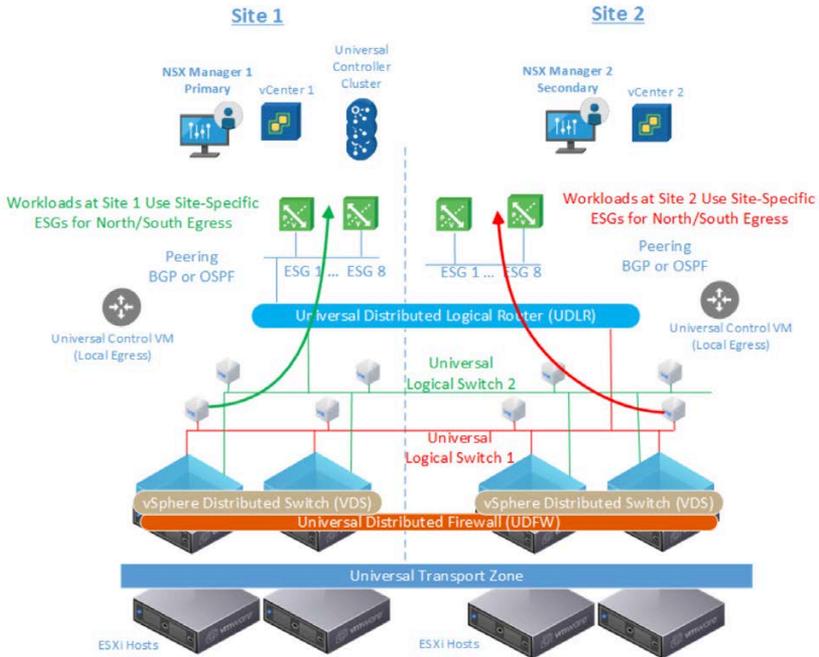


Figure 7.26 Site-specific Active/Active North/South Egress Enabled by Local Egress

Locale ID can be set in various places for specific scenarios:

- **Site Level:** Most common and default behavior. All ESXi hosts at a site inherit the locale ID of the local NSX Manager. No change required.
- **Cluster Level:** Second most common. Useful in DR scenarios. Requires locale ID change on the cluster.
- **Host Level:** Useful for specific scenarios such as single vCenter designs where clusters are stretched across two sites. The locale ID must be changed on the ESXi host.
- **UDLR Level:** Used only for inheritance of locale ID for static routes. Locale ID must be changed in NSX Manager on the UDLR.
- **Static Route Level:** Only supported in scenarios with no control VM. Locale ID must be changed in NSX Manager on the specific static route.

Figure 7.27 shows how the locale ID can be changed at the cluster level. Select the specific cluster and select **Change Locale ID** from the **Actions** button or by clicking on any of the gray configuration icons under the Installation Status, Firewall, or VXLAN columns, then select **Change Locale ID**. The default locale ID inherited by the local NSX Manager is also visible, as shown in Figure 7.28.

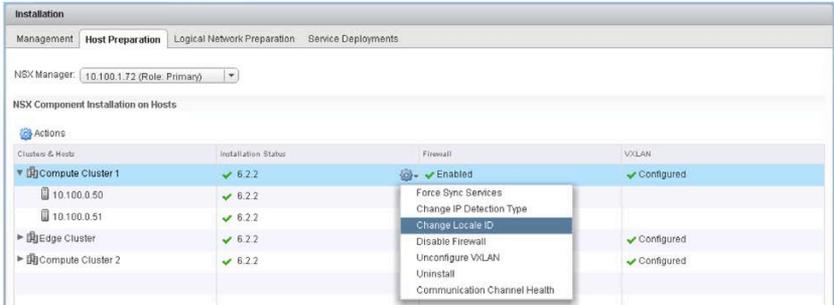


Figure 7.27 Changing locale ID At The Cluster Level

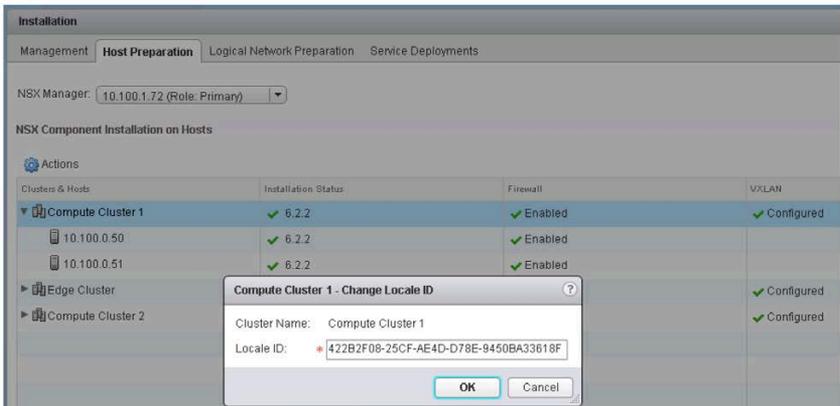


Figure 7.28 Default locale ID Inherited From Local NSX Manager

The UUID used by NSX is based on the IETF RFC 4122 standard. It can be modified manually following the standard guidelines. A programming language can also be used to generate a UUID automatically; there are standard libraries available for popular programming languages that provide this functionality. An example of automatic UUID generation based on the RFC 4122 standard can be

found on the official Python website (<https://docs.python.org/2/library/uuid.html>).

The same procedure can be used to change the locale ID at a host level; instead of selecting a cluster, select the respective ESXi host.



Local egress can provide granular control, allowing one workload to use one egress point and another workload to use another egress point. Local egress can also enable active/passive site egress; this is detailed in the following solution

where it is used in a DR scenarios to orchestrate north/south traffic flow for partial or full application failure.

In Figure 7.29, SRM orchestrates disaster recovery for partial application failure. Since only part of the three tier app failed to site 2, site 1 continues to be used for north/south egress. The locale ID of the cluster at site 2 is reset to match that of the site 1 NSX Manager UUID so that the workload that failed over to site 2 can still leverage site 1 for egress. The workflow for updating the locale ID at the recovery site cluster is part of the SRM recovery plan.

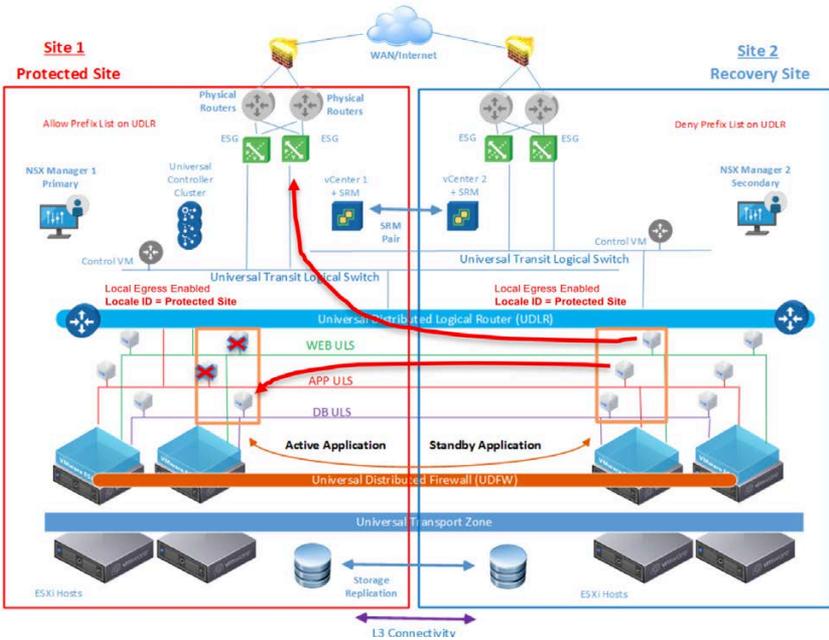
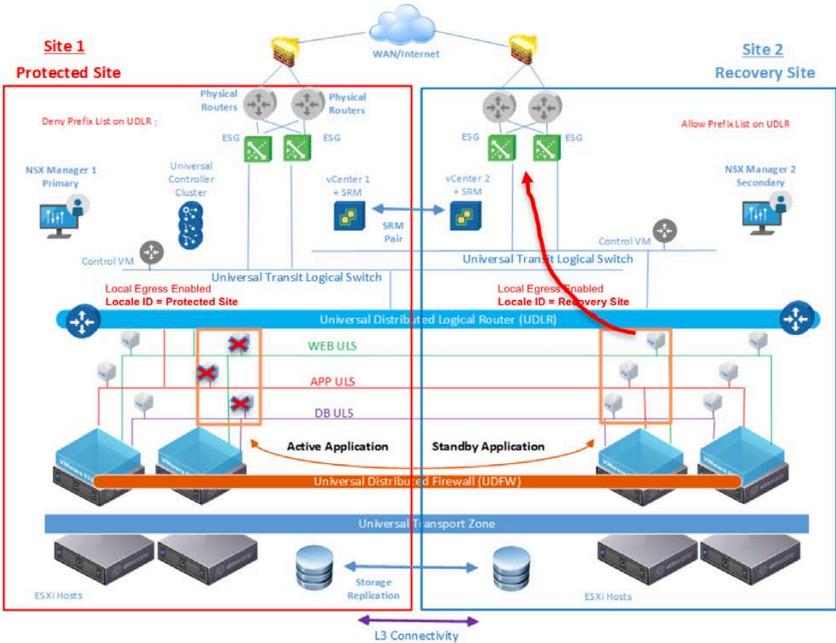


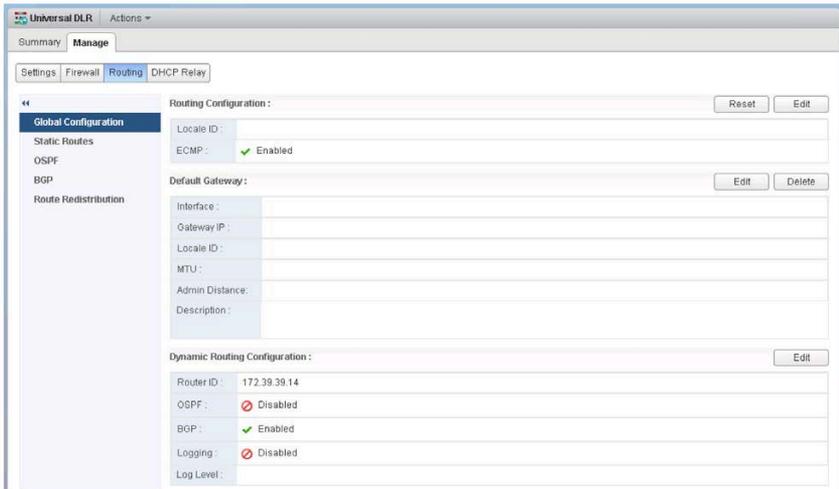
Figure 7.29 Upon Partial Application Failover, Site 1 N/S Egress Still Used

Figure 7.30 shows the entire application failed over to site 2 via SRM disaster recovery orchestration. The SRM workflow/runbook includes changing the locale ID of the cluster to the site 2 NSX Manager locale ID and advertising networks out of site 2 by changing the deny prefix list to an allow prefix list.



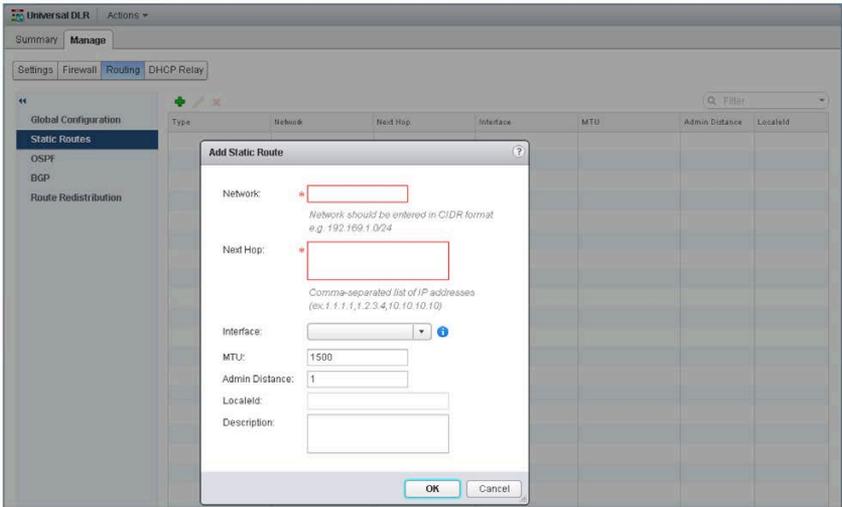
**Figure 7.30** Upon Full Application Failover or Edge Failure, Site 2 N/S Egress Used

Figure 7.31 shows where locale ID can be changed at the UDLR level.



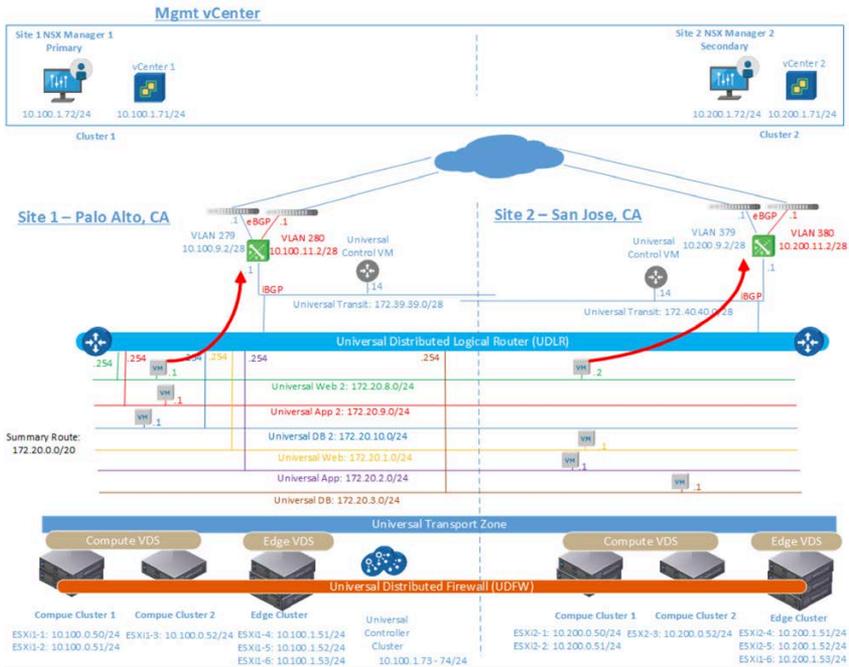
**Figure 7.31** Changing locale ID At The UDLR Level

The locale ID can be changed at the static route level. This is shown in Figure 7.32. Locale ID at a static route level can only be set if there is no universal control VM deployed. If a universal control VM is deployed, the text box to set the locale ID is grayed out.



**Figure 7.32** Changing locale ID At The Static Route Level

For the design shown in Figure 7.33, a universal control VM is deployed at both sites. Workloads at site 1 use site 1 ESGs for egress and north/south, and workloads at site 2 use site 2 ESGs for egress. The status of the UDLRs at the primary and secondary sites all show **Deployed**. For simplicity of demonstration, Figure 7.33 shows only one ESG per site being used, with each ESG providing ECMP northbound. In a production environment, multiple ESGs should be deployed at each site for additional resiliency as discussed in the Edge Component Deployment and Placement section.



**Figure 7.33** Multi-site with Multiple vCenters & Active/Active Site Egress

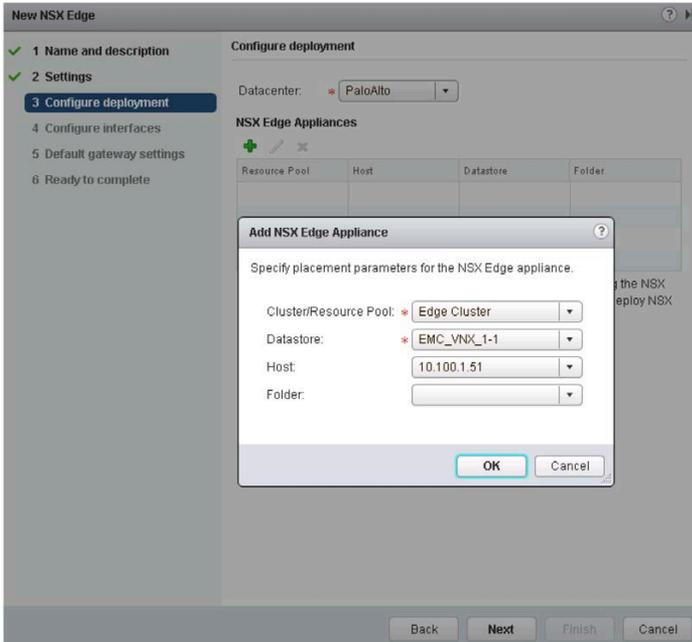
The main difference between this deployment and the deployment of the prior section is that this deployment has active/active site egress whereas the previous one had active/passive site egress. In the previous example, all traffic for both sites uses site 1 ESGs for ingress/egress. Since this example portrays an active/active egress deployment, local egress is required for a universal control VM to be deployed at both sites. Figure 7.33 shows a universal control VM deployed at each site on its own universal transit network/logical switch. The separate transit networks for each universal control VM are necessary to keep the routing information learned from the ESGs isolated to each site.



Local egress must be enabled when the UDLR is initially created by selecting the **Enable Egress Option** as shown in Figure 7.22. The UDLR must be created first with the **Enable Egress Option** and **Deploy Edge Appliance** fields checked; this will deploy a single universal control VM at the primary site. **Enable High Availability** can also be checked to deploy a standby universal control VM for high availability. The universal control VM for the secondary sites will be deployed once the user selects the respective cluster, datastore, host, and HA interface for the universal control VM

on the secondary vCenter. There is only one UDLR, but each site has its own associated universal control VM.

Figure 7.34 shows a universal control VM deployed at the primary site. When the UDLR is deployed, the placement details for the universal control VM are required. An interface for HA communication must be selected regardless of whether HA will be enabled; this can simply be a logical switch designated for HA communication.



**Figure 7.34** Deploying Universal Control VM at Primary Site

Figure 7.35 shows the UDLR deployed with a status of **Deployed** at the primary site. This means a universal control VM has been deployed on the respective NSX Manager.

NSX Edges				
NSX Manager: 10.100.1.72 (Role: Primary)				
Id	Name	Type	Version	Status
edge-b7ffd8dc-dc19-4263-9fef-a...	Universal DLR 1	Universal Distributed Router	6.2.2	Deployed

Figure 7.35 UDLR With Status of Deployed at Primary Site

At the secondary site, the same UDLR reports a status of **Undeployed**, shown in Figure 7.36. Although the local egress option was selected when the UDLR was created, only one universal control VM has been deployed at the primary site. The UDLR instance is active on the hosts at the secondary site but no universal control VM has yet been deployed at the secondary site.

NSX Edges				
NSX Manager: 10.200.1.72 (Role: Secondary)				
Id	Name	Type	Version	Status
edge-b7ffd8dc-dc19-4263-9fef-a...	Universal DLR 1	Universal Distributed Router	6.2.2	Undeployed

Figure 7.36 Local Egress Configured But Universal Control VM Not Yet Deployed at Secondary Site

Double clicking the universal DLR at the secondary site and viewing the **Settings->Configuration** section under the **Manage** tab shows that no universal control VM has been deployed at the secondary site; Figure 7.37 highlights this.

Logical Router Appliances:

Name	Status	Host	Datastore	Folder
0 items				

Figure 7.37 Universal Control VM not Deployed at Secondary Site

Clicking the green '+' symbol, entering the specific location criteria (e.g., cluster, datastore, host), and selecting an HA interface automatically deploys a universal control VM. Once this is complete, the status of the UDLR at the secondary site changes to **'Deployed'** signifying a universal control VM has also been deployed at the secondary site. This status change is depicted in Figure 7.38.



**Figure 7.38** Local Egress Configured And Universal Control VM Deployed at Secondary Site

Similar routing protocol practices discussed in the [prior section](#) can be followed – either BGP or OSPF with the NSX logical network environment acting as a stub network. With BGP, the NSX domain with the UDLR and ESGs can leverage a private ASN, employing iBGP between the ESGs and universal control VMs and eBGP between ESG and physical network.

The earlier examples provide details on traffic flow when the local egress feature is utilized.



In an active/active deployment, give special consideration to asynchronous traffic flows. Workloads exit out of both sites for north/south traffic; this is based on site specific routes enabled by the local egress feature. It leverages locale ID to allow for route filtering by NSX Controllers when forwarding information is distributed to the ESXi hosts across sites. The forwarding information is distributed to the ESXi hosts via the NSX Controllers.

Since both sites are advertising the same logical network that spans both sites, it is possible for asynchronous traffic flows to occur where traffic exits one site and returns to another. This can cause inefficient traffic flows or dropped sessions when stateful services are in the path.

In certain environments with no stateful services, asynchronous traffic flows may not be a concern; thus a local egress solution without any control mechanisms is perfectly acceptable.

For environments where stateful services are in the path, or where asynchronous traffic flows are a concern, solutions such as global server load balancing (GSLB) or route injection can be used.

## Global Server Load Balancing (GSLB)

GSLB is a solution allowing redirection of ingress traffic to the appropriate site when multiple sites advertise the same network or application. A GSLB based solution works like DNS, with the additional ability to monitor the health of applications/virtual IPs (VIPs) and update DNS accordingly.

GSLB directs the ingress traffic to the correct site based on a DNS load balancing policy. In this deployment, GSLB is acting as an authoritative name server and is resolving DNS for incoming requests.

The GSLB device has special software that communicates with a site load balancer (SLB) at each site. Information such as site/application health, response time, and site load/connection is reported back to the GSLB from each SLB. Round-trip time (RTT) or load balancing based on region is used to determine client proximity to the data center. Each data center has a GSLB device from the same vendor. GSLB leverages the learned metrics to determine the path for incoming traffic.

In such deployments, no changes are necessary in the logical IP addressing in the NSX environment. The IP addresses of the web or any Internet-facing applications will have a corresponding public VIP served by the SLB.

A complete GSLB solution discussion on different deployments is beyond the scope of this document. It is highly recommended to refer to vendor documentation for details of GSLB deployments for active/active north/south solutions.

An example deployment using F5 Networks BIG-IP DNS for GSLB functionality is shown in Figure 7.39. F5 BIG-IP LTMs act as the SLBs and are deployed at each site in HA mode. F5 BIG-IP DNS provides the GSLB functionality with one appliance at each site to monitor the local LTMs and VIPs. All appliances are virtual appliances. The local LTMs monitor the local application pools and communicate status to the F5 BIG-IP DNS.

The local SLBs perform source network address translation (SNAT). If client traffic hits the VIP of the SLB, it will automatically return via the SLB, providing for site-local ingress/egress. Since the UDLR remains the gateway of the workloads, all traffic not initiated from a client as well as from workloads not part of the application pools, the path from UDLR to ESG will be used for ingress/egress.

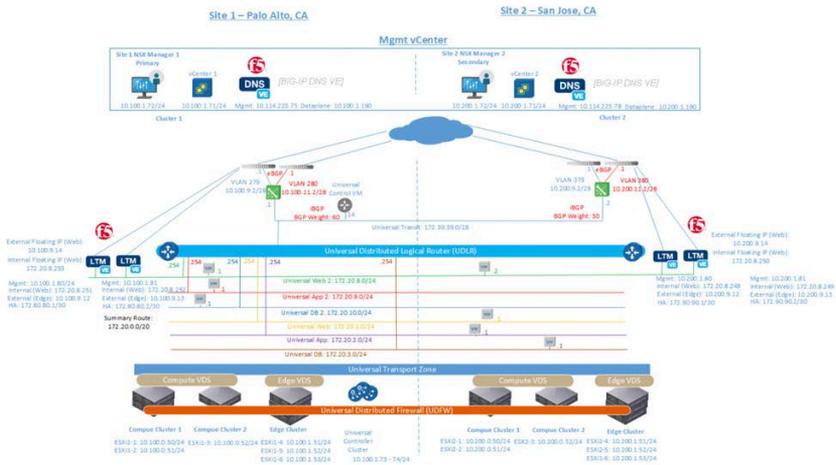


Figure 7.39 Cross-VC NSX Active/Active Deployment Leveraging F5 Networks GSLB

Member Order	Status	Member	Address	Partition	Member Name	Service Port	Ratio	Server Name	Data Center	Ratio
1	On	/Common	10.100.0.52/24	Common	/CommonWeb_VIP	80	1	BIOP1-1	Palo Alto	1
2	On	/Common	10.100.0.52/24	Common	/CommonWeb_VIP	80	1	BIOP1-2	Palo Alto	1
3	On	/Common	10.100.0.52/24	Common	/CommonWeb_VIP	80	1	BIOP1-3	San Jose	1

Figure 7.40 Load Balancing Methods available via F5 Networks GSLB

## Route Injection

It is possible to advertise a /32 route from the NSX ESG to route ingress traffic back based on a specific application IP address. This can be automated via NSX REST API and scripting/orchestration, such that when workload vMotion occurs from one site to another, a /32 route is injected into the network. This is a possible solution for a private network, but it is not practical for Internet facing applications as ISPs will generally not allow the advertisement of /32 routes.

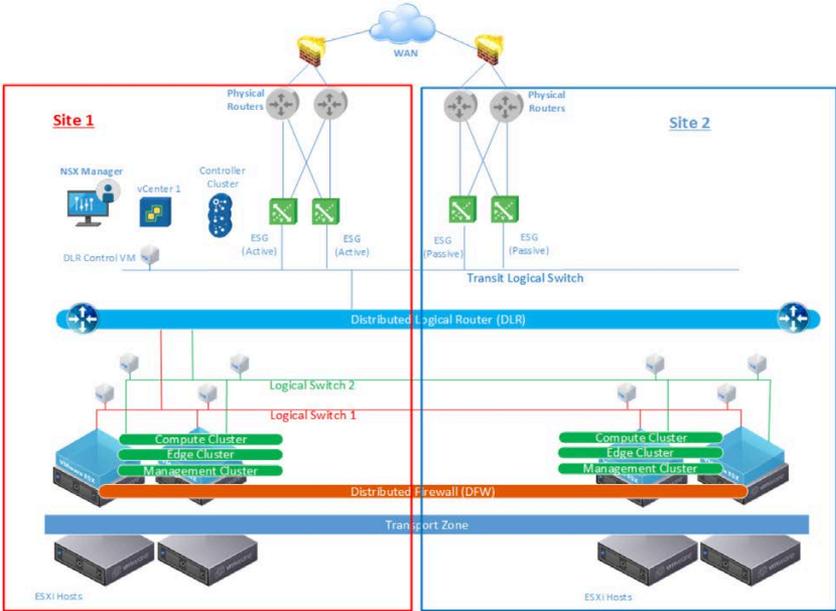
## 2. Multi-site with Single vCenter

### a. Active/Passive Site Egress with Routing Metric or Local Egress

A multi-site solution using a single vCenter with active/passive site egress should utilize the routing metric approach with a single control VM rather than utilizing local egress.

Although local egress can support an active/passive egress model, this scenario only supports static routing. Additionally, path failover upon ESG or upstream connectivity failure is not automatic as the locale ID of each static route would need to be manually changed.

Leveraging dynamic routing with a single universal control VM allows for a more flexible and dynamic model where failover is handled automatically. In this design, a single control VM is deployed and logical networks are stretched across two sites. Since there is only one vCenter, there is no requirement to use universal objects/networks, logical networks are simply stretched across sites, and local egress is not used. Figure 7.41 shows how the desired multi-site solution with consistent networking and security across sites is achieved without using universal objects or Cross-VC NSX.



**Figure 7.41** Multi-Site Solution with One vCenter and No Universal Objects

As discussed in the NSX Multi-site Solutions section, a vMSC multi-site solution can be deployed with NSX, a single vCenter, and no universal objects/networks. Figure 7.42 demonstrates this vMSC solution where vSphere clusters and storage are stretched across sites within metro distances.

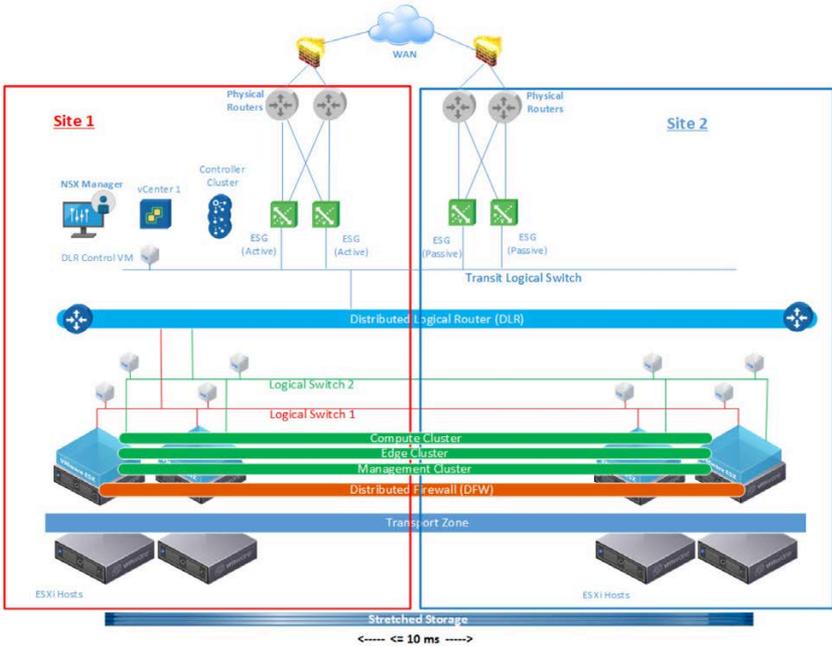


Figure 7.42 vMSC Multi-Site Solution with One vCenter and No Universal Objects

Since the vSphere clusters are stretched across sites in the vMSC with NSX solution, the native vSphere HA and DRS capabilities can be leveraged across sites. Additionally, this solution inherently provides DA and basic DR. Management components (e.g., vCenter, NSX Manager, NSX Controller cluster) and ESGs can run at site 1 and can restart at site 2 via HA if needed. It is still possible to run ECMP ESGs across both sites for resiliency, but failover of ESG to another site is also fully supported via vSphere HA.



If universal networking and security with Cross-VC NSX is under consideration, leveraging universal objects can help future-proof the solution. There is only one vCenter, so there is only one primary NSX Manager and no secondary NSX Managers.

All workloads at both sites would use site 1 for egress north/south, so the UDLR status across sites would be **Deployed**. Also in this example, both ESGs at each site could do ECMP northbound. In a production environment, multiple ESGs should be deployed at each site for resiliency as discussed in previous examples. The ESGs in Figures 7.41 and 7.42 could simply be deployed in HA mode with active/standby if

10 GbE north/south is sufficient and stateful services such as ESG firewall, load balancer, or NAT are desired at the edge.

## b. Active/Active Site Egress with Local Egress

Local egress can be utilized in single vCenter scenarios to achieve active/active site-specific north/south egress. Examples include a vMSC solution or a NSX multi-site solution with separate clusters. In these same scenarios, local egress can be used to achieve active/active egress as shown in Figure 7.43. In this case, there is no universal control VM, and only static routing is supported.



The static routing configuration is done directly on the NSX Manager. The following process must be followed to enable static routing for two different locale IDs:

- Deploy a UDLR instance with local egress enabled and no Edge appliance.
- Enable ECMP on the UDLR.
- Make an NSX REST API to configure a default gateway via the first locale ID.
- Make an NSX REST API to configure a static route for 0.0.0.0/0 via the second locale ID.

In this design, no universal control VM is deployed, and logical networks are stretched across two sites. There is only one vCenter, so there is only the primary NSX Manager and no secondary NSX Managers. Workloads at site 1 use site 1 ESGs for egress and workloads at site 2 use site 2 ESGs. The UDLR status across sites is **Active**. In a production environment, multiple ESGs should be deployed at each site for additional resiliency as discussed in the Edge Component Deployment and Placement section.

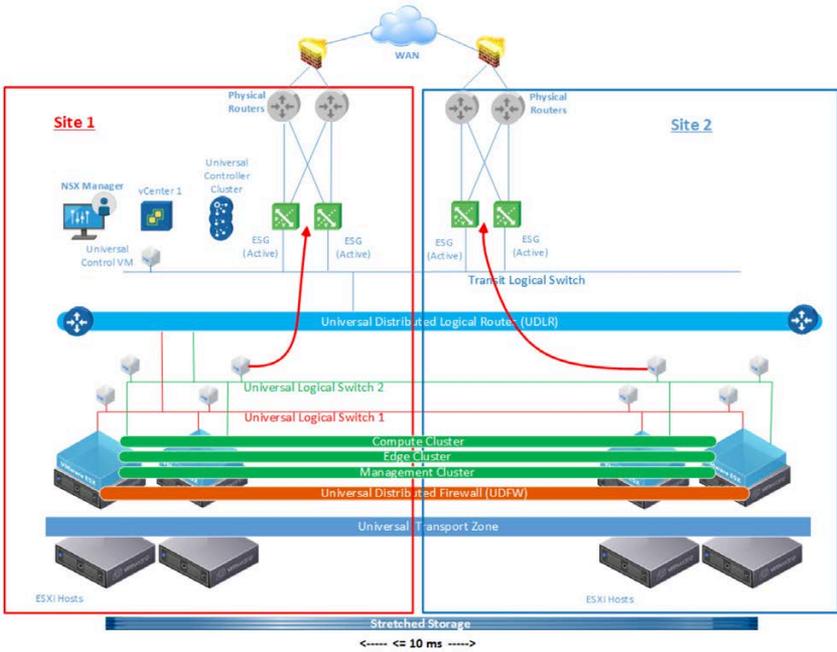


Figure 7.43 Multi-site Solution With One vCenter and Local Egress With Static Routes



# Summary

VMware NSX provides multiple multi-site data center options. The Cross-VC NSX feature offers solutions and flexibility for multi-site networking and security across multiple vCenter domains, which may also be located at different sites. Workloads are no longer constrained to vCenter boundaries. Consistent networking and security policy enforcement can be achieved across a multi-site multi-vCenter environment without additional manual fabric intervention.





# Index

## A

API 54, 82, 83, 121, 130, 131, 164, 167  
ApplyTo 95, 96

## B

BGP 73, 74, 120, 122, 131, 132, 133,  
134, 135, 136, 139, 148, 161  
Broadcast, Unicast, Multicast  
(BUM) traffic 57, 69, 76, 78,  
79, 80, 81, 106, 144

## C

Controller Disconnected Operation  
(CDO) Mode 60, 61, 62, 63, 64,  
68, 69, 144  
Control Plane 22, 30, 42, 45, 53,  
55, 57, 60, 62, 63, 64, 65, 66,  
67, 68, 69, 72, 77, 104, 106, 112,  
115, 118, 140, 144  
Cross vCenter 98  
Cross-VC NSX 2, 3, 18, 19, 23, 24,  
25, 30, 31, 33, 35, 36, 37, 38,  
39, 40, 41, 42, 45, 47, 50, 54,  
59, 60, 61, 62, 71, 72, 74, 75, 76,  
77, 82, 84, 90, 96, 97, 98, 99,  
100, 103, 104, 106, 107, 114, 115,  
117, 118, 119, 123, 125, 126, 127,  
128, 139, 140, 145, 146, 163, 164,  
166, 170

## D

Data Plane 53, 60, 63, 66, 112, 113,  
144  
Deployment Models 3, 19, 72, 103,  
119, 125, 126, 127  
Disaster Avoidance 3, 8, 9, 18, 21,  
36, 37, 166  
Disaster Recovery 3, 7, 8, 9, 16, 18,  
21, 23, 25, 30, 36, 39, 40, 61,  
68, 74, 83, 90, 91, 94, 98, 145,  
146, 152, 154, 155, 166  
Distributed Firewall 17, 35, 39, 42,  
54, 84, 85, 97  
Distributed Logical Router 35, 118,  
122, 123, 129, 139, 140, 147, 148,  
160

## F

F5 Networks 162, 163  
Firewall Rules 42, 84, 86, 87

## G

Graceful Restart 122, 139  
GSLB 161, 162, 163

## I

IPSet 42, 86

## L

L2VPN 2, 7, 18, 19, 26, 27, 28, 29,  
30  
Load Balancer 121, 122, 129, 133,  
162, 167  
Local Egress 21, 23, 25, 39, 42,  
119, 123, 126, 127, 139, 146, 147,  
148, 158, 160, 161, 164, 167  
Locale ID 146, 148, 150, 151, 152,  
153, 154, 155, 156, 161, 164, 167  
Logical Switch 17, 26, 35, 39, 41,  
56, 58, 59, 61, 62, 63, 64, 65,  
66, 67, 68, 69, 71, 72, 73, 75,  
76, 78, 80, 81, 95, 96, 132, 133,  
134, 144, 158, 159

## M

Micro-Segmentation 16, 19, 21, 23,  
25, 83, 95  
MPLS 5, 11, 106  
MTU (Maximum Transmit Unit) 3,  
21, 23, 25, 28, 29, 30, 104, 105,  
106  
Multi-site 2, 3, 5, 6, 7, 8, 9, 10, 12,  
15, 18, 19, 22, 24, 26, 29, 30,  
31, 36, 43, 60, 61, 69, 104, 106,  
114, 119, 120, 122, 125, 126, 127,  
128, 131, 140, 141, 147, 148, 158,  
164, 165, 166, 167, 168, 170

## N

NSX Edge Services Gateway 21, 23, 25, 28, 39, 56, 72, 74, 75, 119, 120, 121, 122, 126, 129, 130, 131, 132, 133, 134, 136, 138, 139, 140, 142, 143, 144, 145, 148, 149, 157, 158, 161, 166, 167

NSX L2 Bridge 76

NSX Manager 38, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 63, 64, 66, 72, 77, 82, 84, 85, 86, 87, 88, 90, 91, 94, 96, 97, 99, 106, 110, 111, 112, 114, 115, 116, 117, 118, 119, 123, 128, 144, 146, 148, 152, 153, 154, 155, 159, 166, 167

## O

OSPF 73, 74, 120, 131, 139, 148, 161

OTV 12

## P

Palo Alto Networks 99, 100

Physical Underlay Network 104

Platform Services Controller 46, 107, 108, 109

Primary NSX Manager 41, 45, 47, 49, 51, 56, 88, 91, 110, 111, 116, 117, 119

## R

Resource Pooling 8, 9, 18, 30, 36

REST API 54, 82, 83, 121, 130, 164, 167

Route Injection 161, 164

Routing 15, 21, 23, 25, 53, 56, 72, 75, 80, 121, 122, 126, 129, 131, 132, 133, 134, 139, 146, 147, 148, 149, 158, 161, 164, 167

## S

Secondary NSX Manager 41, 43, 46, 48, 49, 50, 51, 52, 54, 55, 85, 88, 90, 116, 117, 118, 144

Security 2, 3, 9, 15, 16, 17, 18, 19, 21, 23, 24, 25, 30, 33, 34, 35, 36, 37, 38, 39, 42, 46, 48, 82, 83, 84, 86, 87, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 104, 106, 164, 166, 170

Distributed Firewall 39, 58, 84, 86

Groups 17, 39, 42, 81, 82, 86, 87, 89, 92, 94, 97, 98, 100

Policy 35, 38, 39, 93, 94, 95, 96, 162, 170

Tags 16, 42, 86, 90, 91, 94, 95

Service Composer 87, 97, 100

Spanning Tree Protocol 10, 11, 106

SpoofGuard 97

Stateful Services 21, 23, 25, 74, 121, 122, 129, 130, 133, 139, 161, 167

## T

Traffic

North/South 5, 18, 28, 29, 74, 98, 106, 119, 120, 121, 123, 126, 127, 128, 129, 130, 139, 140, 142, 143, 146, 147, 148, 151, 154, 157, 161, 162, 166, 167

## U

Universal Controller Cluster 50, 53, 112, 115, 120, 148

Universal Control VM 25, 118, 119, 120, 122, 131, 132, 139, 141, 143, 144, 146, 148, 149, 150, 156, 157, 158, 159, 160, 161, 164, 167

Universal Distributed Firewall 35, 42, 58, 82, 83, 84, 85, 86, 87, 95, 96, 98, 99

Universal Distributed Logical Router 23, 35, 41, 42, 56, 72, 73, 74, 118, 119, 120, 123, 126, 128, 129, 131, 132, 139, 145, 147, 148, 152, 155, 156, 158, 159, 160, 161, 162, 166, 167

Universal Security Groups 42, 86, 89, 90, 92, 94, 97, 100

Universal Security Tags 42, 86, 90, 91, 92, 94, 95

Universal Synchronization Service 41, 44

## V

vCenter 3, 18, 20, 21, 22, 23, 24, 25,  
34, 35, 36, 37, 38, 39, 41, 42,  
44, 46, 48, 50, 52, 53, 54, 58,  
59, 77, 81, 82, 83, 86, 87, 91, 96,  
98, 104, 106, 107, 108, 109, 110,  
112, 113, 114, 115, 116, 117, 118, 119,  
120, 122, 125, 126, 127, 128, 129,  
140, 147, 148, 152, 159, 164, 165,  
166, 167, 168, 170

Virtual Network 1

vMSC 19, 20, 30, 104, 165, 166, 167

VPLS 10, 11

vRealize 130

vSphere 17, 19, 20, 21, 22, 23, 24,  
25, 30, 35, 41, 54, 60, 86, 91,  
107, 115, 165, 166

VXLAN 16, 17, 19, 21, 23, 25, 26, 27,  
28, 29, 30, 49, 79, 80, 81, 82,  
104, 105, 106, 113, 115, 130, 153

## W

Workload Mobility 3, 9, 16, 18, 36,  
37

Customers deploy multi-site data center solutions for use cases such as workload mobility, resource pooling, unified logical networking and security policies, and disaster avoidance/recovery.

Unfortunately, traditional multi-site solutions have been inadequate in helping customers solve the many challenges they face when deploying applications across multiple data centers/sites. Furthermore, these traditional multi-site solutions have been network-only focused, failing to provide a holistic solution covering all aspects of the application including security and automation.

VMware NSX solves these traditional challenges by decoupling the network services and intelligence from the physical infrastructure by leveraging a network overlay and providing a complete network, security, and automation platform. VMware NSX provides a complete holistic multi-site solution for customers.

NSX offers many options for multi-site data center connectivity to allow workload mobility or disaster recovery across sites, and, these solutions are discussed and compared in this book. However, the main focus of this book is on a feature called Cross-vCenter NSX (Cross-VC NSX). Cross-VC NSX allows for consistent networking and security across multiple sites and across multiple vCenter domains.

In this book we walk through what multi-site is and its traditional challenges, discuss some of the short-comings of traditional multi-site solutions, compare different multi-site solutions provided by NSX, discuss the advantages of NSX and how it can be used for different use cases, and go into detail on Cross-VC NSX for multi-site solutions.

With Cross-VC NSX, we start from the basics of understanding what it is and then dig deeper into Cross-VC NSX architecture, features, and design. The goal of this book is to provide a comprehensive overview of why NSX has become so popular for multi-site deployments, and also provide a good understanding of Cross-VC NSX for you to get started designing and deploying your own successful multi-site solutions with NSX.

**vmware** PRESS

Cover design: VMware  
Cover photo: iStock / alexsl

[www.vmware.com/go/run-nsx](http://www.vmware.com/go/run-nsx)

ISBN-13: 978-0-9986104-6-7  
ISBN-10: 0-9986104-6-1



\$12.99