Virtual Design Master Season 2

Final Submission: The Rainmaker - Single Host Nested Private Cloud

Challenge 4: Beyond The Clouds

Byron Schaller
8-9-2014

Contents

1 Overview

1.1 Executive Summary

The preparations are complete and the time has come to make our way from our temporary lunar home to our new lives on Mars. The ships making the journey will serve as housing, lab space, and transport. Once on Mars they will also serve temporary shelter until the colony can be built.

A system has been designed to run a non-flight related functions during the journey and provide necessary support during the build phase. Resources on the lunar base are tightly constrained and as such each ship has been allocated one Dell M610 server to house this system. For several reasons, including the small size and cloud technology it serves (as well as our dreams for Mars) we have come to call this systems “Rainmakers”.

Each ship will only contain one Rainmaker system, for this reason they are self-contained. However once we have established the New Netherlands colony on Mars all Rainmakers will be networked into the Elysium Grid, so they are built with that future in mind.

One of the last things discovered on Earth before it was left behind was a Dell warehouse with the M610 servers. Luckily there were just enough servers to allocate one per ship. The downside of the limited numbers is that each server has 16 GB of RAM. The Rainmaker systems have been customized to account for this limitation.

With only one server per ship and several services needed the design team has specified that each server should run a nested VMware vSphere environment to best use the limited resources while still maintaining service separation. To further containerize the application servers, Docker has been implemented.

The application development and operation teams on the space ship have been trained in agile continuous development processes and will run applications in a development environment before deploying to production.

1.2 Requirements

Identifier	Requirement
R01	The design must be a nested vSphere infrastructure.
R02	The design must include VMware vCenter.
R03	The design must include VMware vCenter Orchestrator.
R04	The design must include VMware vCloud Director.
R05	The infrastructure must host at least one accessible Windows guest.
R06	The infrastructure must host as least on accessible Linux guest.
R07	The design must include a working network overlay.
R08	The design must contain a Docker deployment.

1.3 Constraints

Identifier	Constraint
C01	A Dell M610 has been chosen as the physical host for the environment.
C02	The M610 is configured with 1 processor with 4 cores.
C03	The M610 is configured with 16 GB of RAM.
C04	The M610 is configured with 1 physical uplink NIC.
C05	The system can use 5 public IP addresses.
C06	All Management and Host Cluster networking must be done with distributed switches.

1.4 Assumptions

Identifier	Assumption
A01	Networking not on a nested vSphere host does not need to use distributed switches.
A02	Clusters can contain only one host.
A03	Application will need a development and production environment.

1.5 Risks

Identifier	Risk
K01	The single physical host is a single point of failure.
K02	The management cluster only contains one host, making it a single point of failure.
K03	The resource cluster only contains one host making it a single point of failure.
K04	The physical host only contains one physical NIC. No reasonable way has been found to use distributed switches outside the nested Management and Resource clusters due to this constraint.
K05	Powering on the Orchestrator Appliance causes massive instability.

2 Physical Design

2.1 Physical Host Design

It has been stipulated in constraints C01-C04 that the single physical host of the Rainmaker system must be a Dell M610 server with the following configuration:

Type	Component
CPU	1x2.13 GHz Harpertown E5506
Memory	16GB DDR3-1066
Hard Disk	1 x 250.0GB 2.5" SATA 7200RPM
NIC	Gigabit Ethernet

The following table details the total compute capacity for the system.

Metric	Capacity
CPU Cores	4
CPU Ghz	8.52
Logical CPUs (w/ Hyper-Threading Enabled)	8
Memory	16 GB

2.1.1 Operating System

The host will be installed with VMware vSphere 5.5 Enterprise Plus.

2.1.2 Naming Convention

Each unique Rainmaker system will be configured with a UUID for a hostname at the domain rainmaker.local. This is to ensure that when the systems join the Elysium Grid on the Martian Colony each server's host name will be unique.

2.2 Physical Storage Design

The host is configured with a single 250GB SATA drive. This will be formatted with VMFS 5 when vSphere is installed. The name of the local datastore has been changed to "BaseDisk" to avoid confusion with the nested hosts "local" datastores.

2.3 Physical Networking Design

As documented in constraint C04 and referenced in Risk K04 the Dell M610 server only has one physical network interface card (NIC).

With a single NIC connected and the vCenter virtual machine hosted on the same server there is no reasonable way to use vSphere Distributed Switches on the physical host (Risk K04). Several possible methods to accomplish this goal were conducted in the proof of concept lab to work around this limitation. None were successful, and some efforts went so far as to greatly impact the stability of the test host and require a complete rebuild of the test environment.

With this limitation it was decided to use two vSphere Standard Switches. vSwitch0 is configured with two port groups, one for the Management VMKernel interface and one VM Networking named "Core".

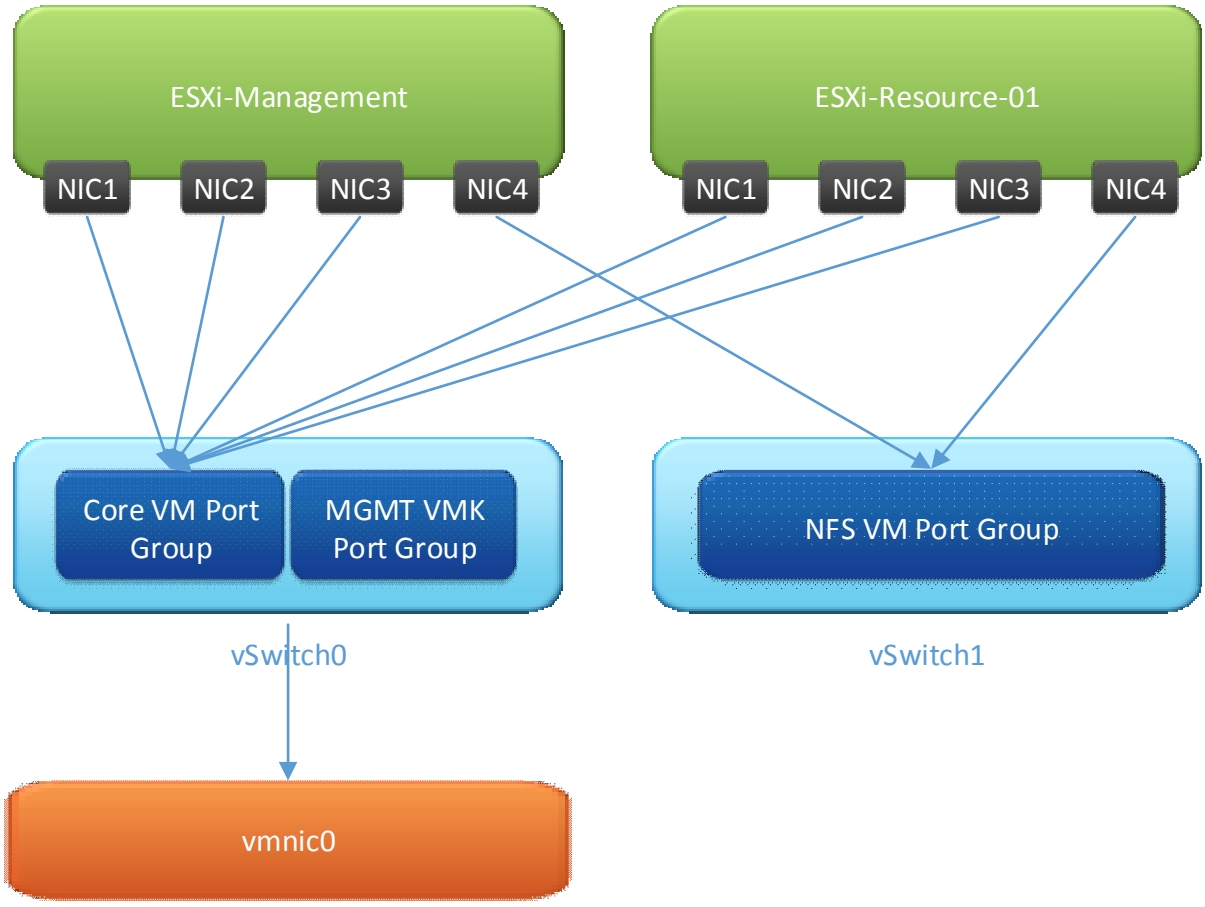
vSwitch1 contains one VM Networking Port Group named “NFS“ for the NFS. vSwitch0 is connected to the vmnic0 uplink. vSwitch2 is connected to no uplink as it is for internal NFS traffic only.

NICs 1, 2 and 3 of both nested vSphere hosts are connected to the Core VM Port Group with NIC 4 being connected NFS port group on vSwich1.

It is important to note a few design considerations about the vSphere networking on the physical host:

1. When passing traffic from a virtual NIC to a port group on a nested vSphere host then, routing to an uplink that is a virtual NIC connected to a VM Network port on the physical host, VLANs don’t work.
2. Every path in the network aside from traffic destined for a public address in virtual and operates at internal bus speed, not external line speed.
3. Failover and redundant paths become much less important when dealing with virtual NICs as a “hardware” failure is several orders of magnitude less likely to happen.
4. Link Aggregation of virtual links for bandwidth reasons is not needed due to the chance of the virtual line being saturated is basically zero.

The configuration is detailed in the following diagram:



All traffic on the “Core” Port Group resides on the 10.0.0.0/24 network.

All traffic on the “NFS” Port Group resides on the 10.0.10.0/24 network.

3 Management Cluster Design

3.1 Design Decision on Availability

Due to constraints imposed by the physical limits of the Dell M610 hardware several compromises had to be made in regards to the design of the nested vSphere hosts and the Clusters that contain them. The largest area of compromise is that of availability, or potential lack thereof.

The goal of maximizing availability is a potential gain and has steep diminishing returns. The extreme limit on physical memory (per constraint C03) made the return even more costly. After several tests in the proof of concept lab environment it became clear that adding any nested hosts beyond one per cluster would not increase the overall system availability and would in fact make the systems less stable and lower the system's uptime.

At the conclusion of the lab based proof of concept the decision was made to build both clusters with one nested host due to the overwhelming evidence observed.

3.2 Nested Host Design

The Management Cluster is comprised of one host named "management.rainmaker.local". The host runs VMware vSphere 5.5 for its operating system and is configured as shown in the following table:

Component	Configuration
vCPU	4
vRAM	12 GB
SCSI Controller	LSI Logic Parallel
Hard Disk 1	250GB Thin
NIC 1	Switch0 > Port Group "Core"
NIC 2	Switch0 > Port Group "Core"
NIC 3	Switch0 > Port Group "Core"
NIC 4	Switch1 > Port Group "NFS"
Configuration > vhv.enable	TRUE
Configuration > hypervisor.cpuid.v0	FALSE

CPU and Memory were sized to meet the needs of the Management Cluster VMs while keeping them stable.

Due to all network traffic being internal the physical server. All network traffic aside from NFS is connected back to the "Core" port group on Switch0.

Storage is all direct attach. The decision was made to Thin Provision the data store as the total volume of data in the future is unknown.

The vhv.enable and hypervisor.cpuid.v0 configuration items are set to allow the guest to host and run 64 bit virtual machines.

3.2.1 Network Design

3.2.1.1 Virtual Switch

Per Constraint C06 the Management Cluster has been designed to use only vSphere Distributed Switches. Only one switch is deployed named "Management".

3.2.1.2 Uplink Group

There are 4 Uplinks in the Uplink Group. The following table details the mapping to vNICs

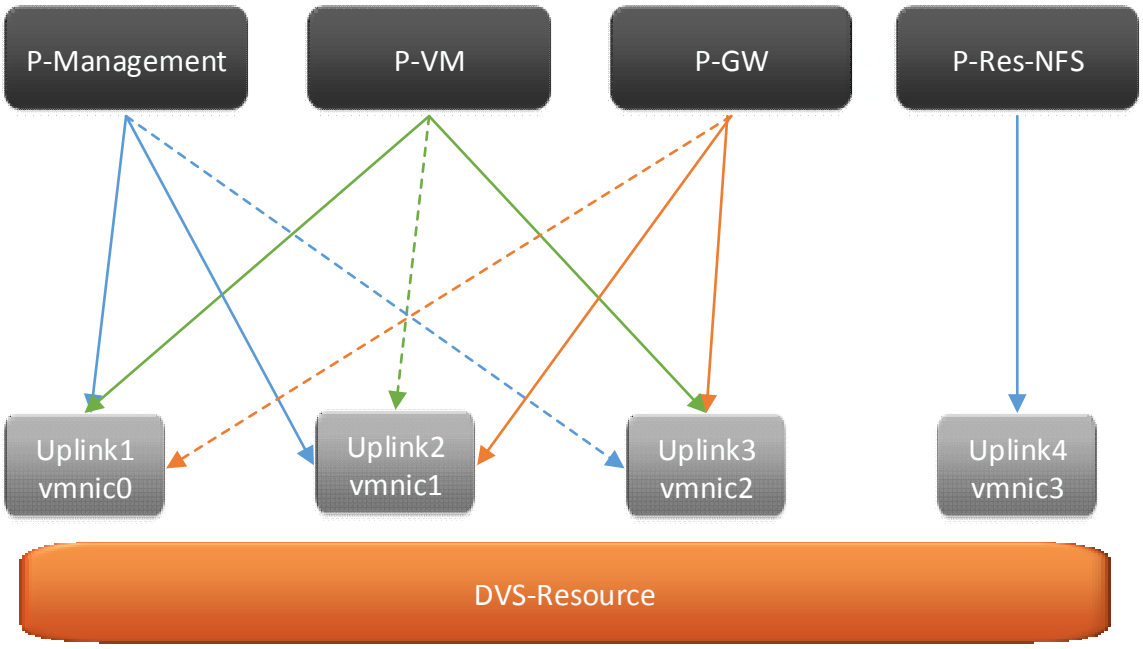
Uplink	vNIC
dvUpklink1	vnic0
dvUpklink2	vnic1
dvUpklink3	vnic2
dvUpklink4	vnic3

3.2.1.3 Port Groups

There are 4 Port Groups. They are detailed I the following table:

Virtual Switch	DVPortGroup	Network Ports	Load Balancing
DVS-Management	P-Management	dvUpklink1 dvUpklink2 dvUpklink3 (standby)	Route Based on Virtual Port ID
DVS-Management	P-NFS	dvUpklink4	None
DVS-Management	P-VM	dvUpklink1 dvUpklink2 (standby) dvUpklink3	Route Based on Virtual Port ID
DVS-Management	P-GW	dvUpklink1 (standby) dvUpklink2 dvUpklink3	Route Based on Virtual Port ID

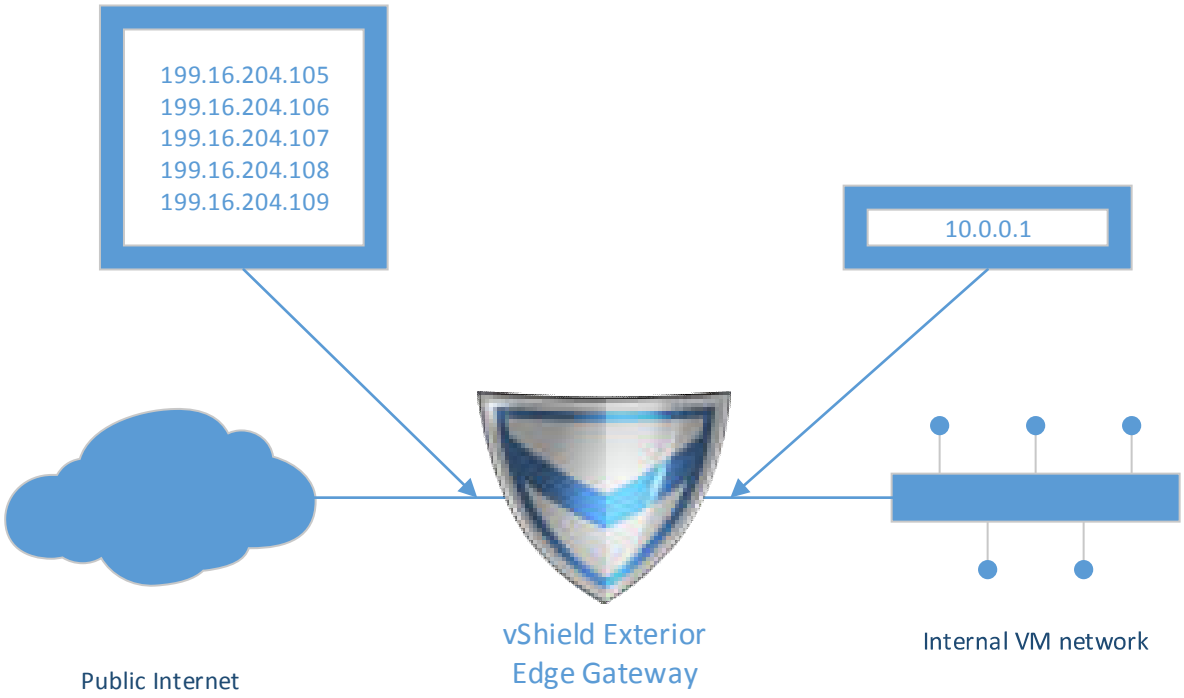
The following diagram illustrates the configuration:



3.2.2 vShield Gateway

A vShield Gateway is configured on the 10.0.0.0/24 network and hosted on the Management Cluster. The Gateway provides access to the public Internet as well as DNAT and SNAT functionality.

The following diagram illustrates this configuration:



The NAT rules are detailed in the following table:

Rule	Source	Destination
DNAT	199.16.204.107	10.0.0.10
DNAT	199.16.204.108	10.0.0.12
DNAT	199.16.204.109	10.0.0.13
SNAT	10.0.0.2-10.0.0.254	199.16.204.106

3.2.3 VM Sizing

The following table details all VMs hosted on the Management Cluster.

VM Name	vCPU	vRAM	Disk
EXT-edge-0	1	256MB	448MB
vCenter Server	2	4GB	125 Gb Thin
vCloud Director	1	2560MB	30 GB Thin
vShield Manager	2	4GB	60 GB Thin
vCenter Orchestrator	2	3GB	12 GB Thin
OpenFiler	1	512MB	170 GB Thin

3.2.4 Orchestration

Requirement R03 states that vCenter Orchestrator must be deployed in the environment. While it technically has been, it has been powered off to conserver memory.

When it is powered on it causes such constraint with the other VMs that is renders vCenter unusable as well as Orchestrator thus negating its use. The OVF has been kept as part of the system in case of future expansion.

This has been detailed as Risk K05.

4 Resource Cluster Design

4.1 Nested Host Design

The Resource Cluster is comprised of one host named “resource-01.rainmaker.local”. The host runs VMware vSphere 5.5 for its operating system and is configured as shown in the following table:

Component	Configuration
vCPU	2
vRAM	4 GB
SCSI Controller	LSI Logic Parallel
Hard Disk 1	250GB Thin
NIC 1	Switch0 > Port Group “Core”
NIC 2	Switch0 > Port Group “Core”
NIC 3	Switch0 > Port Group “Core”
NIC 4	Switch1 > Port Group “NFS”
Configuration > vhv.enable	TRUE
Configuration > hypervisor.cpuid.v0	FALSE

CPU and Memory were sized to meet the needs of the Resource Cluster VMs while keeping them stable.

The vhv.enable and hypervisor.cpuid.v0 configuration items are set to allow the guest to host and run 64 bit virtual machines.

4.1.1 Storage Design

Due to all network traffic being internal the physical server. All network traffic aside from NFS is connected back to the "Core" port group on Switch0.

Storage is direct attach and provided via NFS. The decision was made to Thin Provision the data stores as the total volume of data in the future is unknown.

The NFS mount is served by the OpenFIler NAS appliance hosted in the Management Cluster at 10.0.10.250.

The NFS Mount is located at /mnt/iscsi/nfs/esx. The entire 10.0.10.0 subnet has been grated read and write permission to the share in case of future expansion.

4.1.2 Network Design

4.1.2.1 Virtual Switch

Per Constraint C06 the Management Cluster has been designed to use only vSphere Distributed Switches. Only one switch is deployed named "DVS-Management".

4.1.2.2 Uplink Group

There are 4 Uplinks in the Uplink Group. The following table details the mapping to vNICs

Uplink	vNIC
dvUpklink1	vnic0
dvUpklink2	vnic1
dvUpklink3	vnic2
dvUpklink4	vnic3

4.1.2.3 Port Groups

There are 7 Port Groups. The 3 backed my VXLAN are not listed due to their transience. The permanent port groups are detailed In the following table:

Virtual Switch	DVPortGroup	Network Ports	Load Balancing
DVS-Management	P-Management	dvUpklink1 dvUpklink2 dvUpklink3 (standby)	Route Based on Virtual Port ID
DVS-Management	P-NFS	dvUpklink4	None
DVS-Management	P-VM	dvUpklink1 dvUpklink2 (standby) dvUpklink3	Route Based on Virtual Port ID
DVS-Management	P-GW	dvUpklink1 (standby) dvUpklink2	Route Based on Virtual Port ID

		dvUpklink3	

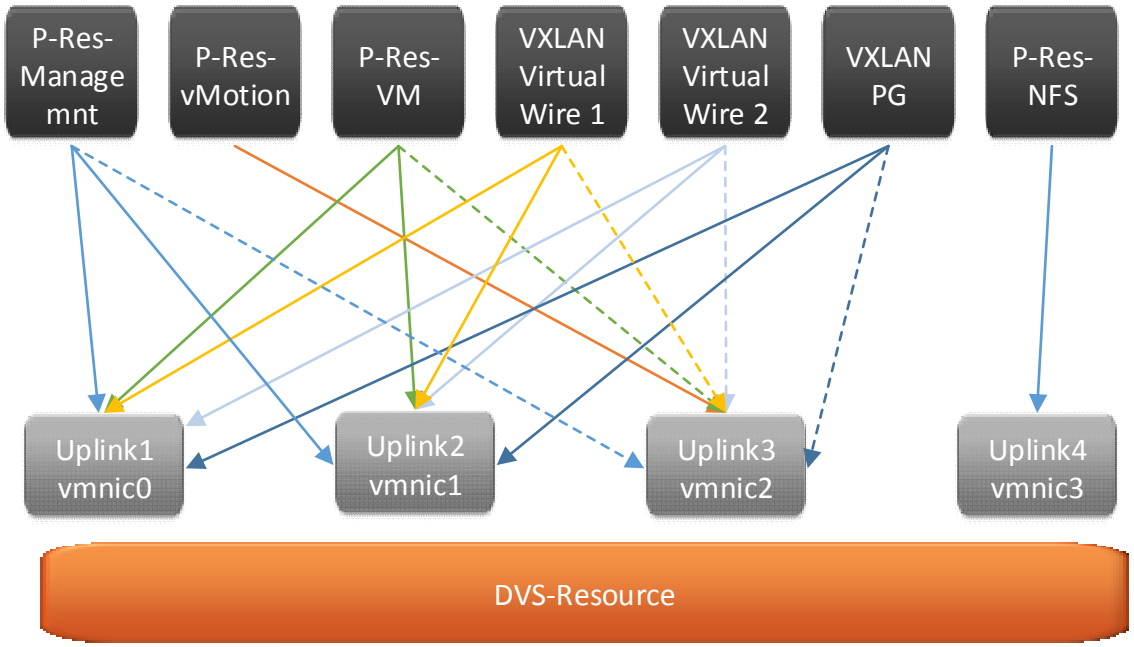
4.1.2.4 VXLAN

The Resource Cluster has been prepared by vCloud Networking and Security to present VXLAN backed network pools to vCloud Director.

The following table details the VXLAN settings

Setting	Value
Segment ID Pool	7000-9000
Multicast Addresses	232.1.42.0-232.1.62.0
VLAN	2101
Teaming Policy	Failover

The following diagram details the Network design:



4.1.3 VM Sizing

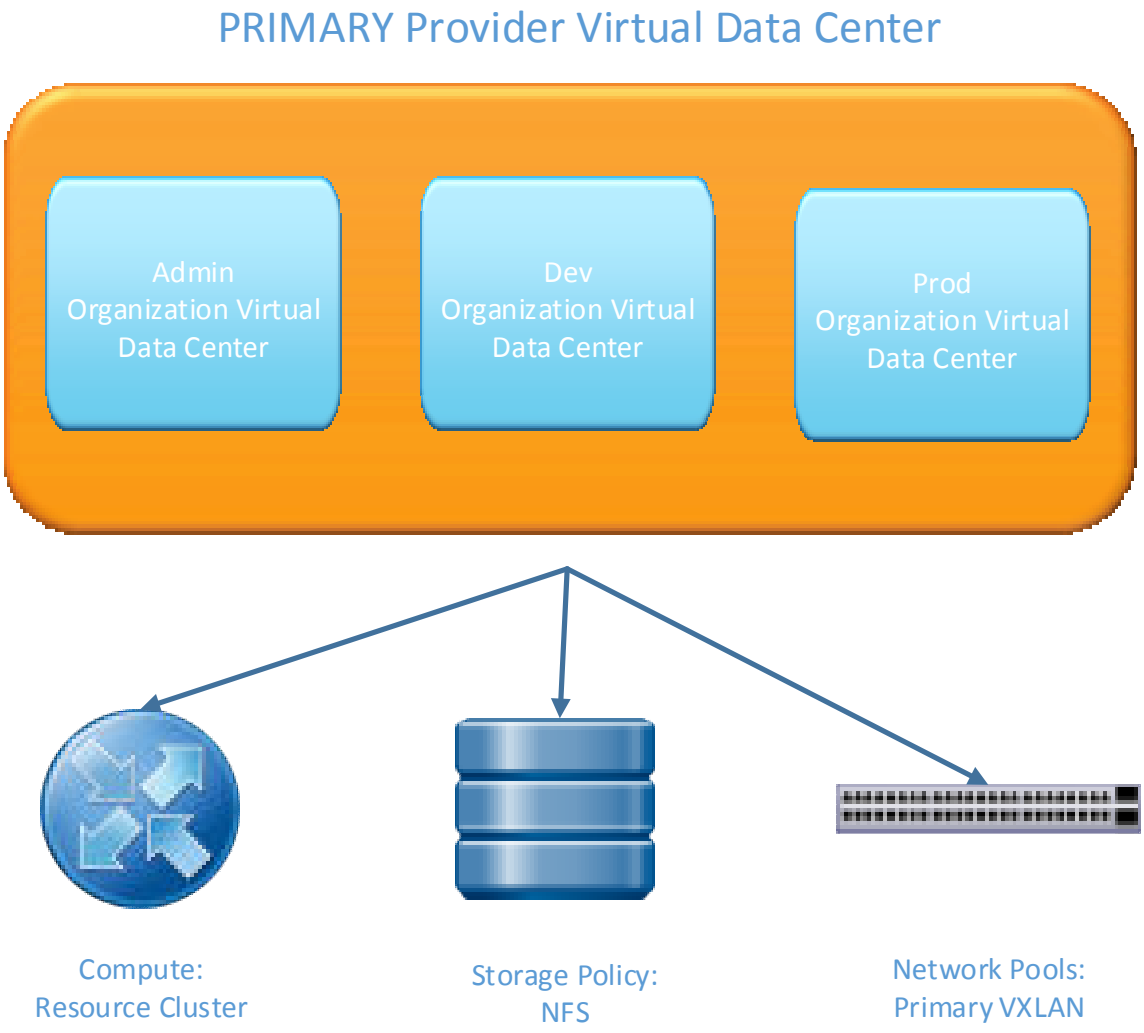
The Resource Cluster has hosts no VMs natively. All VMs are dynamically created through vCloud Director.

5 vCloud Design

5.1 Provider Virtual Data Center

There is one Provider Virtual Data Center (vDC) in the Rainmaker vCloud Design, its name is “Primary”.

Primary is backed by the vcenter.rainmaker.local VMware vCenter (v5.5u1) server and the vsm.rainmaker.local VMware vShield Manager (v5.5.2.1). This is illustrated in the following diagram:



The Resource Cluster provides its base resource pool ("Resource") of 2vCPUs and 4GB vRAM for the compute resources. Storage is presented by the NFS Storage Policy backed by the NFS-01 Data Store connected to the OpenFiler NFS share.

There is one External Network available to the Primary Provider vDC, "Res-EXT".

5.2 External and Internal Networks

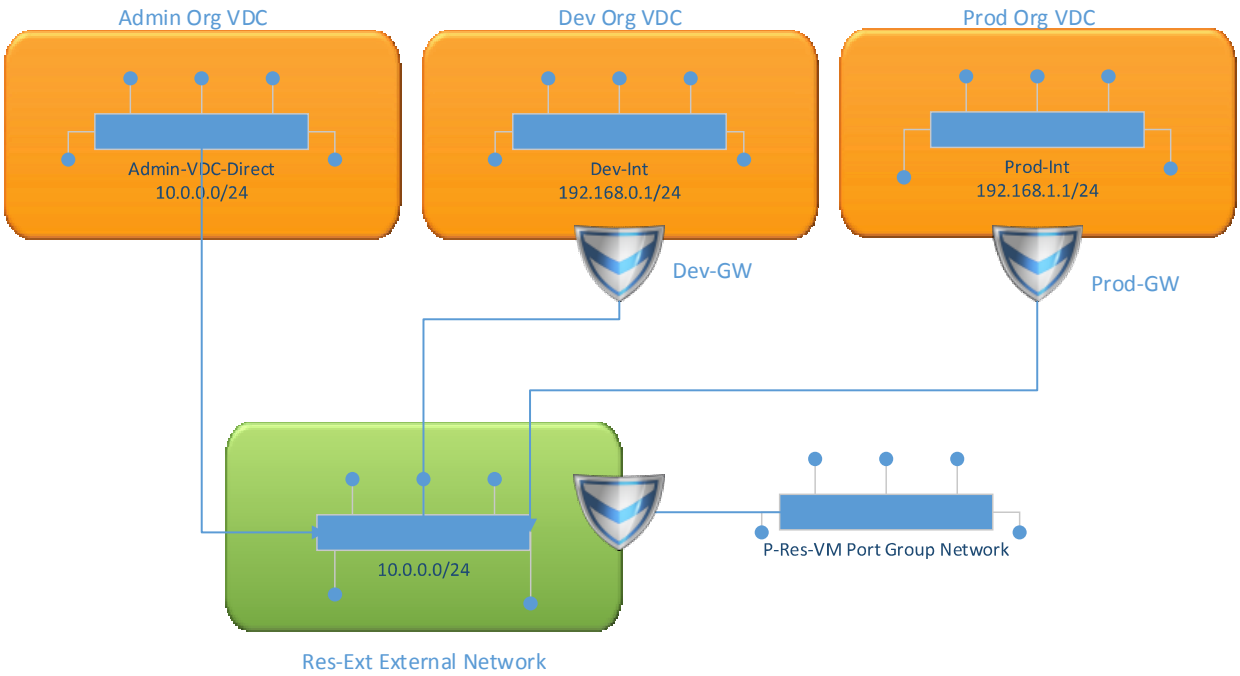
The Res-EXT External Network provides a gateway to the "P-Res-VM" Port group on the 10.0.0.0/24 subnet.

Res-EXT is connected to 3 Org vDC networks. The "Admin-VDC-Direct" Network is direct attached as shares the 10.0.0.0/24 subnet.

The "Dev-Int" network is routed through a vShield Edge Gateway, the internal subnet is 192.168.0.0/24.

The "Prod-Int" network is routed through a vShield Edge Gateway. The internal subnet is 192.168.1.0/24.

The following diagram details the vCloud network connectivity:



5.3 Organizations

There are three tenants of the Rainmaker vCloud: Admin, Dev, and Prod. All three contain one Organization vDC each and use the “Pay-As-You-Go” Allocation Model.

5.3.1 Admin

The Admin Organization owns the Rainmaker vCloud. The purpose of this tenant is for administration on the vCloud as well as testing and piloting changes.

The organization administrator account for the Admin Organization is named “admin”

The Admin Organization contains one Catalog named “Admin-Cat” that publishes vApps to an Org vDC named “Admin-VDC”.

5.3.2 Dev

The Dev Organization is owned by the development department. They are responsible for the software development life cycle of the containerized applications hosted by the Rainmaker vCloud.

The organization administrator account for the Dev Organization is named “dev-admin”

The Dev Organization contains one Catalog named “Dev-Cat” that publishes vApps to an Org vDC named “Dev-VDC”.

5.3.3 Prod

The Prod Organization is owned by the Operations Team. They are responsible for the day to day Rainmaker production environment.

The organization administrator account for the Prod Organization is named “prod-admin”

The Prod Organization contains one Catalog named “Prod-Cat” that publishes vApps to an Org vDC named “Prod-VDC”.

5.4 Staffing Considerations

5.4.1 The vCloud Center of Excellence Model

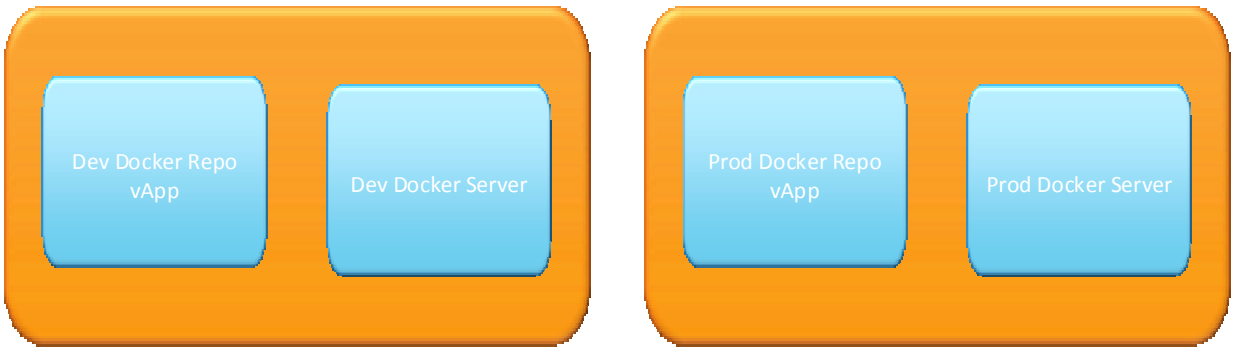
When operating production systems in a multi-tenant vCloud it is highly recommended to organize personnel in to “Center of Excellence” model. This will ensure that optimizations gained in process and tools will not be lost due to a failure of people.

More information on this model and how to organize fir a CoE can be found in chapter 4 of the VMware Press vCloud Architecture Toolkit (vCAT) Version 3.

6 Docker Application Design

The Docker infrastructure consist of 4 kinds of vApps in two Organizational VDCs.

The following diagram illustrates the vApps in both the Dev and Prod Org vCDs.



6.1 Container Lifecycle

6.1.1 Development

Developers create and test new containers in Dev Docker Server vApps. Images are saved on the Dev Docker Repo vApp.

6.1.2 Promotion

When a container has pasted testing, it is promoted to production, This happens by the image being copied from the Dev Docker Repo to the Prod Docker Repo and a new container being created from the promoted image.

6.1.3 Steady State

New containers are deployed based on demand from the Prod Docker Repo.

6.1.4 End of Life

When an image is no longer used for containers it is deleted from the Prod Docker Repo.

7 The Protocol Foxtrot Contingency System

7.1 System Overview

In an effort to mitigate the risks posed by K01, K02, and K03 and lack of redundancy to ensure availability the Protocol Foxtrot Contingency System (PFCS) has been developed and deployed with Rainmaker system.

Be default the PFCS is powered down as not to consume already constrained resources and is intended only to be powered on and accessed in the event of complete Whiskey Tango Foxtrot class event.

That being said it can also be used as a “jump box” to access the systems. This is accomplished by login directly into the physical vSphere host using the vSphere C# Client, powering on the system and connecting to the console. Once connected and logged in you will find a Microsoft Windows 7 desktop with a complete tool suite to troubleshoot and correct issues with the Rainmaker Cloud Systems.

7.2 The PFCS Tool Suite

The following tools and scripts are installed and available on the PFCS system:

7.2.1 Veeam Backup and Replication 7

With Veeam Backup and Replication backup images can be taken of the nested hosts. The images can be archived for restoration later in case of a nested host failure.

Once the colony on Mars has been established and more resources are available, Veeam Backup and Replication can be used as a full time backup and disaster recovery tool. The Elysium Grid is a requirement as an external backup or replication target is needed.

7.2.2 VMware vSphere C# Client 5.5 and PowerCLI for vSphere and vCloud Director 5.5

Both graphical based and command line tools are provided to access the physical and nested vSphere systems. Alan Renouf’s vCheck suite of Powershell scripts for vSphere and vCloud Director are also included as they provide quick access to power troubleshooting and reporting information.

7.2.3 RVTools

To further assist with troubleshooting Bobware's RVTools has also been included. This is another must have tool in the PFCS troubleshooting and correction toolkit.

7.2.4 vCloud Director Rest API Shell (RAS)

For extreme Whiskey Tango Foxtrot events when the only way to resolve issues is by making manual API calls to vCloud Director, the VMware Labs vCloud Director RAS Fling is the best tool to have.

7.2.5 Secure Shell Access

Both Putty and WinSCP have been installed to provide SSH and SCP access to the vSphere hosts and Virtual Appliances.

7.2.6 Web Browsers

Internet Explorer, Firefox, and Chrome (and their respective flash plugins) are all installed as some web browsers are better at certain tasks than others.