CHALLENGE 3 – DR IS EVIL!

(BACKUPS OUT OF THIS WORLD)

BY JAMES BROWN (@JBCOMPVM)

Contents

Executive Summary

Mars has been the primary focus. We now need to provide an offsite backup recovery and disaster recovery (DRBR) solution. Without a DRBR solution the human race could be in jeopardy. The offsite location has been designated as the Moon. The human race also has colonies on the moon surface. An advanced laser communication system is operational with a consistent 10ms latency all hours of the day.

The virtualization team must define PTO and RPO on for each application. We also need to explain what process and procedures will be required.

Requirements

1. Each colony must be capable of working fully independently.
2. Life support systems must meet a 99.9% service-level availability.
3. Greenhouse systems need to meet the same service-level availability.
4. The design must be highly reliable and easily deployable.
5. DR process and procedures must be in place.

Constraints

1. Network latency between the mars and the moon is 10ms
2. The design must incorporate the use of reliable, serviceable technology that can degrade gracefully over time.

Assumptions

1. Advanced laser communications has an uplink speed of 1GB/s.
2. Advanced laser communication is considered high redundant.
3. There is a Layer 2 stretched network
    a. 10.0.0.1/20 is used across Mars and the Moon
4. Three data centers have been established on Mars
5. Three data centers have been established on the Moon.
6. Data center are connect via a 40GB fiber ring on each planet.
7. All specified hardware and software can be acquired and will work on Mars.
8. Appropriate licensing for all vendor products (VMware, Microsoft, Red Hat, etc.).
9. vSphere administrators have and can maintain the skillsets required to implement and maintain the solution.

Risks

1. There is not existing disaster recovery solution in place.
2. No backup resource requirements have been collected.

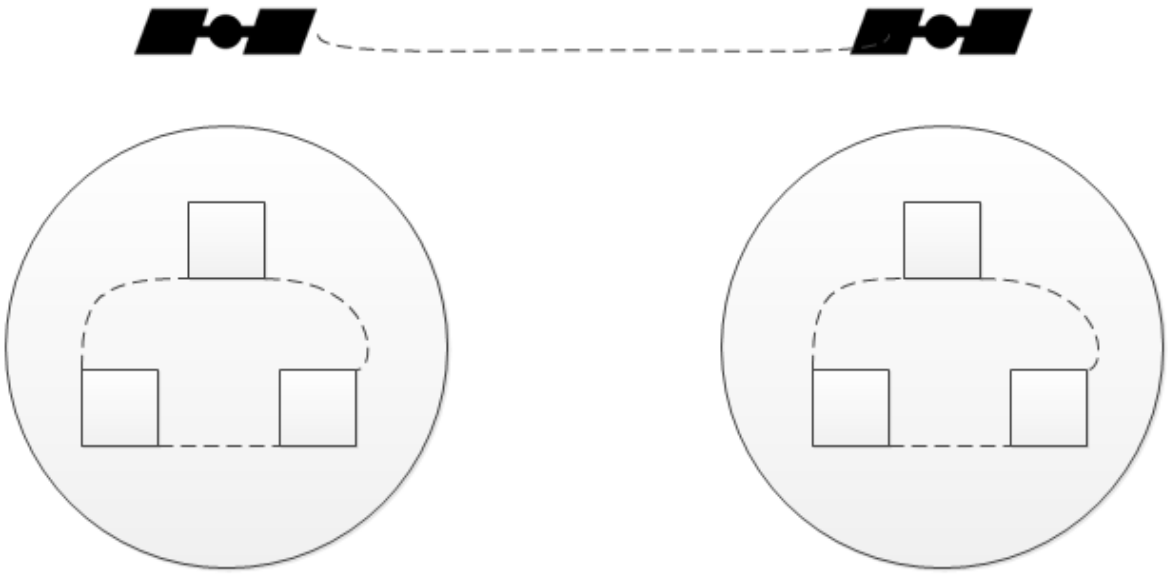
3. A lack of appropriate personnel may jeopardize the ability to maintain and improve the solution.
4. An alien ship disrupts the advanced laser communications.
5. Aliens attack.

Conceptual Design

Disaster Recovery

Building off challenge 1 and challenge 2, we know need to use the knowledge we have from these challenges to build a backup and disaster recovery plan. Mars has an advance laser network connection that is consistence at 10ms. The virtualization team has the choice to use the public cloud or build the infrastructure.

I choose the private cloud. I will design two infrastructures. One on the Moon and one on Mars. These two infrastructures will replication across the 1GB advance laser communication connection.



Moon

Mars

These servers have been identified:

Servers

Quantity	Function	vCPU	Memory (GB)	Storage (GB)	Total vCPU	Total memory
2	Front-End (Exchange)	4	12		8	24
3	Mailbox (Exchange)	4	12		12	36
15	Web Servers	2	8		30	120
3	Maria DB Cluster	1	4		3	12
5	File Server	2	12	2048	10	60
15	CoreOS (Docker)	4	12		60	180
2	NT 4	1	4		2	8

Architecture Design

Physical Design

Server Hardware

Nutanix NX-8000 Series rack mount systems have been chosen as the standard server hardware. The Nutanix platform is gaining popularity, and is becoming well known. There is very little training required to manage and maintain a Nutanix system. Basic users find the Prism interface very easy to navigate and use. This would make it an ideal choice for the vSphere admins and other admins who will be trained to support it. Rack systems allow greater flexibility to scale up and down without the constraints and limited points of failure of a chassis-based system. If a single node fails, service levels gracefully degrade on the remaining systems.

Based on hypervisor design specifications, the Nutanix NX-8035-G4 server has been chosen. This model has 2 10-Gigabit Ethernet interfaces for all. Each server has 2 E5-2697v3 (28 cores) CPUs for a total of 2 physical cores, and 256 GB RAM, two 1.6GB SDD Drivers, and four 6TB SATA drives. The system only supports 2 nodes per block. This number can increase or decrease depending on the true workload. The specific hardware configuration is found in the table below. The hardware will be connected to the network as described in the Networking Configuration section.

NX-8035-G4 base chassis w/o CPU/DIMM/HDD

Hardware	Quantity
E5-2697 V3	2
16 GB Memory	16
6 TB HSS	2
1.6 TB SSD	4
10 GB Network adapters	2
Nutanix Pro License	1

Networking Configuration

Brocade VDX switches work well with the Nutanix system and are offer a large number of Gigabit Ethernet ports, FCoE capability, and generous internal bandwidth. Like the Nutanix product line, the popularity of the Brocade switches ensures the systems are well known and training material exists for those unfamiliar with the product.

The VMware cloud's core will be a pair of Brocade VDX 6940 switches. The switches are capable for 36 40-Gigabit QSFP+ Ethernet interfaces, each 40-Gigabit interface can be broken out into 4 10-Gigabit interfaces, five fan modules and two power-supplies, providing redundancy within each unit. Each compute device will be connected to both switches and the switches will cross-connect to ensure that complete or partial chassis failure of either switch does not constitute an outage. The switches will be configured according to Brocades best practices guidelines to ensure efficient performance.

The models were chosen to provide room for growth or component failure. If workloads increase beyond 75% port utilization of any tier, the design will need to be revisited to properly accommodate growth without impairing long-term operational abilities.

Server Placement

Two Nutanix NX-8035 Blocks will be deployed at all six data centers. Each block contains enough compute and storage to run the entire infrastructure. Three data centers on Mars and three data centers are the Moon. Each data center will also have a Brocade VDX 6940 at the top of the rack.

Total RU used will be five.

Logical Design

Hypervisor Design

ESXi v6 will be installed on each Nutanix node via their foundation installation software. The hosts will have 2 physical CPUs, 256 GB of RAM, two 1.6TB SDD Drivers, and four 6TB SATA drives. Management, VM traffic, backup, and vMotion will be carried on redundant 10-Gigabit Ethernet interfaces per node. Storage is internal so it does not require any networking protocols. Special settings are required on the CVM to keep them on their respected controllers. Each host will be joined to a cluster and managed from vCenter. Local access is for last resort in case of emergencies.

vSphere Management Layer

vCenter v6 with vSphere Enterprise Plus will provide centralized management of the ESXi hosts, VMs, and features. vCenter will be installed in a VM to ensure availability via vSphere HA. Because the current workload is unknown the Windows Server version will be used.

vCenter Server will be installed with an External Platform Services Controller (PSC). The extended PSC is meant to allows multiple vCenter Servers to link to a PSC.

vCetner Server for Windows installation provides scalability and maintains low complexity. If VM guest numbers exceed the ability of this install, the components can be separated by the migrating the SSO and database components to additional VM guests.

vCenter SSO will connect to a Windows Active Directory domain. Users will be able to use the VMware vSphere Client and/or vSphere Web Client and their AD username and password for all access into vCenter.

vSphere Update Manager (VUM) will be used for both VMware and Windows patches. All patches will be approved and tested with 30 days of being released by vSphere administrator and then patches must be deployed no later than 60 days after testing.

VM Design

Initial system VMs are described here.

Microsoft licenses have been acquired. Windows Server 2012 R2 Datacenter Edition is the most recent server edition. Windows licensing allows the installation of 2012 R2 or any previous Windows Server edition, including NT 4. Exchange 2013 Enterprise licenses have been acquired for the mail server.

MariaDB is an open source license, no license will need to be acquired.

CentOS is an open source license, no license will need to be acquired.

Docker license has been acquired. The Production Edition will allow 1 Trusted Registry, 10 Docker engines, and business critical support. For the administrator among us that have no Docker experience.

As there is no current benchmarks to compare to, all resource allocations are estimations based on a combination of vendor guidelines and community best practices. Resource usages will be recorded via vCenter and requirements will be revisited every 30 days.

Management services

There is one installed set of management VMs. Clustering or distributed service guidelines will be followed according to vendor best practices if the workload determines that service levels are insufficient.

Application Design

Windows Domain and vCenter

General network and end user access will require integration with new or existing Active Directory forest. A new domain and forest will be created and two Windows 2012 R2 guests will be provisioned as domain controllers. Windows 2012 R2 Datacenter licenses have been acquired and all additional Windows guests will also run 2012 R2 unless otherwise specified. Additional domain-related VMs include Exchange and Skype for business. The vCenter server will be installed on Windows. This table lists the initial resource allocations and VM quantities.

Server	vCPU	RAM (GB)	OS Disk (GB)	Data Disk (GB)	Quantity
Domain Controllers	1	6	80	0	4
vCenter	2	16	80	100	2

Windows Files Server

These windows files servers will be used to store @vmiss33 collections of cat pictures and videos. Officials have tried to convince her to that 1TB of these artifacts is enough, but she has refused to listen.

These Windows files servers are setup with DFS. The 2TB of data is replicated every 30 seconds between these DFS servers.

Server	vCPU	RAM (GB)	OS Disk (GB)	Data Disk (GB)	Quantity
File Server	2	12	80	2000	2

Exchange Server

Windows 2012 R2 and Exchange 2013 have been selected for the mail system.

Server	vCPU	RAM (GB)	OS Disk (GB)	Data Disk (GB)	Quantity
Exchange - Front End	4	12	80	0	2
Exchange - Mailbox	4	12	80	3500	3

CentOS and Docker

CentOS and Docker will be setup and configure by the development team.

Server	vCPU	RAM (GB)	OS Disk (GB)	Data Disk (GB)	Quantity
CoreOS - Docker	4	12	100	0	15

MariaDB

CentOS will be the choice OS for the MariaDB install. To create a Galera Cluster we will need a total of three servers with CentOS and MariaDB installed and operational.

Server	vCPU	RAM (GB)	OS Disk (GB)	Data Disk (GB)	Quantity
MariaDB - CentOS	1	4	600	0	3

Webservers

CentOS will be installed with Apache.

Server	vCPU	RAM (GB)	OS Disk (GB)	Data Disk (GB)	Quantity
Webservers	2	8	120		15

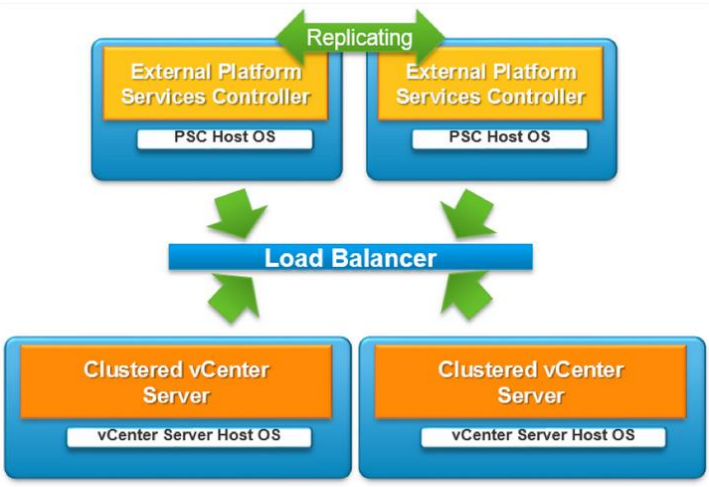
Legacy Software

Windows NT 4 is being used for legacy software. Support for this operating system ended on Dec 30, 2004. Someone really needs to re-write this application.

Server	vCPU	RAM (GB)	OS Disk (GB)	Data Disk (GB)	Quantity
Windows NT 4	1	4	60	0	2

VMware Datacenter Design

Two vCenter Servers will be defined. One on the Moon and one on Mars. They will be setup in a windows server cluster. External Platform Service Controllers (PSC) will be used on both planets. A failure of a PSC will not impact the usage of the infrastructure. The PSCs will be separated from each other physically using anti-affinity rules. The PSCs replicate state information vCenter Server nodes are individually clustered with WSFC for HA. The vCenter Servers interact with the PSCs through a load balancer.



To meet the 99.9% SLA, the cluster(s) will be configured with High Availability (HA) and Distributed Resource Scheduling (DRS). Due to the homogenous hardware that Nutanix uses in there systems, Enhanced vMotion Capability (EVC) is not required at this time.

If an EVC need did arise, cluster performance capabilities would be impaired without justification. The probability that EVC would be required would be low and the risk to the support systems will be higher. If future iterations of the design require EVC, risk can be mitigated and support system can conserved providing the current homogenous system and implementing a replacement heterogeneous system in the data centers.

HA will have an initial admission control policy of 20% of cluster resources to provide for 1 host failure (1/6 * 100) and will be revisited every 30 days as capacities increase and cluster size varies. Host Monitoring will be enabled with the default VM restart priority (Medium) and Host isolation response (Leave powered on). Critical VMs will have their restart priority increased. VM Monitoring will be disabled initially. The Monitoring settings will help avoid false positives that could negatively affect manufacturing and violate the SLA. They will be revisited within 24 hours of any HA-related outage to determine if changes are required to continue to meet the SLA, and again at the 30, 60 and 90 day marks.

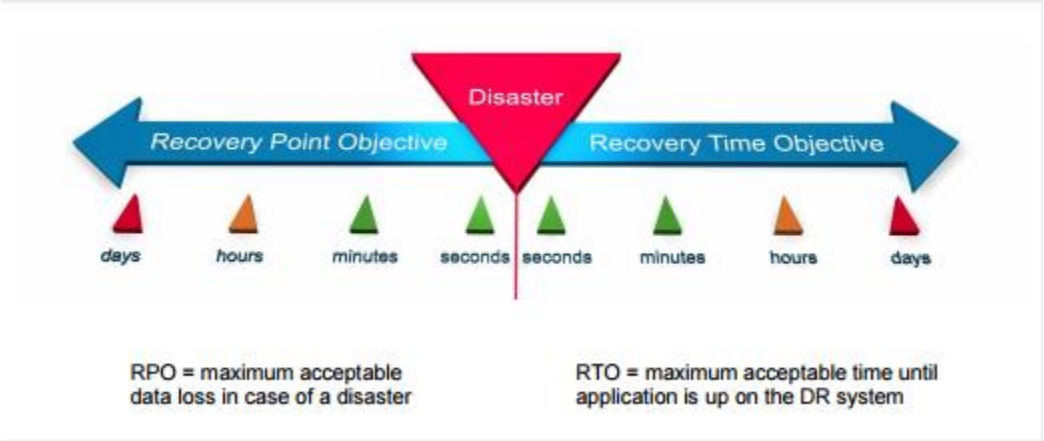
DRS will be configured as Fully Automated and to act on three star recommendations or greater. This will ensure the vSphere loads remain balanced across ESXi hosts as the support system scale.

A summary of initial HA and DRS rules are in the table below.

Rule Types	VMs
DRS VM-VM Anti-Affinity	DC1, DC2
DRS VM-VM Anti_Affinity	PSC1, PSC2
DRS VM-VM Anti-Affinity	MariaDB1, MariaDB2, MariaDB3
DRS VM-VM Anti-Affinity	Exchange – Front End
DRS VM-VM Anti-Affinity	Exchange – Mailbox Servers
DRS VM-VM Anti-Affinity	DFS1, DFS2
DRS VM-VM Anti-Affinity	WindowsNT1, WIndowsNT2
VM Override VM Restart Policy – High	Management - vCenter, DCs
VM Override VM Restart Policy – High	WindowsNT1, WindowsNT2
VM Override VM Restart Policy - Low	WebServers (15)

Backup and DR Architecture

DR project all start with the question of the potential data loss in case of a disaster (Recovery Point Objective or "RPO") and the time necessary to have the systems up and running on the DR server (Recovery Time Objective or "RTO").



Local DR Solution

Nutanix's native replication infrastructure and management supports a wide variety of enterprise topologies to meet real-world requirements, including:

1. Two-way mirroring
2. One-to-Many
3. Many-to-One
4. Many-to-Many

We will be using the One-to-Many metro replication for the data centers on Mars and on the moon. This replication cannot be used between Mars and the Moon, because the latency is over 5ms. In the One-to-Many replication there is one central site with multiple remote locations. The main tier-one systems run at data center 1, and data center 2 and 3 serves as remote back-up locations. The data center 1 systems can then be replicated to both 2 and 3 data centers. In the event of a local DR event on a planet, the protected systems can be started on either the desired replication sites for greater overall VM availability.

RPO	RTO
>1 minute	> 5 minutes

Remote DR Solution

The latency of the advance laser communications system is 10ms and it has a speed of 1GB/s. This connections allows the use of the Nutanix remote replication. Nutanix remote replication is an asynchronies replication. This remote replication has been optimized for WAN connections. 70% of the data is deduped and compressed before it is set to the remote location. Remote replication will be used between Mars and the Moon for DR. In the event of a remote DR event,

the protected systems can be started on either the desired replication sites for greater overall VM availability.

RPO	RTO
>15 minute	> 10 minutes

If this advanced laser connection would happen to become saturated, we will have to redesign the remote DR solution.

DR procedures

Setup

Within Prism a protect domain of Mars will be created.

The following machines will be selected to be replicated:

- Files Server
- All Exchange Systems
- CentOS – Docker
- MariaDB
- Windows NT Systems
- Webservers

Application consistence will need to be enabled for the following systems:

- Windows NT Systems
- All Exchange Servers

Local replication will be every 5 minutes. Remote replication will be every 15 minutes. The local and remote retention policy will be set for 30 days.

Failure

During failover, local or remote, the vSphere administrator has the flexibility to failover the entire protection domain or any number of VMs. This is all done through the Nutanix Prism interface.

Future Architecture

The Mars and Moon colony have no existing matrix to evaluate or use for future planning. vSphere administrators will review the vCenter performance graphs for both real-time monitoring and basic historical analysis/trending.

vCenter Operations Management Suite has been procured and will be installed and configured within the next 90 days to provide enhanced monitoring and future planning

Revision History

Date	Revision Number	Author	Comment
16-Jul	1	J. Brown	Started the Design
18-Jul	1.1	J. Brown	Finalized draft
19-Jul	1.2	J. Brown	Design comments and grammar issues pointed out by Chris Chua