CHALLENGE 1 – LET’S START SLOW

(A NEW DESIGN ON MARS)

BY JAMES BROWN (@JBCOMPVM)

Contents

Executive Summary

We are now settled on Mars, and ready to build a more permanent infrastructure. Keep in mind that power, cooling, and space are extremely expensive resources on Mars. In order to save space, we have decided not to use a traditional Fiber Channel infrastructure, meaning there will be no dedicated Fiber Channel Switches. We do however have plenty of 10G Ethernet switches, with some 40G Ethernet switches. We have three data centers on the planet, in order to provide high availability for our most critical applications. Our most critical system is our Environmental system, which is responsible for production of water and oxygen, as well as reclamation of waste. Should the environmental systems fail, the pods we live in work in can be sustained for only 20 minutes with the existing oxygen in the pod. We rely on this environmental system to control these resources, as well as to warn us when mechanical components throughout the system are failing or have failed. Our second most critical system is the system which controls our greenhouses. Any failure in this system will likely lead to major issues with our food supply. While we have the ability to communicate via radio if needed, many of the residents on Mars are used to e-mail and collaborative applications and prefer to use them when possible, since it makes them feel more at home. Your infrastructure should also be able to support the deployment of an unknown business critical application in the future. Please design an infrastructure you think will meet these requirements. You will not have the opportunity to defend your design in front of the judges, as their expertise is needed to keep the existing systems running until a new architecture is decided upon. Your design should speak for itself.

Background

The years is 2015. President Donald Trump has just taken office. The Mars colonies are starting to add numbers to a small but strong human race. President Trump has always look out for the needs of the colony.

Last year if you can remember that the human race narrowly escape for the zombie apocalypse on Earth. With the help from virtual geeks, they were able to establish a colony on Mars.

Overview

The virtualization solution primarily supports the two main support systems. Secondary systems include collaboration system such as email and instant messaging. These systems are hosted within one of these three data centers at any given time with a raised floor and a conservative HVAC capacity. The datacenter space also includes redundant power and 10GB network connections to every rack. All three data centers are connected with a 40GB fiber ring connection.

After the support system and secondary systems are online, a replica of the support systems will be sent to each data center. Similar to the Microsoft and Amazon datacenters, your data could be located in one of these data centers at any given time.

The system is designed to be able to scale both upwards as demand increases and down as demand decreases, for instance if unknown business applications and/or collaboration software is required for population growth.

Requirements

1. The solution cannot use fiber channel
2. Each colony must be capable of working fully independently.
3. A minimum of three additional datacenters are planned.
4. Environmental systems must meet a 98.6111% 24x7 service-level availability.
    a. Maximum downtime of 20 minutes a day
5. Greenhouse systems need to meet the same service-level availability.
6. The design must be highly reliable and easily deployable.
7. Email and collaboration software need to be available.

Constraints

1. Power, cooling, and space are very expensive resource. They needed to be used sparingly.
2. The design must incorporate the use of reliable, serviceable technology that can degrade gracefully over time.

Assumptions

1. Use more than one data center for redundancy.
2. Data center connections are 40 GB with QinQ enabled.
3. Appropriate licensing for all vendor products (VMware, Microsoft, Red Hat, etc.).
4. vSphere administrators have and can maintain the skillsets required to implement and maintain the solution.

Risks

1. There is an existing solution but there were no anticipated resource requirements.
2. A lack of appropriate personnel may jeopardize the ability to maintain and improve the solution.
3. All specified hardware and software can be acquired and will work on Mars. If this hardware is no longer available, a new solution will need to be designed. (Unless Nutanix will allow their community edition to run in production without limitations)
4. Unknown business application to be announced at a later date

Hypervisor Design

ESXi v5.5 will be installed on each Nutanix node via their foundation installation software. The latest version is ESXi is 6.0. This version has only been publicly available for seven months. With the critical nature of this deployment, the decision was made to use ESXi 5.5 for stability. The hosts will have 2 physical CPUs, 256 GB of RAM, two 480GB SDD Drivers, and four 2TB SATA drives. Management, VM traffic, and vMotion will be carried on redundant 10-Gigabit Ethernet interfaces. Storage is internal so it does not require any networking protocols. 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 v5.5 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 and there is no migration path from VCSA to Windows, the Windows Server version will be used.

A Simple Install – all components on a single VM – provides scalability and maintains low complexity. If VM guest numbers exceed the ability of a Simple Install, the components can be separated by 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. See Security Architecture for more details.

vSphere Update Manager (VUM) will be used for both VMware(vSphere, vCenter, and vShield) 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.

Server Hardware

Nutanix NX-3000 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-3360-G4 server has been chosen. This model has 2 10-Gigabit Ethernet interfaces for all. Each server has 2 E5-2660v3 (20 cores) CPUs for a total of 2 physical cores, and 256 GB RAM, two 480GB SDD Drivers, and four 2TB SATA drives. The system will be configured with 3 nodes. 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-3360-G4 base chassis w/o CPU/DIMM/HDD

Hardware	Quantity
2660 V3 Processors	6
16GB Meory	48
2 TB 2.5 HDD	12
480 GB SSD	6
10GBE Daul SFP+ Network adapters	3
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.

This system will include a firewall (see Security Architecture) to connect to the greater Mars network.

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. All workloads are supported on this platform, which will be used throughout the design.

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 after 30 days.

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	2
vCenter	2	16	80	100	1

vSphere Systems

Nutanix has built in native replication. 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. 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 DR event, the protected systems can be started on either the desired replication sites for greater overall VM availability.

This table shows all vSphere systems and their initial resource allocations and quantities.

Server	**vCPU**	**RAM (GB)**	**OS Disk (GB)**	**Data Disk (GB)**	**Quantity**
vShield Manager	2	8	60	0	1
vShield Edge	2	1	0.5	0	1
vShield EndPoint	1	0.5	512	0	3

Email and Instant Messaging Systems

The email and Instant messaging system has been developed by Microsoft. These systems are designed for high scalability to support larger designs. We will base the server needs on the vendor best practices.

These systems are not mission critical, but they are highly utilized by the Mars colony to keep in touch with the different areas.

Service	vCPUs	RAM (GB)	OS Disk (GB)	Data Disk (GB)	Quantity
Exchange (400 mailboxes or less)	4	16	80	500	1
Skype	2	32	80	100	1

The cumulative totals of vCPU, RAM, and disk allocations and VM count for the initial turn up are:

vCPUs	RAM (GB)	DISK (GB)	Quantity
14	86.5	2696.5	10

VMware Datacenter Design

The Mars vCenter server will define one datacenter for ThinkBig City. A single Nutanix node of ESXi hosts will be provisioned immediately.

To meet the 98.61% 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 7% of cluster resources to provide for 1 host failure (1/6 * 100) and will be revisited every 30 days as manufacturing capacity increases 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
VM Override VM Restart Policy - High	Management - vCenter, DCs

Security Architecture

The security of the support system is vital. Any security compromises, accidental or purposeful, risk the entire human race. Defense in depth (or layers) will mitigate nearly all security gaps.

Security is an ongoing concern and the steps outlined here define an initial security policy only. The architecture, policy, and implementation will immediately and continually evolve to meet the demands of the system and its users. Therefore this document is NOT be considered authoritative for the production system.

All VMs will use the OS's included host firewall (Windows Firewall or iptables) to manage inbound traffic. The template VMs will include a very conservative policy (inbound ssh and established connections only). Outbound VM traffic will not be managed with host firewalls unless an application specifically calls for it (currently none do).

Inter-VLAN traffic will be managed and protected with VMware vCloud Networking and Security 5.5.4.1. VMware NSX is being considered for a future upgrade. vShield Manager, vShield Edge and vShield Endpoint will provide central management, protection between networking segments, and in-VM protection from viruses and malware. vShield App is not required due to Guest OS firewalls and vShield Data Security is not a relevant concern to the isolated VMware cloud.

The system's edge, between the support network and the Mars colonies LAN/WAN, will be protected with a Fortigate-300D in routed, single-VDOM mode. The FortiGate license allows for multi-VDOM mode (multiple logical firewall instances per interface/VLAN) if necessary in the future. Initial policy will allow unrestricted outbound common services and more restricted inbound services according to the table below.

	System to LAN/WAN		
SRC	**DST**	**Service**	**Action**
Internet Network	External Network	HTTP, HTTPS, SSH, SMB, DNS	Permit
	LAN/WAN to Systems		
SRC	**DST**	**Service**	**Action**
vSphere Admins	vCenter	9443/TCP	Permit

vCenter SSO will use the Active Directory as the primary namespace. All vSphere administrators and the chosen colonist will be in the Administrator group. The administrator@vsphere.local account information is made known to the vSphere team lead and the colonist in order to reduce the potential for unaudited actions that could cause harm.

The ESXi hosts will have lockdown mode enabled. The local root account password will be shared with all hosts (applied via host profile) and will be made known to the vSphere team lead and colonist in case local DCUI/shell access is required.

Monitoring and Capacity Planning

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

Day one monitoring will consist of vCenter's default alerts. vCenter alerts will be configured to send email notification to the vSphere administrators as well as the support plant's on-shift manager and a designated liaison. Notification to three parties will ensure that alert responses are timely. After workloads are in place for fourteen days and the vSphere administrators have a baseline utilization to reference, the alerts will be customized. A balance between too much alerting, which breeds complacency and ignorance, and too little alerting, which may result in outages or impacts occurring, must be maintained.

The lack of existing baselines prevents more accurate forecasting of monitoring and capacity planning needs. Ongoing management of the infrastructure and other efforts may preclude vSphere administrators having excess time to manage a monitoring system, especially if the workloads are stable and within thresholds. However, vCenter Operations Management Suite has been procured and, time and effort permitting, will be installed and configured within the next 90 days to provide enhanced monitoring and capacity planning