Virtual Design Master

SEASON 5 – CHALLENGE 1

Dale Handley
@DALEMHANDLEY

04/07/2017

Table of Contents

1 Executive Summary 2
1.1 Mission objective 2
1.2 Requirements 2
1.3 Constraints 2
1.4 Assumptions 2
1.5 Risks 2
2 Conceptual Design 3
2.1 Datacenter Locations 3
2.1.1 DC1: Goonhilly Satellite Earth Station 3
2.1.2 DC2: Cheyenne Mountain Complex 3
2.1.3 DC3: Pine Gap 3
3 Physical Design 3
3.1 Servers 4
3.2 Storage 5
3.3 Networking 5
3.3.1 Physical Networking 5
3.3.2 Virtual Networking 5
3.4 Management Workloads 6
3.4.1 NTP 6
3.4.2 Directory Services (incl. DHCP, DNS & Certificate Services) 6
3.4.2.1 Active Directory 6
3.4.2.2 DHCP 6
3.4.2.3 DNS 7
3.4.2.4 Certificate Services 7
3.4.3 vCenter Server (incl. ESXi) 7
3.4.4 Container Services 8
3.4.4.1 Container Registry 8
3.4.4.2 Container Management 8
3.4.4.3 Container Engine 8
3.5 Monitoring 9
3.5.1 vRealize Operations 9
3.5.2 vRealize Log Insight 9
3.6 Backups 9
3.7 Security 10
3.7.1 Physical Security 10

3.7.2 Management Access 10
3.7.3 Anti-Virus / Anti-Malware 10
3.7.4 Updates / Patching 10

1 Executive Summary

1.1 Mission objective

It is assumed that the zombie population has now finally died out and mankind is ready to retake the Earth. To do this the Earth will need to be terraformed by an army of robots. Why does it need to be terraformed you ask – presumably as a way of informing the participants of a technology that may come in handy in a later challenge (https://www.terraform.io/)

1.2 Requirements

	Description
RE01	Choose three locations for our Datacenters
RE02	Choose appropriate hardware and software
RE03	Create a resilient three site architecture
RE04	Resiliency should be deployed in as many layers as possible
RE05	Uptime and performance of HumanityLink 2.0 is paramount
RE06	Terraforming robots should be available 24/7 (except during scheduled maintenance)
RE07	Scale workloads in all directions & deal with unknown workloads

1.3 Constraints

	Description
C01	Initial deployment on Earth

1.4 Assumptions

	Description
A01	There is sufficient budget to procure and build out the design
A02	The Datacenters chosen are still in full working order with adequate power and cooling
A03	A redundant, high-speed, low latency link is already in place between the three sites
A04	Earth-like environmental conditions are available
A05	Hardware identified is available to us
A06	Licenses have been procured
A07	HumanityLink 2.0 is responsible for the scaling of the application
A08	Robots are running on an OS that can be effectively secured
A09	VMs have been sized using hardware requirements supplied by the vendor where available
A10	Humanity has already been through a lot – the infrastructure should be easy to manage

1.5 Risks

	Description	Risk Mitigation
RI01	The Zombie population has not died out like we have assumed	Datacenter locations have been chosen that are secure and/or remote

RI02	One or more “bad actors” – humans with nefarious intent	The design incorporates a level of resilience and separation that reduces the chances of unauthorised access
RI03	The size and performance characteristics of the workloads are unknown	Servers have been sized to provide a level of headroom to deal with scale-up or scale-out of the HumanityLink 2.0 application
RI04	Viruses and/or malware	Antivirus agents will be deployed to the VMs NSX Distributed Firewall will be used to limit unnecessary server to server communications limiting the potential for viruses/malware to spread
RI05	Failure of the Certificate Authority will cause communication issues	Subordinate CAs will be deployed at each site to ensure that a failure of one will not shut down all of the sites

2 Conceptual Design

2.1 Datacenter Locations

2.1.1 DC1: Goonhilly Satellite Earth Station

Located in Goonhilly Downs in Cornwall, Goonhilly Earth Satellite Station is connected directly to a key submarine cable landing point which houses TAT-14 (a transatlantic fibre optic cable system designed to operate at up to 9.38 Tbit/s.) It is also powered directly by the National Grid with battery and diesel generator backups. This makes this site ideally suited to be the main base of operations.

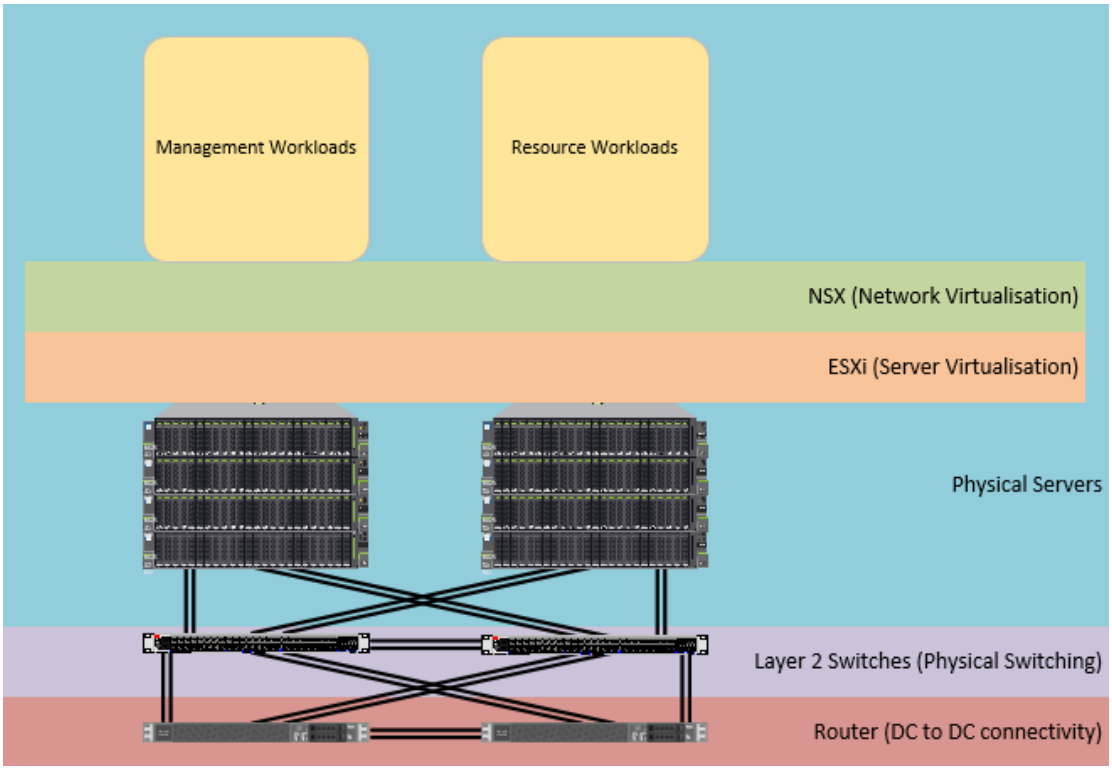
2.1.2 DC2: Cheyenne Mountain Complex

The Cheyenne Mountain Complex will be the location for the second Datacenter because… well… why wouldn’t you? It’s an underground Nuclear bunker with Air Defence & Space Defence (VDM Season 6 anyone?) systems and a blast door capable of withstanding a 30 megaton explosion.

2.1.3 DC3: Pine Gap

The third Datacenter will be located at Pine Gap – an Earth Station located in Australia. It has excellent communication links and its remote location will serve us well in the event of another Zombie Apocalypse.

3 Physical Design



3.1 Servers

Four management servers and four resource servers will be deployed at each site (using the same specification for simplicity.) The rack that the servers are in will be fed by two independent power feeds. The hardware specification is as follows:

Component	Description	Quantity
System	Fujitsu PY RX2540 M2 16x 2,5" expandable (AF-6 Series)	24
CPU	Intel Xeon E5-2650v4 12C/24T 2.20 GHz	2
Memory	16GB (1x16GB) 2Rx4 DDR4-2400 R ECC	16
Network Interface	PLAN EM 2x10Gb T OCl14000-LOM interface (SFP+) Eth Ctrl 2x10Gbit PCIe x8 D2755 SFP+ lp	1 1
Controller	Fujitsu PSAS CP400i SAS	2
Caching Tier	SSD SAS 12G 200GB Main 2.5'' H-P EP SAS	2
Capacity Tier	SSD SAS 12G 800GB Main 2.5'' H-P EP SAS	10
Boot Device	Embedded UFM 8 GB Device	1

This provides a total of 24 Cores (48 Threads with Hyperthreading enabled) and 256GB of RAM per server (storage details below)

Justification:

Fujitsu are a Tier 1 server manufacturer with a number of servers certified by the VMware vSAN Ready Node program. vSAN requires a minimum of 3 nodes per cluster – clusters have been sized at 4 nodes to ensure that VMs can still be deployed if a host is unavailable due to a fault or scheduled maintenance. The use of Blade servers was investigated due to the speed at which extra capacity could be added if required but for all of the layers of resiliency that are included they have one very significant single point of failure – the backplane.

3.2 Storage

Storage will be provided by VMware vSAN 6.6 in an All-Flash setup. The storage setup will be (for each cluster):

Raw Storage Requirement (without FTT)	12083 GB
Raw Storage Requirement (with FTT using Erasure Coding)	16070 GB
Space reserved for Slack	30%
Raw Unformatted Storage Capacity	20891 GB
Number of Hosts per site	4
Cache Required per Host	522 GB
Capacity Required per Host	5222 GB

Justification:

Using VMware vSAN simplifies the design, the level of administration overhead, provides a consistent performance level and reduces the amount of power and cooling required versus a traditional Storage Array. The sizing design was done using the VMs that have been identified for the Management Cluster, the same setup will be used for the Resource Cluster for simplicity. [RE04]

3.3 Networking

3.3.1 Physical Networking

It is assumed that a redundant connection between the three Datacenters is in place [A03] and provides sufficient connectivity for a pair of Layer 2 switches.

Two Brocade 6740 switches have been selected to provide the physical networking connectivity. The switches each have 48 x 10Gbe SFP+ ports and 4 x 40Gbe QSFP+ ports. 2 x 40Gbe ports will connect the switches whilst the other 2 x 40Gbe connections will provide uplink connectivity (one to each router.)

VLAN	Portgroup	NIC
100	Management	vmnic0, vmnic2
101	vCenter HA	vmnic0, vmnic2
200	vMotion	vmnic0, vmnic2
300	vSAN	vmnic1, vmnic3
400	Virtual Machines	vmnic1, vmnic3

Justification:

The switches have been chosen due to their large port count and ease of use (for example connecting the switches together using 2 x 40Gbe ports will automatically form a resilient ISL connection.)

3.3.2 Virtual Networking

Virtual networking will be provided through the use of VMware NSX

VXLAN	Description
401	Web Tier
402	Application Tier

403	Database Tier
404	Communication with Robots

Justification:

NSX allows for rapid deployment of virtual network services such as load balancers & firewalls, orchestration of new networks and vastly increases the security of the platform through the use of the Distributed Firewall.

3.4 Management Workloads

3.4.1 NTP

A pair of CentOS 7 servers will be configured in DC1 to provide a time source to all of the infrastructure incl. Switches, Management workloads, Resource workloads and Robots.

Name	Description	vCPU	vRAM (GB)	vDisk (GB)
DC1NTP1	DC1 NTP Server 1	2	4	20
DC1NTP2	DC1 NTP Server 2	2	4	20

Justification:

All of the infrastructure components rely heavily on a single, accurate time source. Configuring a pair of NTP servers will provide a simple to support, highly available time service.

3.4.2 Directory Services (incl. DHCP, DNS & Certificate Services)

3.4.2.1 Active Directory

Two Active Directory Domain Controllers will be deployed at each site running on Windows 2012 R2. The servers will all be part of the same domain.

Name	Description	vCPU	vRAM (GB)	vDisk (GB)
DCxAD1	DCx Active Directory 1	2	4	70
DCxAD2	DCx Active Directory 2	2	4	70

Justification:

Deploying a pair of DCs at each site will ensure a resilient authentication service. These servers will also supply a platform for resilient IP address, name resolution and certificate services. [RE04]

3.4.2.2 DHCP

Management workloads will be deployed using Static IP addresses – this ensures availability in the case of a complete failure of DHCP services. Resource workloads due to their potentially transient and scale in/out nature will receive their IP address using DHCP.

The two Active Directory servers will be responsible for local IP addressing through the use of the Microsoft DHCP Server service. The two servers will be configured using the DHCP Failover service in a Hot Standby setup. The Primary DHCP server in each site will be responsible for responding to DHCP requests, allocating IP Addresses and notifying the Secondary DHCP server of lease status. [RE04]

Justification:

In the above setup failure of the Primary DHCP server would see the Secondary DHCP server take over the role (once the Primary DHCP server is back online the role will fail-back.) In a split-brain scenario both servers may respond to DHCP requests but they will use a different portion of the free IP address pool to prevent IP conflicts. [RE04]

3.4.2.3 DNS

The DNS server role will be installed on each of the Domain Controllers

Justification:

A highly available name resolution service (forwards and backwards) is key to the operation of most (if not all) of the management and resource workloads.) [RE04]

3.4.2.4 Certificate Services

Digital certificates will be installed on the following workloads:

- Communication between Management Workloads (vCenter, vROPS, vRLI etc.)
- Communication between the tiers of the HumanityLink 2.0 application
- Communication between HumanityLink 2.0 and the terraforming robots

A Root CA will be installed in Datacenter 1 with Subordinate CAs installed into the three Datacenters.

Justification:

Replacing the self-signed certificates with CA-signed ones increases the overall security of the solution and ensures that only authorised endpoints are communicating with each other. The use of multiple subordinate DCs prevents a failure of all sites in the event of a failure or unavailability of a single CA. [RE04]

3.4.3 vCenter Server (incl. ESXi)

The vCenter Server Appliance (6.5) will be used to manage the ESXi hosts. It will consist of a HA pair of vCSA nodes and a vCSA witness (to prevent a split-brain scenario) deployed using the Small appliance profile. vCenter server will be deployed using the embedded Platform Services Controller as there is no need to provide an external SSO service.

Name	Description	vCPU	vRAM (GB)	vDisk (GB)
DCxVCSA1	DCx vCenter Server 1	4	16	250
DCxVCSA2	DCx vCenter Server 1	4	16	250
DCxVCSAW	DCx vCenter Server Witness	2	10	-

Due to the small size of the environment ESXi servers will be built manually using an ISO attached to the Fujitsu iRMC. Host Profiles will be used to set the basic configuration parameters (NTP, DNS, Password policy, firewall policy etc.) and vSphere Update Manager (embedded) will be used to ensure that patches are deployed in a timely and consistent manner.

Justification:

vCSA 6.5 will be used so that we can take advantage of the new VCSA only features (High Availability, simple backup/restore/migration etc.) An external PSC service will not be deployed as there is no requirement for a single PSC domain across the three sites or to provide a PSC service to other VMware components (which would also need a complex load balancing setup.)

Even though the environment only calls for a Tiny appliance (10 hosts, 100 VMs) a Small appliance will be used to satisfy the requirements of vCenter HA (to provide an RTO of 5 minutes.)

3.4.4 Container Services

3.4.4.1 Container Registry

VMware Harbor will be used as it is an enterprise-class Docker compatible registry. Two instances will be deployed at each site and load balanced using an NSX load balancer configured for HA. Changes will be made on the primary Harbor instance in DC1 which will be configured to replicate to the instances in DC2 and DC3.

Name	Description	vCPU	vRAM (GB)	vDisk (GB)
DCxHADMIRAL1	DCx Harbor/Admiral 1	2	4	500
DCxHADMIRAL2	DCx Harbor/Admiral 2	2	4	500

Justification:

VMware Harbor provides features such as AD support, Role Based Access Control, Auditing etc. Using two instances of Harbor will allow running a resilient, synchronised Docker registry. [RE04, RE07]

3.4.4.2 Container Management

VMware Admiral will be used as the Container Management Portal. Two VMware Admiral hosts will be deployed at each site to provision and manage the container lifecycle.

Justification:

VMware Admiral can be deployed in an Active/Active Cluster where any of the Admiral nodes can handle container provisioning requests using a shared state which is replicated between the hosts. [RE04, RE07]

3.4.4.3 Container Engine

vSphere Integrated Containers has been chosen as the Docker compatible engine. A single Virtual Container Host will be deployed at each site which will act as the Docker endpoint.

Justification:

VIC has been chosen as it provides significant benefits over other container engines such as:

Availability – Running multiple containers on top of a monolithic VM would cause all containers to be unavailable if the VM fails.

Manageability – Each container is running as an individual VM which provides visibility of each individual workload and allows us to take advantage of vSphere features such as HA, vMotion and DRS.

Performance – Each container has the resources of the cluster available to it and are not constrained by the capacity allocated to a single VM.

Security – Running each container in a separate VM allows us to exploit the NSX Distributed Firewall to secure each workload individually.

vSphere HA will be used to protect the VIC engine. In the event of a failure of the host that the VIC engine is running on existing containers will continue to run but actions against new or existing containers will fail until vSPhere HA restarts the VCH. In the event of a total failure of the VCH a new one can be provisioned and registered with Admiral. [RE07]

3.5 Monitoring

3.5.1 vRealize Operations

vRealize Operations 6.6 will be deployed in a two node Cluster configuration (a Master and a Replica.)

Name	Description	vCPU	vRAM (GB)	vDisk (GB)
DC1VROPS1	DC1 vRealize Operaions 1	8	32	250
DC1VROPS2	DC1 vRealize Operaions 2	8	32	250

Endpoint Operations agents will be installed on the Management Workloads and the Robots, with dashboards created to monitor them at a Service level.

Justification:

To monitor performance and ensure that the environment is running efficiently vROPS will monitor all workloads on the platform. Using a two node cluster ensures that vROPS will still be available in the event of the failure of one server and there will be no loss of data. The vROPS sizing guide has been used to determine the appliance sizes which will cater for 24 ESXi Hosts, 200 VMs and 1200 Endpoint Operations agents with room for growth. [RE04]

3.5.2 vRealize Log Insight

vRealize Log Insight 4.5 will be deployed in a three node configuration using the Medium size appliance (the minimum required for HA) and integrated with vRealize Operations.

Name	Description	vCPU	vRAM (GB)	vDisk (GB)
DC1VRLI1	DC1 vRealize Log Insight 1	8	16	1500
DC1VRLI2	DC1 vRealize Log Insight 2	8	16	1500
DC1VRLI3	DC1 vRealize Log Insight 3	8	16	1500

Unstructured data will be collected from each device - Syslog based (Routers, Switches and Servers) or Agent based (Windows and Linux Management/Resource workloads.) Content packs will be installed for: Windows and Linux OS, Active Directory, vSphere, vSAN and NSX.

Justification:

Log Insight will provide visibility into the actual workloads running in the environment through syslog/agent based communications. This data will be analysed by the various content packs that are installed and surface alerts through the integration with vROPS providing a “Single Glass of Pain” for the administration team

3.6 Backups

Veeam will be installed on a Windows 2012 R2 server of the following specification:

Name	Description	vCPU	vRAM (GB)	vDisk (GB)
DCxVEEAM1	DCx Veeam Backup Server 1	4	8	5000

The Backups taken by Veeam will be Reverse Incrementals and the retention policy will be set to 14 days. There will be two jobs setup – Management Workloads and Resource Workloads. Resource Workloads will be backed up first starting at 8pm.

Justification:

Veeam has been chosen as it was designed from the ground-up to backup Virtual environments. It is simple to use, fast to backup, has excellent compression technology and provides a wealth of features. Reverse incremental backups have been chosen as restores in most instances will be done from the previous nights backup. Sizing has been chosen based on the minimum sizing (2 vCPU and 4GB RAM + 500MB for each concurrent backup job. Storage has been allocated based on the total storage size of VMs being a total of around 7TB, 5 TB has been allocated to the VM due to the space saving of deduplication and compression.)

3.7 Security

3.7.1 Physical Security

The sites identified have been specially selected because they will already have in place a number of physical security defences which at a minimum would include security fences and CCTV cameras. Additional security features will be retrofitted to the building to ensure only authorised personnel can get on site. These will include identity passes and fingerprint scanners.

3.7.2 Management Access

A separate Management LAN has been provisioned which should ensure that only authorised administrators can connect to systems such as the network switches and the ESXi hosts.

3.7.3 Anti-Virus / Anti-Malware

McAfee Antivirus will be installed onto all Windows and Linux endpoints.

3.7.4 Updates / Patching

Microsoft WSUS and Redhat Spacewalk will be deployed into the environment to provide patching services for Windows and Linux workloads respectively.