Architecture Design Document
VirtualDesignMaster - Season 2

Challenge 2
Lunar Base Communications Infrastructure

Prepared by: Daemon Behr

Date: 2014–07-21

Revision History

Date	Rev	Author	Comments	Reviewers
2014-07-14	R1	Rob Nelson	Challenge 1 Submission	VDM Judges
2014-07-21	R2	Daemon Behr	Challenge 2 Submission	VDM Judges

Design Subject Matter Experts

The following people provided key input into this design.

Name	Twitter Address	Role/Comments
Daemon Behr	@VMUG_Vancouver	Infrastructure Architect
Rob Nelson	@Rnelson0	Security Engineer / Systems Administrator

Contents

1. Purpose and Overview 6
1.1 Executive Summary 6
1.2 Summary Analysis 6
1.3 Design Interpretation 6
1.4 Intended Audience 6
1.5 Requirements 6
1.5.1 Fault Tolerance 7
1.5.2 Delay 7
1.5.3 Mobility 7
1.5.4 Connectivity 8
1.5.5 Energy 8
1.5.6 Environmental Factors 8
1.6 Constraints 9
1.7 Risks 9
1.8 Assumptions 10
2. Architecture Design 11
2.1 Design Decisions 11
2.2 Conceptual Design 13
2.2.1 IPN Conceptual Design 13
2.2.2 Lunar base conceptual design 18
2.2.3 Communications Infrastructure conceptual design 19
2.3 Logical Design 20
2.3.1 Logical design for communications infrastructure in A-BLD-7 20
2.3.2 Logical design for communications infrastructure in A-BLD-4 21
2.3.3 Logical design for infrastructure power system 21
2.3.4 Logical design for virtual infrastructure 23
VM Design 24
Management services 25
Windows Domain and vCenter 25
vSphere Systems and Appliances 26
Communications System 27
Web-Commander Portal 28
Puppet System 29
VMware Datacenter Design 30
Security Architecture 32

Network Design 33
2.3 Physical Design 34

1. Purpose and Overview

1.1 Executive Summary

After much planning, design and a successful infrastructure deployment, the Depot at Cape Canaveral is almost online, and the depots in the Netherlands, Australia, and New Zealand will be coming online soon as well.

The next obstacle is the infrastructure for the Anacreon Lunar base. The Moon will be used as stopping point on the way to Mars, in addition to serving as the human race's new home until the colony on Mars is finished.

Due to the way the base was constructed, there are serious power, cooling, and space limitations. This requires scaling the design down to fit into half of a datacenter rack, or 21U. In addition, there is only an IPV6 network infrastructure on the base.

The same vendors will be used as in the previous challenge, but different product lines will be used.

Configuration maximums must be provided for the new designs (for example, maximum hosts, VMs, storage capacity), and what the failure tolerance is.

1.2 Summary Analysis

The purpose of this design is to implement an extremely robust and fault resistant infrastructure that is able to operate on minimal hardware and many constraints. The design has to make use of very limited space; power and technical resources but provide the highest capability for compute and storage.

1.3 Design Interpretation

Due to the constraints provided by the main factors, the design will be a highly converged infrastructure, fault-tolerance, and have graceful degradation. In addition the lack of replacement parts, or additional resources need to factor into the design blueprint.

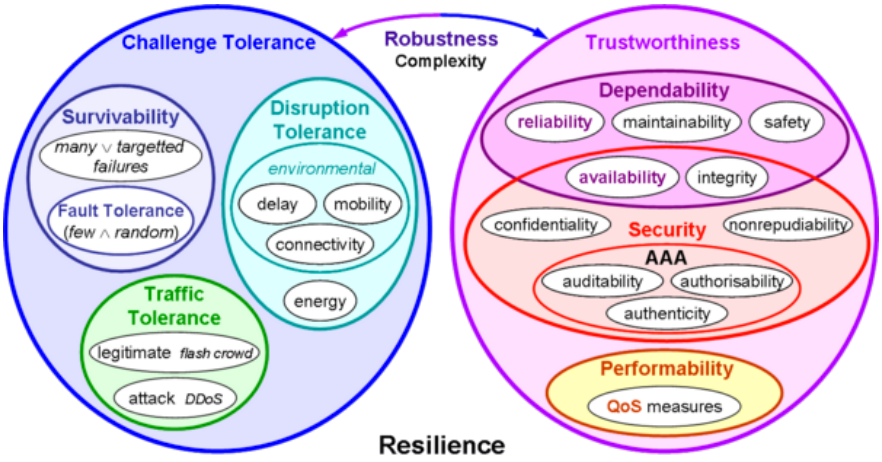
1.4 Intended Audience

This document is meant for the key stakeholders in the project as well as terrestrial and lunar technical staff.

1.5 Requirements

Below are the requirements as defined by the scenario document as well as additional communication with judges and creators in clarification emails.

The requirements defined will be mapped to the Challenge Tolerance discipline of a resilient system. Since the main focus of the project is survivability and disruption tolerance, we will focus on those areas for the requirement definitions.



1.5.1 Fault Tolerance

A fault-tolerant design enables a system to continue its intended operation, possibly at a reduced level, rather than failing completely, when some part of the system fails

R001	**All infrastructure building blocks must be able to continue operation after failure.**

1.5.2 Delay

Delay can be defined by latency of undetermined, or indefinite periods. This may be seconds to minutes, to days, or years.

R002	**The infrastructure must be able to store and forward communications in the event of extended communication loss**

1.5.3 Mobility

Mobile computing involves mobile communication, mobile hardware, and mobile software. In this context it refers to the ability to ability of communications between terrestrial and lunar bodies during orbital movement.

R003	**The Inter Planetary Network (IPN) communication must account for lunar orbital (including eccentricity) and weather systems that disrupt "line of sight" communications.**

1.5.4 Connectivity

In this context, connectivity refers to the transmission links connect the nodes together. The nodes use circuit switching, message switching or packet switching to pass the signal through the correct links and nodes to reach the correct destination terminal.

R004	**The Inter Planetary Network (IPN) communication must be able to communicate with multiple regions on the earth where the launch facilities are located.**

1.5.5 Energy

Power availability is not infinite and must be taken into consideration when designing the infrastructure. Consumption will vary depending on workload and "active" hardware.

R005	**Predictable output of power systems is required.**
R006	**Redundant power sources and backup power are required.**
R007	**Equipment must be able to power down when not active to reduce usage.**

1.5.6 Environmental Factors

Due to the harsh environmental conditions of not having an atmosphere, additional considerations are required.

R008	**Opto-isolation is required from antennae systems / transceivers to routing equipment to prevent surges from adverse solar weather patterns.**
R009	**Climate control is required in communications infrastructure pod to ensure equipment stays within operating temperature limits.**

1.6 Constraints

C001	**The infrastructure must fit in 21U of rack space**
	This is the maximum space and cannot be expanded
C002	**There will be no additional replacement parts that will be sent in a timely fashion.**
	No four-hour parts replacement SLA here. Maybe 4 months, or never.
C003	**IP networking on lunar base is IPv6 only**
	All current networking on the base is IPv6 only and is a requirement for the IPN.
C004	**Power potential is limited**
	Power is restricted and requires limited use.
C005	**The Foundation's mandate is to preserve humanity's history**
	This is in the form of the Encyclopedia Galactica, which attempts to capture impressions of the colonists and Earth to send with each ship.
C006	**Bandwidth from launch sites to the lunar base is limited**
	When the LLCD is the preferred route, upstream bandwidth can be maintained reliably at 20Mbps up and a peak of 622Mbps down. Relayed satellite communications are limited to an average of 10Mbps with a greater latency.

1.7 Risks

R001	**The virus may spread to lunar colonists**
	If lunar base staff contract the virus, get space madness, or depression, then no on-site support will be available.
R002	**If total system failure occurs, then further missions may be difficult or impossible.**
	Delays due to communication errors will cost human lives.
R003	**Adequate power may not be available to support all equipment**
	This would impact the ability to operate critical equipment for communication.
R004	**Inclement weather conditions.**

	Terrestrial storms obscuring LOS (Line of Site), or Solar storms affecting ionosphere needs to be accounted for. Debris or meteoric impacts also need to be considered
R005	**Communication issues may prevent adequate bandwidth for data replication.**
	If limited bandwidth is available, then timely data transfer may not be possible.

1.8 Assumptions

A001	**Existing terrestrial infrastructure is capable of minor configuration changes.**
	This is required to support the IPN.
A002	**No additional hardware or personnel will reach the equipment for a very long period of time.**
	Once installed, it will be left alone for months, or years, without any physical interaction.
A003	**Adequate power is available for 21U of rack space.**
	Estimated power is 15A at 110V
A004	**There is a high speed, low latency WAN link between all three launch site locations.**
	This is 1Gbps at <100ms latency
A005	**All three launch sites will have identical infrastructure**
	Compute, network, storage and application stack will be identical.
A006	**Support staff will still be available**
	No virus infections have happened at the launch sites and the world cup is not occurring.
A007	**The Foundation has acquired appropriate licensing**
	This is for all vendor products (VMware, Microsoft, Red Hat, etc.) via government orders.
A008	**vSphere administrators and developers have and can maintain the skillsets required to implement the solution.**
	Runbooks and workflow documents are available in the case of sudden staff shortage due to zombie over-runs and facility reclamation.
A009	**Each colony ship will include one trained technician who can use, maintain, and repair the ship's virtualization platform**
	If this is not possible, then an offline Pluralsight library will be available for training.

2. Architecture Design

2.1 Design Decisions

The architecture is described by a logical design, which is independent of hardware-specific details.

The following design decisions are presented with their justification.

D001	**A management cluster will be added to all launch sites**
	To simplify troubleshooting of management components, they will be contained in a small and manageable 3 host cluster. This also allows scalability of end user compute to achieve its cluster maximums without management overhead.
D002	**Memory on all terrestrial hosts will be increased to 512GB**
	This is for VMs that may have large memory configurations. This allows for 128GB per CPU socket with 4 sockets, increasing the distribution per NUMA node.
D003	**VDPA will be replaced by VSC / SnapManager at the terrestrial launch sites**
	This is to take advantage of native snapshotting features and increase the speed of recovery.
D004	**Reduce storage vendors to only NetApp**
	This is to reduce storage complexity and interoperability. A NetApp FAS8020 will replace the Synology device with an equivalent amount of storage capacity. Replication will be done via native snapmirror capability.
D005	**Change primary storage protocol to NFS instead of FCoE for lunar site**
	NFS will be used for datastores in vSphere. This is to make use of diskless booting and gain the benefits of NFS such as file level snapshots and simplified management.
D006	**Interplanetary communications will route through geostationary satellites.**
	Direct communication with the lunar base from terrestrial launch sites is difficult because of distance and atmospheric interference. By relaying communication through one or more satellites, reliability and redundancy are increased.
D007	**Only Windows vCenter servers will be used.**
	This is because of lack of support for IPv6 in VCSA.
D008	**Replication will not use snapmirror**
	Inter-cluster snapmirror is not supported on IPv6. However, NDMPCOPY and RSYNC are both supported. The one that is more tolerant to latency and retransmissions will be used. This can only be determined after adequate testing.

D010	**Lunar infrastructure will not use any spinning disk**
	Due to the average lifespan of even the most reliable spinning drives, they do not have adequate longevity for a long-term solution. In the best of conditions, 5-10 years is the maximum lifespan. Under load, it lessens even more to 3-5 years.
D011	**Only SLC SSD drives will be used in the lunar infrastructure**
	SLC drives have a greater longevity that MLC drives due to a number of factors including size under-provisioning and wear leveling. Cost is not a constraint in this design.
D012	**Lunar infrastructure will have multiple power sources**
	The power cells will be charged by two means. Solar cells and a radioisotope thermoelectric generator. The power cells will also have dual charge controllers and a dummy load for runaway power surges. There will be two power cell banks, A and B.
D013	**An emergency failover infrastructure will be located at the opposite end of the lunar base**
	This will consist of a single 2U Cisco ISR router with 2 UCS-E 160W blades and local disk. This is for use in the case of a complete failure or loss of the primary lunar infrastructure.
D014	**DPM will be used to reduce power consumption**
	When not in use, the unrequired servers will power down to a standby state to preserve power.
D015	**No automation engine will be used.**
	This decision is due to fact that the environment will not scale or change very often. The components required for vCAC and Puppet put additional load on the system.
D016	**An Expert System will be used for human-machine interaction.**
	This can be used instead of a keyboard and touchpad for controlling the infrastructure via PowerCLI and Python scripts from voice macros. The system can perform self-diagnostics by singing the song “Daisy Bell”.
D017	**No SAN will be used in the lunar site**
	VSAN will be used in order to reduce the amount of space being used just for storage. A converged infrastructure using UCS-B and UCS-E series will be used instead.
D018	**There will be a replication VM that acts as the destination for terrestrial uploads.**
	This VM will run NetApp OnTap Edge and act as an intermediary LUN before it’s replicated to long-term storage.

D019	Certain components that were installed in the terrestrial sites will not be installed in the lunar sites because of the limitations of IPv6.
	The following components are not supported by IPv6: - vCloud Automation Center - NetApp Inter-cluster replication on clustered mode

2.2 Conceptual Design

2.2.1 IPN Conceptual Design

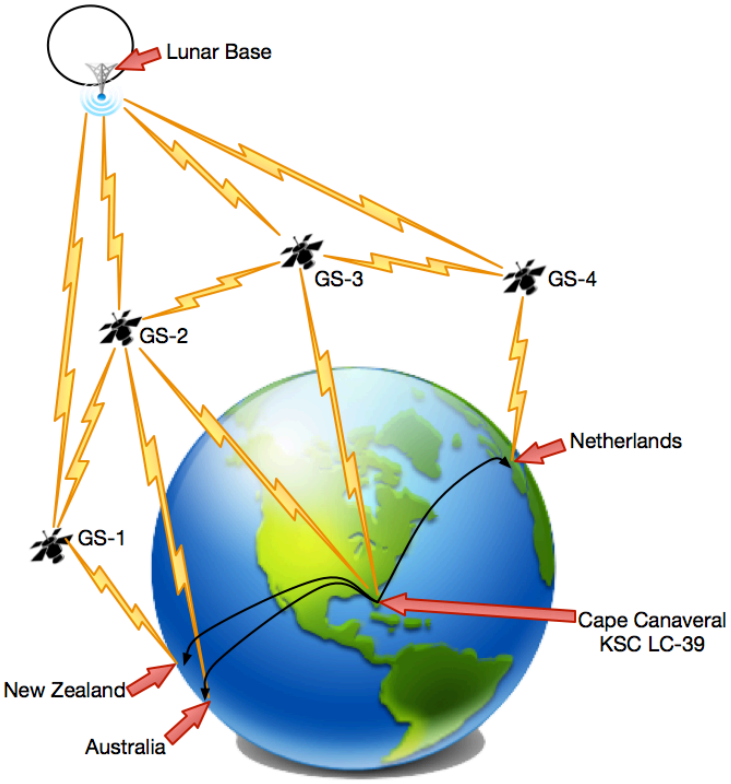


Figure 1. Geostationary satellites for lunar base communication.

Communication between terrestrial and lunar base stations is done via the IPN, which routes through several orbiting satellite layer. The uplink satellite depends on which has the most reliable connection. Routing occurs through long haul terrestrial high-speed links, then to the satellite that is the best path at that point in time.

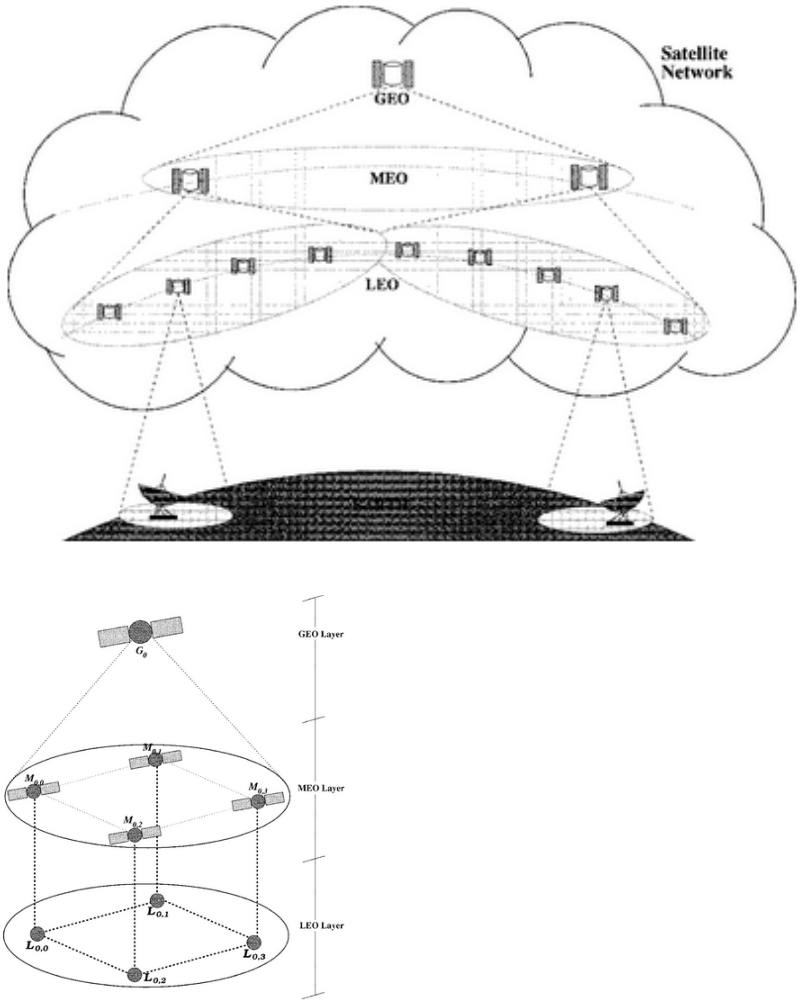


Figure 2. Satellite communication layers

Communication to geostationary satellites occurs in layers, from “low Earth orbit” (LEO) to “medium Earth orbit” (MEO) to “geostationary orbit” (GEO). Once in geostationary orbit, the satellites will have line of site with other GEO satellites or the lunar base.

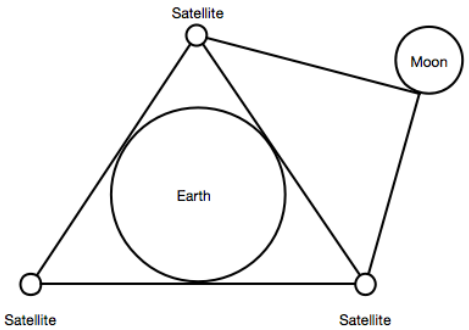


Figure 3. Satellite Line of Sight

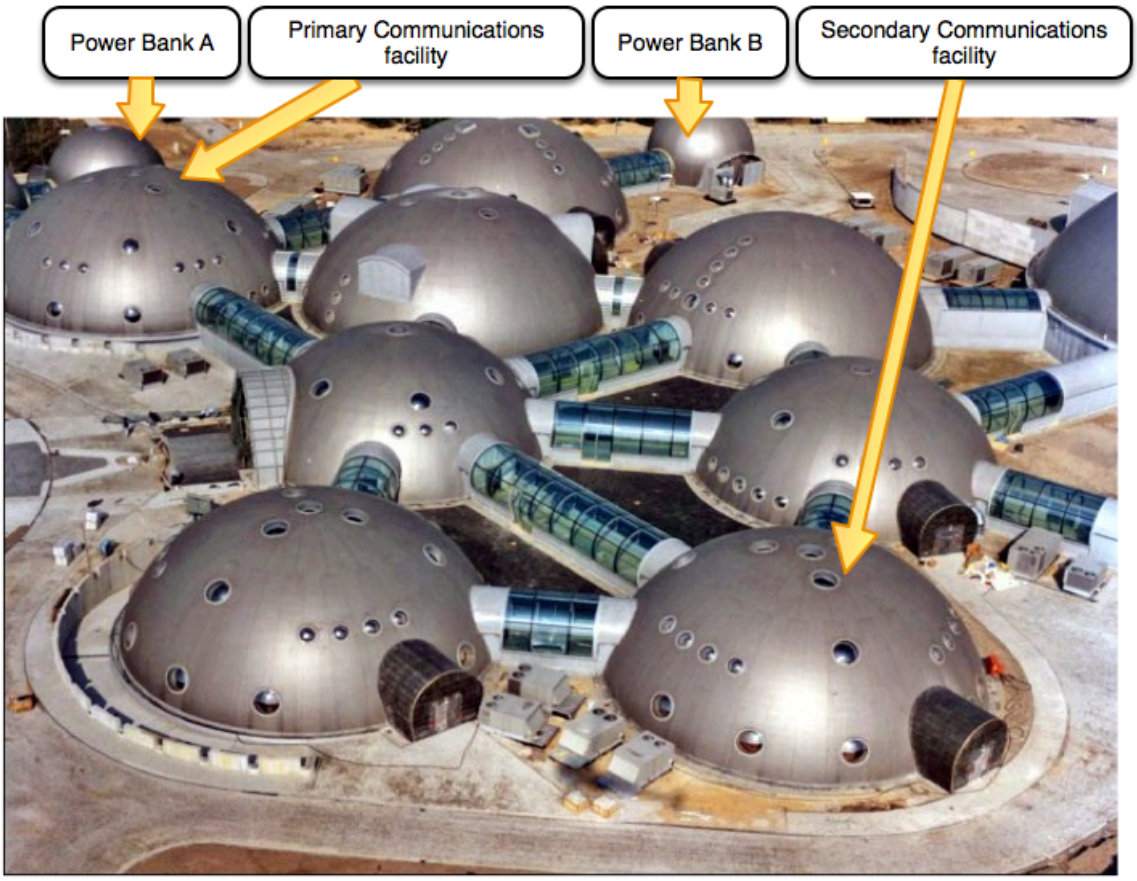


Figure 4. Lunar base communications facilities

IPN communications will occur via RFC 4838, which details the architecture for delay-tolerant and disruption-tolerant networks. One aspect of the DTNs are the use of the "Bundle protocol" or RFC 5050. IPv6 will be layered on the bundle protocol. This can be seen in figure 4 below.

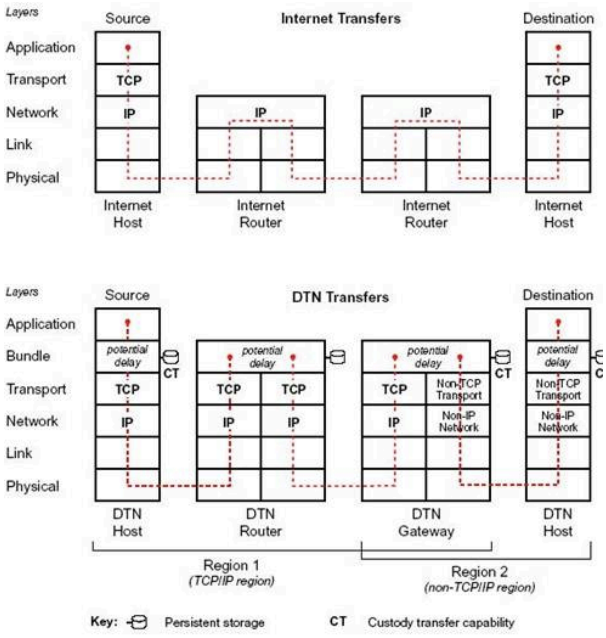


Figure 5. IP transfers over DTNs.

There will also be a high-speed laser link from each one of the launch sites. This will be used in optimal weather conditions as the preferred route. It will use the Lunar Laser Communication Demonstration (LLCD) system. This is a relatively new communications system and has not been tested extensively over time.

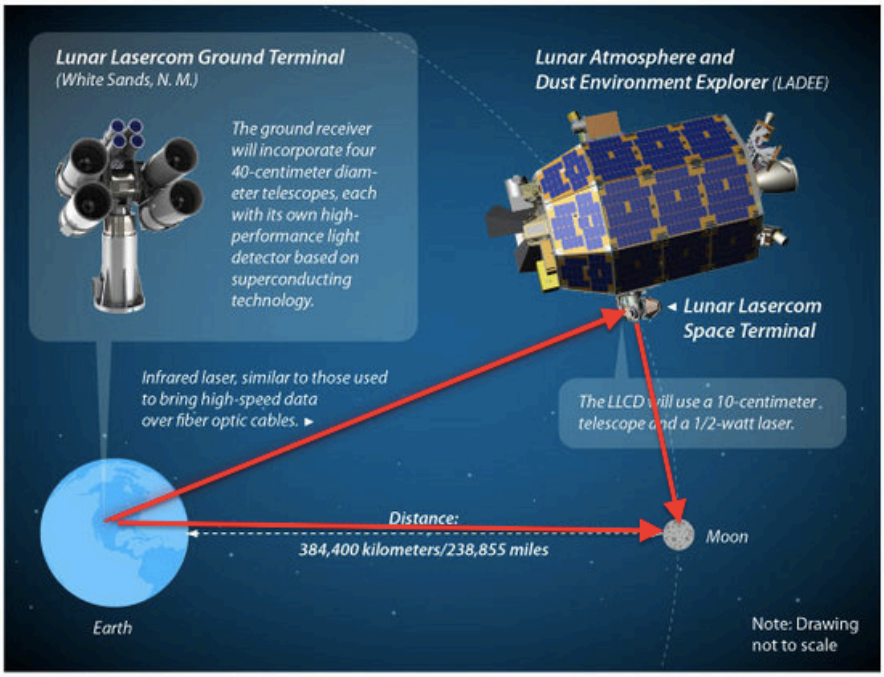


Figure 6. Lunar Laser Communication Demonstration (LLCD) system

The LLCD will attempt a direct link with the lunar base is possible. If there is no direct line of site, then it will relay through the Lunar Atmosphere and Dust Environment Explorer (LADEE).

A mobile version of the LLGT, which has built-in zombie defense mechanisms, is still in development but can be deployed if required. Currently target acquisitioning is a problem in areas populated by seals and surfers.

Figure 7. Prototype of the MLLTG (Mobile Lunar Lasercom Ground Terminal)

The number of hops is reduced in a LLCD path as opposed to a satellite path. The routing protocols are different as well. A satellite-based route will use MLSR (multilayered satellite routing protocol), whereas an LLCD route will use OSPF.

2.2.2 Lunar base conceptual design

Below is the conceptual design of the lunar base infrastructure.

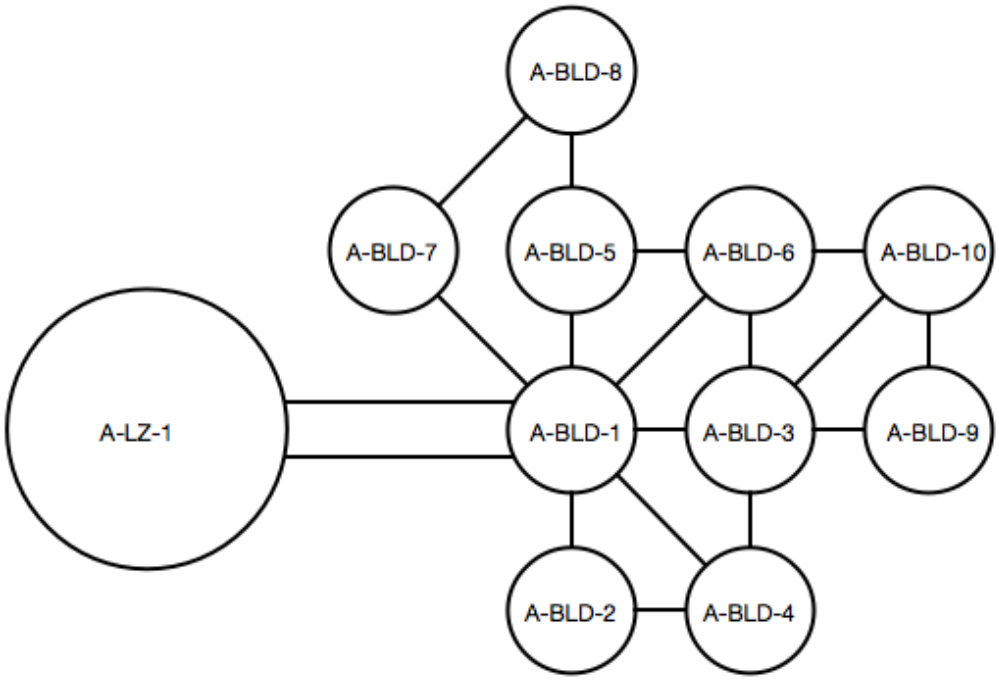


Figure 8. The lunar base Anacreon. Landing zone and buildings.

The primary communications infrastructure will be located in building A-BLD-7. The redundant infrastructure will be located in building A-BLD-4. This is to separate the locations in the case of a catastrophic incident in A-BLD-7 that destroys all the equipment.

2.2.3 Communications Infrastructure conceptual design

Below is the conceptual design for the communications infrastructure in A-BLD-7. It details a high level view of the various component areas required and the inter-relations.

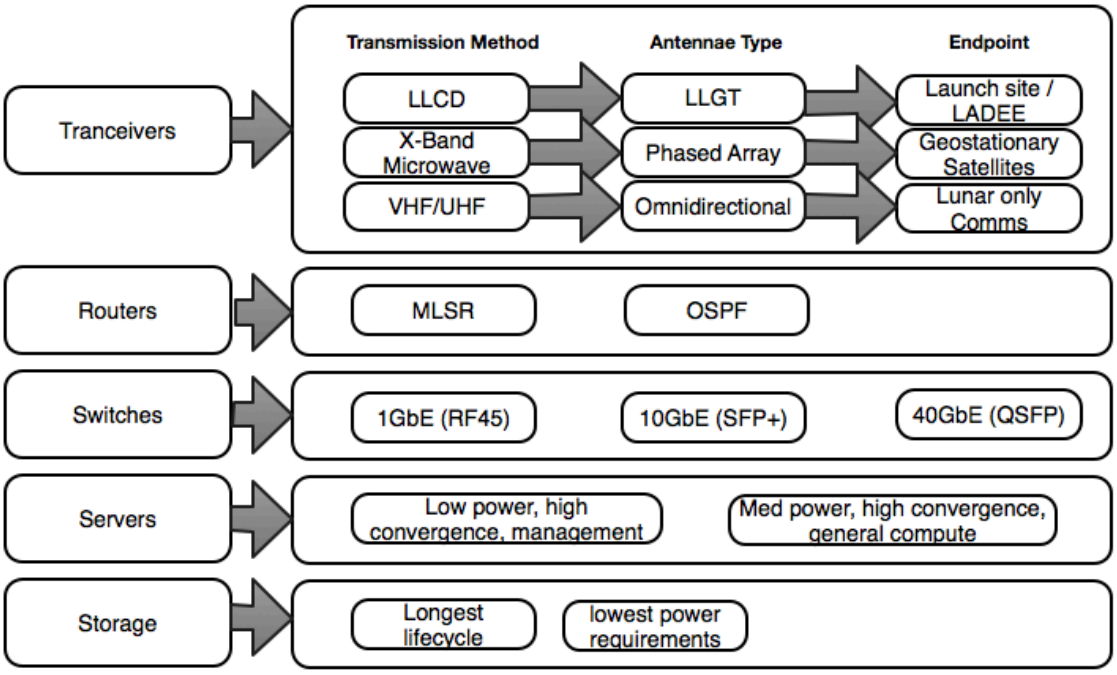


Figure 9. Communications Infrastructure conceptual design

The transceivers will connect to three types of transmission systems. The primary communication with the launch sites will be via optical link with the LLCD.

The secondary link will use X-Band microwave frequencies in the range assigned to the NASA Deep Space Network (DSN). The antenna used is a Phased Array, which allows for software-defined alignment by electrical beam tilting.

The VHF and UHF bands will be used for communications with other regions of the lunar base, as well as mobile communications with rovers.

The routers will use two routing protocols depending on the path they take. MLSR (Multi-Layer Satellite Routing) will be used for the lower speed fail-over route on the satellites.

OSPF (Open Shortest Path First) will be used for LLCD links.

Switching will have several types of ports based on their role, access or aggregation.

Servers will be put into 2 categories, low power and high power. Low power provides core functions and management. High power provides additional compute capability for

large workloads. This separation allows for power conservation and management cluster separation.

The components used for storage are selected based on longevity, resiliency to failure and graceful degradation. Low power utilization is also a key factor.

2.3 Logical Design

2.3.1 Logical design for communications infrastructure in A-BLD-7

Below is the logical design of the communications infrastructure in A-BLD-7

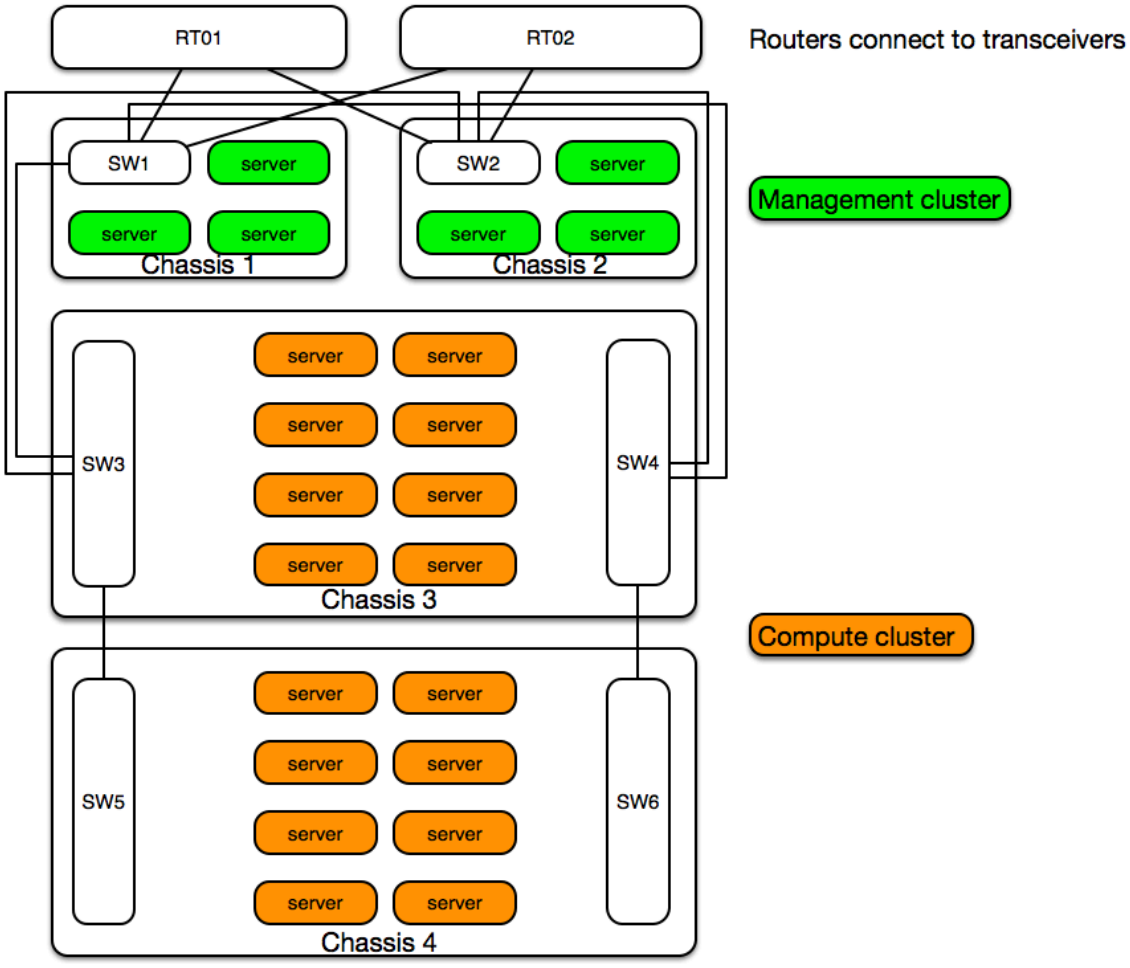


Figure 10. Communications infrastructure in A-BLD-7

The blade servers will be divided into management and compute clusters. Cluster 1 is comprised of blade chassis 1 and 2. These are highly converged units that also incorporate switching and routing in them.

Cluster 2 is comprised of the larger chassis 3 and 4. Chassis 3 will have integrated management and fabric switching. Chassis 4 will connect to the fabric switching on Chassis 3 (cascaded).

2.3.2 Logical design for communications infrastructure in A-BLD-4

Below is the logical design of the communications infrastructure in A-BLD-4

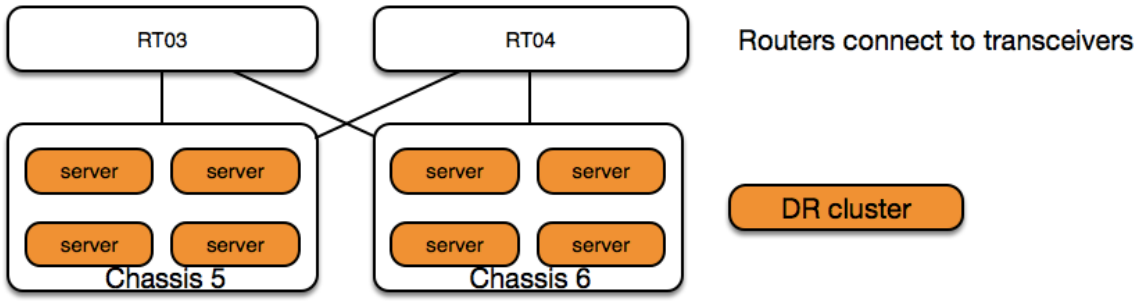


Figure 11. Communications infrastructure in A-BLD-4

This two chassis configuration will provide a DR location for critical systems. It will be a cluster managed by the same vCenter instance as the infrastructure in A-BLD-7.

2.3.3 Logical design for infrastructure power system

The power systems will be on A / B power banks. Each bank is comprised of nickel-hydrogen batteries. The primary power comes from a solar array, then is passed through a charge controller to power bank A. Secondary power comes from a radioisotope thermoelectric generator (RITEG). The decay of the isotope that is encased creates heat, which is converted to electricity via the Seebeck effect.

If a power bank is fully charged then the charge controller can redirect the power source to the other bank. In the case that both banks are full, dummy loads are used to disperse excess energy and heat.

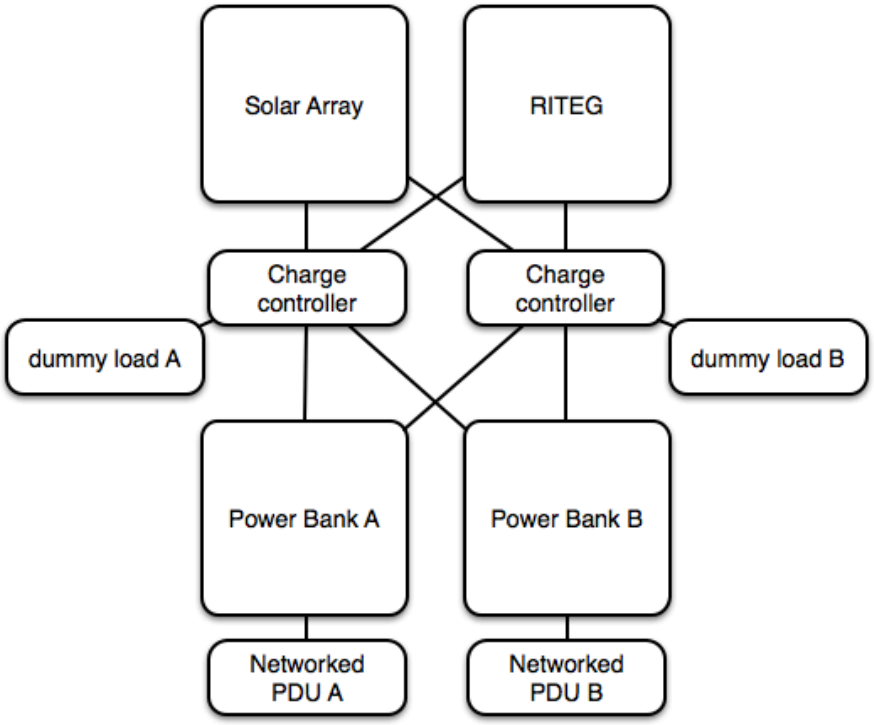


Figure 12. Infrastructure power system logical design

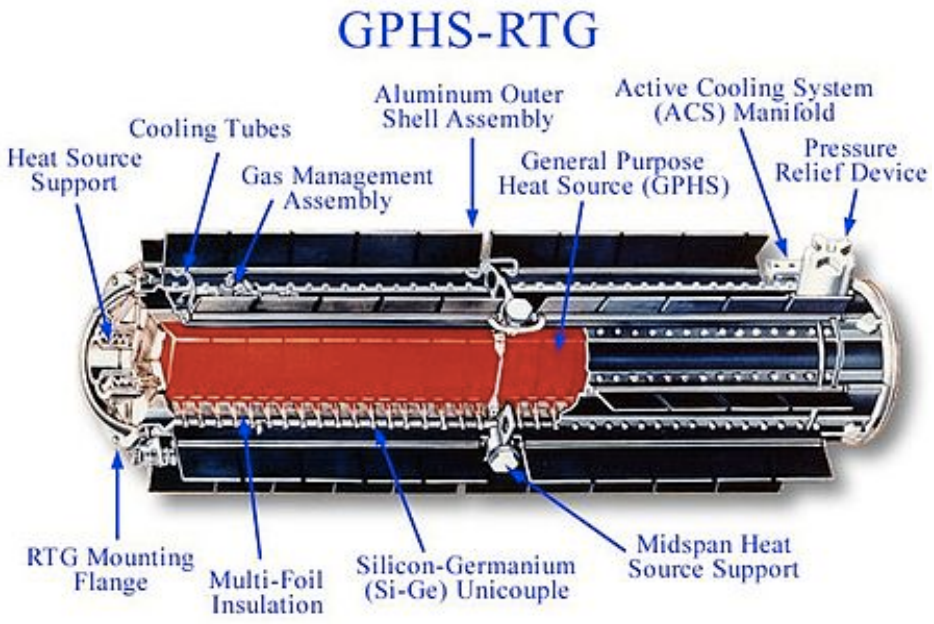


Figure 13. A radioisotope thermoelectric generator (RITEG)

2.3.4 Logical design for virtual infrastructure

The virtual infrastructure will use vSphere 5.5 Enterprise Plus as the hypervisor and licensing level. There will be two separate infrastructures, one production and one for testing, development and DR.

There will be one management cluster and two compute clusters. In the case of a DR scenario, the DR cluster will take over management functions as well.

vCenter will provide centralized management of the ESXi hosts, VMs, and features. vCenter will be installed in a VM to ensure availability via vSphere HA. Since VCSA is not supported in an IPv6 environment, the Windows Server version will be used (see VM Design for OS specifics).

A Simple Install – all components on a single VM – provides scalability and maintains low complexity. If VM numbers exceed the ability of a Simple Install to manage, the components can be broken out by migrating the SSO and Database components to additional VMs. The Auto Deploy component will be installed on the same VM. Auto Deploy is used to deploy ESXi to new hosts.

vCenter SSO will connect to an Active Directory domain. Users will use the VMware vSphere Client or vSphere Web Client and their AD account information for all access to vCenter. See Security Architecture for more details

There will be no vSphere Update Manager (VUM). All patches will be provided via Auto Deploy and the use of an offline software depot.

VMware web-commander will be deployed to provide a self-service catalog where users can provision “X as a Service”, including additional instances of the communications applications, new Active Directory accounts, change passwords, etc. Web-commander provides the front-end to PowerCLI and Powershell scripts. The scripts can perform actions on their own or incorporate the workflow engine of vCenter Orchestrator through a plugin.

Below is the virtual infrastructure object hierarchy.

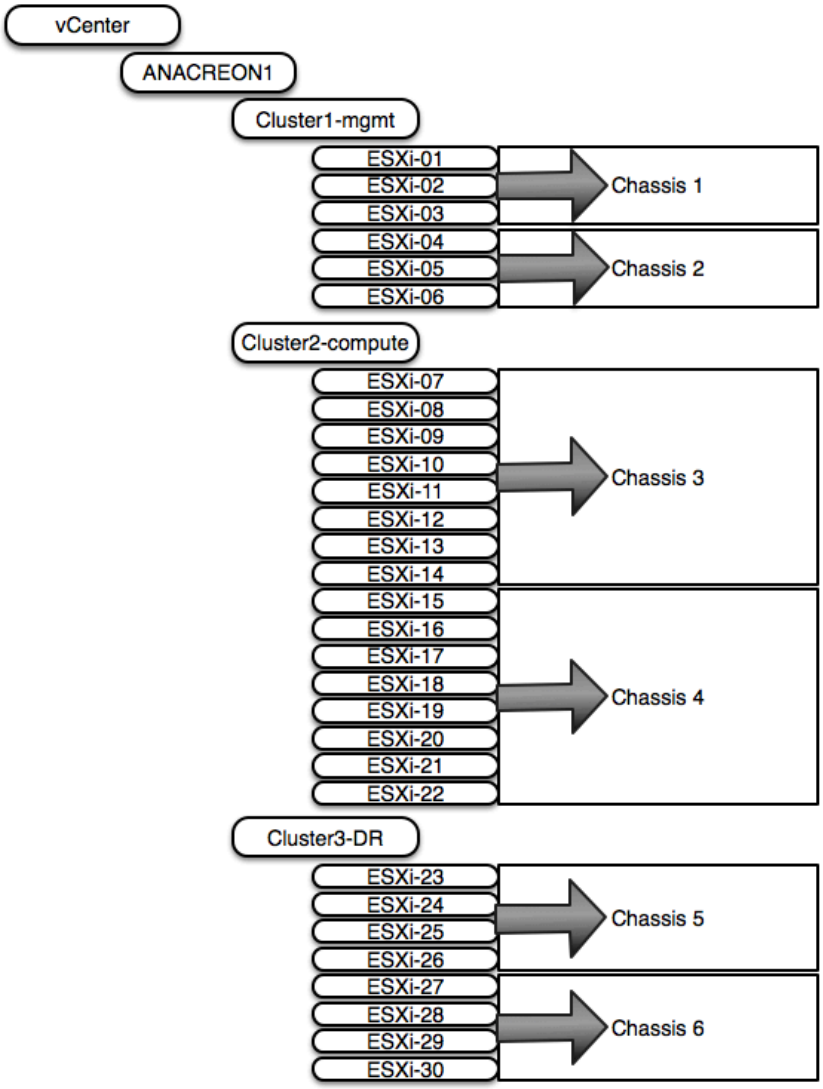


Figure 14. Virtual infrastructure object hierarchy

VM Design

Initial systems VMs are described here.

As described, Microsoft and Red Hat Enterprise licenses have been acquired. Windows Server 2012R2 Datacenter Edition and Red Hat Enterprise Linux (RHEL) 7 are the most recent server editions from each vendor. Windows licensing allows the installation of 2012 R2 or any previous Windows Server edition. All workloads are supported on these two platforms, which will be used throughout the design.

All resource allocations are estimations based on a combination of vendor guidelines and community best practices. Resource usage will be recorded via vCenter and requirements will be revisited after 30 days.

The VMs in the vSphere cloud can be broken into two general groups. Management includes the vSphere centric VMs as well as the automation and orchestration engine. Communications encompasses the communications applications and their attendant services. Some services may involve both systems; these will be classified as Management.

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.

Windows Domain and vCenter

General LAN and End User access requires integration with the existing Active Directory forest. A new domain tree will be created and two Windows 2012R2 VMs will be provisioned as domain controllers for this tree. Windows 2012 R2 Datacenter licenses have been acquired and all additional Windows VMs will also run 2012 R2 unless otherwise specified. Additional domain-related VMs include RDP servers for remote access and management stations, an RDP Licensing Server. The vCenter server will be installed on Windows as well. This table lists the initial resource allocations and VM quantities.

Service	vCPUs	RAM (GB)	System disk (GB)	Data disk (GB)	Quantity
Domain Controller	2	8	60	0	2
RDP Session Host	2	32	60	300	4
RDP Licensing	1	4	60	0	2
vCenter	4	32	100	1000	1

vSphere Systems and Appliances

VMware Data Protection Advanced will provide backup services. Initially, two VDPA instances are required. Additional instances will be deployed as required to maintain a 25:1 ratio of VMs to VDPA instances of 25:1, as suggested by the VDPA Administration Guide (page 18).

All management VMs will be backed up daily and backups retained for 90 days. Of the communications VMs, only the database servers will be backed up daily and backups retained for 90 days. Other communications VMs are considered “cattle” and will only hold data in transit to or from the database. These services will be redeployed in brand new VMs as necessary.

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

Service	vCPUs	RAM (GB)	System disk (GB)	Data disk (GB)	Quantity
IaaS Components (Windows 2012)	2	8	30	0	1
VDPA	4	4	3100	0	2
vShield Manager	2	8	60	0	1
vShield Edge	2	1	0.5	0	1
vShield Endpoint	1	0.5	512	0	6

Additional RHEL VMs are required to complete the management system. The automation and orchestration system require Kickstart, Puppet Master, Gitolite, and Jenkins CI VMs (see Puppet System). Only one of each server is required to complete the system. The Puppet Master may have scaling issues if it needs to support over 1000 VMs, at which point additional master systems would need to be created. Puppet includes no built-in synchronization methods when there are multiple masters and this would introduce unnecessary complexity if it were not needed. The number of masters will be revisited after 30 days and adjusted if necessary. This table shows these VMs and their initial resource allocations and quantities.

Service	vCPUs	RAM (GB)	System disk (GB)	Data disk (GB)	Quantity
Kickstart	1	0.5	100	0	1
Puppet master	4	8	100	0	1
Gitolite	2	8	500	0	1
Jenkins CI	2	8	500	0	1

Communications System

The communications system has been developed by the Foundation. It is a three-tiered system consisting of a Web Front End (nginx), a Message Queue tier (RestMQ), and a Database tier (MongoDB). This system is designed for high scalability to support ever-larger space-station designs, thus it will require a Load Balancer to manage connections to the various VMs at each tier. There are also two Monitor (watchdog) VMs that monitor system load and control how many VMs are deployed at each tier. Each VM has a specific resource allocation. The watchdogs will monitor utilization and create or destroy VMs as needed to maintain services levels. There is no defined upper bound but there is a minimum of two service and one load balancer VMs per tier.

The communications infrastructure relies on continuous integration and continuous deployment processes to improve the system bandwidth, reliability, range and correct errors through rapid code deployments. In order to ensure these rapid deployments do not have an adverse effect on the communications system, Development adheres to a strict change control process that involves three environments: Development, QA, and Production. Any change must be promoted upward through each environment and test successfully before promotion to Production. Any code issues or oversights are caught before the production manufacturing equipment is tested.

To achieve this goal, the vSphere cloud must deploy the communications system three times in these environments. This table shows the VMs and their initial resource allocations, plus per-environment and total quantities.

Service	vCPUs	RAM (GB)	System disk (GB)	Data disk (GB)	Quantity
Web Front End	1	1	50	0	6
Message Queue	1	2	50	0	6
Database	2	4	50	200	6
Watchdog	1	0.5	50	0	6
Load Balancer	1	4	50	0	9

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

vCPUs	RAM (GB)	Disk (GB)	Quantity
89	328	12833	50

Web-Commander Portal

The web-commander system components are installed according to the VM Design section. The portal allows authorized users (vSphere admins, manufacturing application developers, and certain high level Foundation officers) to provision numerous objects types ("X as a Service").

Day one blueprints in the Service catalog will include all of the standardized templates for the existing VMs (see VM Design), Active Directory objects and actions (new users, reset password, etc.).

vSphere admins will have the ability to manage the web-commander catalog, adding and updating scripts and entries as appropriate. Developers will NOT have the ability to update the catalog, as Puppet will be used to determine the correct application load out for each VM (see Puppet System), and will program the communications system to use the REST API to provision objects as needed.

Management VMs will be restricted to vSphere admins only and can be deployed in any environment/network. Communications VMs are restricted to developers and can be deployed in any of the communications environments/networks (initially Development, QA, Production). Web-commander scripts will rely upon vCenter templates (Windows), kickstart processes (RHEL) and Puppet services to provision new VMs. Other objects will rely on various services specific to the object type, such as ADS for user services.

Puppet System

The Puppet Configuration Management (CM) product is at the heart of the automation and orchestration engines and the manufacturing continuous delivery/continuous integration processes. The Foundation has chosen Puppet Open Source based on its popularity and familiarity to our personnel, and because of the well documented features and modules that allow it to manage both Windows and Linux VMs (or nodes). Puppet allows the developers to define the desired state of the node and ensure that state is achieved without worrying about how to get there.

Because Puppet applies a desired state to a system, it can start with a bare-bones template and add the required configuration during provisioning. This allows the use of a single Windows VM template per OS level (2012 and 2012R2) and the kickstart of all RHEL VMs. Node definitions can be adjusted and very rapidly, Puppet will bring all of the desired nodes into compliance with the new desired state. This “infrastructure as code” process allows for rapid code development and rapid application deployment.

All system-provisioning requests (Windows or Linux) will be made through web-commander. When the user selects a VM via the catalog, the appropriate scripts will initiate a workflow to provision the VM.

1. User interactively selects an item in the catalog and submits the request.
2. Selected script creates the kickstart file for provisioning (Linux only) and deposits it on the Kickstart VM.
3. vCenter Orchestrator initiates the VM provisioning workflow
4. The workflow builds the VM from template (Windows) or creates an new VM of the correct size (Linux)
5. (Linux) The VM receives a kickstart file and downloads/installs the appropriate files.
6. The workflow forces the VM to register with puppet and signs the certificate (ref).
7. The vCO workflow forces the new VM to check in to Puppet and receive its initial setup.
8. vCO repeats steps 2-6 for any additional VMs required by the selected catalog entry.
9. vCO notifies the requestor the catalog request has been fulfilled.

The vSphere administrators are responsible for configuring Web-Commander, the Kickstart server, the vCO workflows, and ensuring communication between all components shown above.

The Kickstart server provides RHEL system contents for newly provisioned VMs, plus the kickstart file for the new VM. Each RHEL VM receives nearly the same basic kickstart file, only the node elements such as networking vary. Any two provisioned VMs will start with the exact same minimum level of software required to run Puppet.

The Web-commander scripts communicate between vCO, Kickstart, and newly provisioned nodes. The vSphere administrators and the developers work together to create the proper workflows.

The vSphere administrators will manage the puppet master instance(s) and the puppet code for management VMs. The developers will manage the puppet code for all remaining VMs. Both groups will provide oversight and assistance to each other, ensuring basic sanity checks and adherence to process, which ensures the manufacturing SLA (99.99%) is maintained.

Developers are responsible for aligning the puppet code with manufacturing releases. The puppet code determines the system configuration performed during provisioning step #6. The Web-commander catalog will allow selection of a number of environments, including Development, QA, and Production. Developers are responsible for ensuring the Watchdog services provision in the correct environment.

All code changes also involve the Gitolite and Jenkins CI VMs. Gitolite is the authoritative version control source for all Git repositories. Developers will have their own Git workspaces, but only Gitolite and the RDP servers are backed up by VDPA. Jenkins CI is a Continuous Integration tool that is used to test changes to the communications application code.

The development workflow defines that all code must be tested by Jenkins CI and pass all tests to be merged upstream. This includes both puppet and communications code, further validating code before deploying to any environment and assisting in maintaining the 99.99% SLA.

vSphere administrators will also follow the Gitolite/Jenkins CI and Dev->Production workflow when modifying puppet code for the management systems. Puppet will also be used to ensure a consistent state and make changes to vCenter. Developers will provide assistance as needed.

The threat from Zeds to humanity's very survival is real and the communications capabilities of the Foundation are vital to colonizing Mars and beyond. The use of Dev->Production promotion ensures that all changes, even those rapidly required to adapt to future manufacturing requirements, are thoroughly vetted for software bugs AND material output quality before being introduced to Production. The potential for outages can never be eliminated but combined with other availability measures; the possibility continues to reduce to a near-0 value.

VMware Datacenter Design

The Foundation's vCenter server will define one datacenter for the Anacreon lunar base . A single cluster of 6 UCS ESXi hosts will be provisioned immediately. This cluster will be for the management VMs only. The second cluster will have 16 ESXi hosts across 2 blade chassis. The cluster can scale up to 20 ESXi hosts, which is the limit of the UCS mini domain. The third cluster will have 8 ESXi hosts and will be for DR, test/dev and QA.

If additional ESXi host are required, then additional power is required to make it work. The cluster can also scale down when needed to a minimum viable level of 4 hosts to maintain the minimum number of Management and communications VMs for the anticipated workloads.

To meet the 99.99% SLA, the cluster(s) will be configured with High Availability (HA) and Distributed Resource Scheduling (DRS). Due to the different hardware within the chassis, Enhanced vMotion Capability (EVC) is required and will be enabled at the maximum compatible level..

HA will have an initial admission control policy of 50% of cluster resources to provide for 3 host failures on the management cluster, 50% for the compute cluster (to allow for a chassis failure of 8 hosts). Admission control will be disabled for the DR cluster, as violations may be required in certain emergency scenarios. This will be revisited every 30 days as communications capacity increases and cluster sizes vary. 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 communications system scales itself. DRS rules will help ensure availability of management VMs with multiple instances.

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

Rule Type	VMs
DRS VM-VM Anti-Affinity	DC1, DC2
DRS VM-VM Anti-Affinity	RDPLicense01, RDPLicense02
DRS VM-VM Anti-Affinity	RDPSH01, RDPSH02
DRS VM-VM Anti-Affinity	RDPSH03, RDPSH04
DRS VM-VM Anti-Affinity	RDPSH01, RDPSH04
DRS VM-VM Anti-Affinity	VDPA01, VDPA02
DRS VM-VM Anti-Affinity	DevWatchdog01, DevWatchdog02
DRS VM-VM Anti-Affinity	QAWatchdog01, QAWatchdog02
DRS VM-VM Anti-Affinity	ProdWatchdog01, ProdWatchdog02
VM Override VM Restart Policy - High	Management - vCenter, DCs
VM Override VM Restart	Automation - vCO App, Gitolite VM

Policy - High	
VM Override VM Restart Policy - High	Communications - Database and Watchdog VMs
VM Override VM Restart - Low	Web Front End VMs

Security Architecture

The security of the manufacturing security is extremely 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 to 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) and the puppet manifests will manage additional rules for each VMs installed applications. 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 5.5.2. 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.

Lunar Communications to Foundation terrestrial links			
SRC	DST	SERVICE	ACTION
Internal Networks	External Networks	http, https, ssh, smb, dns	Permit
Terrestrial links to Lunar Communications System			
SRC	DST	SERVICE	ACTION

vSphere Admins	vCenter	9443/tcp	PERMIT
vSphere Admins, Developers	Puppet System	ssh	PERMIT
vSphere Admins, Developers	RDP Session Hosts	rdp	PERMIT

The system's edge, between the manufacturing network and the Foundation's LAN/WAN, will be protected with a pair of Cisco ASA class devices. Initial policy will allow unrestricted outbound common services and more restricted inbound services according to the table below.

vCenter SSO will use the Active Directory domain as the primary namespace. All vSphere administrators and the chosen colonist will be in the Administrator group. Developer team leads will be in the Power User role. Other developers will have read-only access unless specifically requested and granted. The administrator@vsphere.local account information is made known to the vSphere team lead and the colonist (prior to launch) 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 (again prior to launch) in case local DCUI/shell access is required.

Network Design

The network configuration will be the same as in challenge 1 with 4 VLANs

VM Network VLAN 10

vMotion VLAN 20

Storage VLAN 30

WAN / External networks VLAN 1001-1999

Each server will have one standard vSwitch with a passive LAG and 2 x uplinks. QoS will be managed by using NIOC. VLANs will be assigned to the associated port groups.

2.3 Physical Design

The blade architecture is comprised of Cisco UCS-E series and UCS-B (Mini) series servers. Here is an overview of UCS-E.

Cisco UCS-E series servers are a class of highly converged micro-blades that reside in Cisco ISR G2 routers. They are available in two form factors, single and double wide. The single wide can accommodate up to 16GB of RAM, whereas the double-wide can accommodate up to 48GB of RAM. An ISR G2 router can contain 2 double wide, or 4 single wide micro-blades.

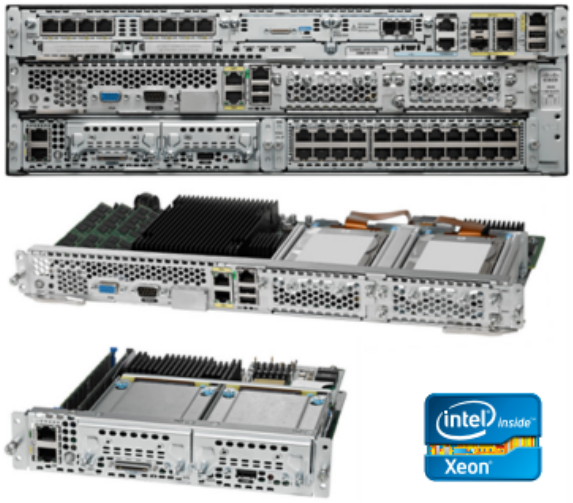


Figure 15. Cisco UCS-E servers with a Cisco 3945 ISR.

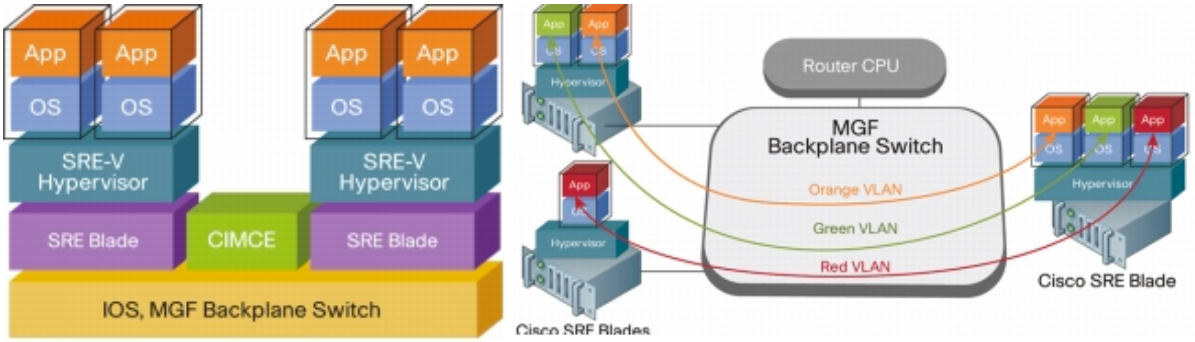


Figure 16. Cisco UCS-E components

Here is an overview of the UCS-B mini solution.

The new Fabric Interconnect (FI-6324) is the main difference from a standard UCS-B solution. The form factor is much smaller and fits into the IO Module slots in the chassis. The result is a very compact solution

A single baby UCS domain based on the 6324 Fabric Interconnect can still support up to 20 servers spread across 2 chassis (8 blades per chassis) and 4 Cisco C-series rack servers.

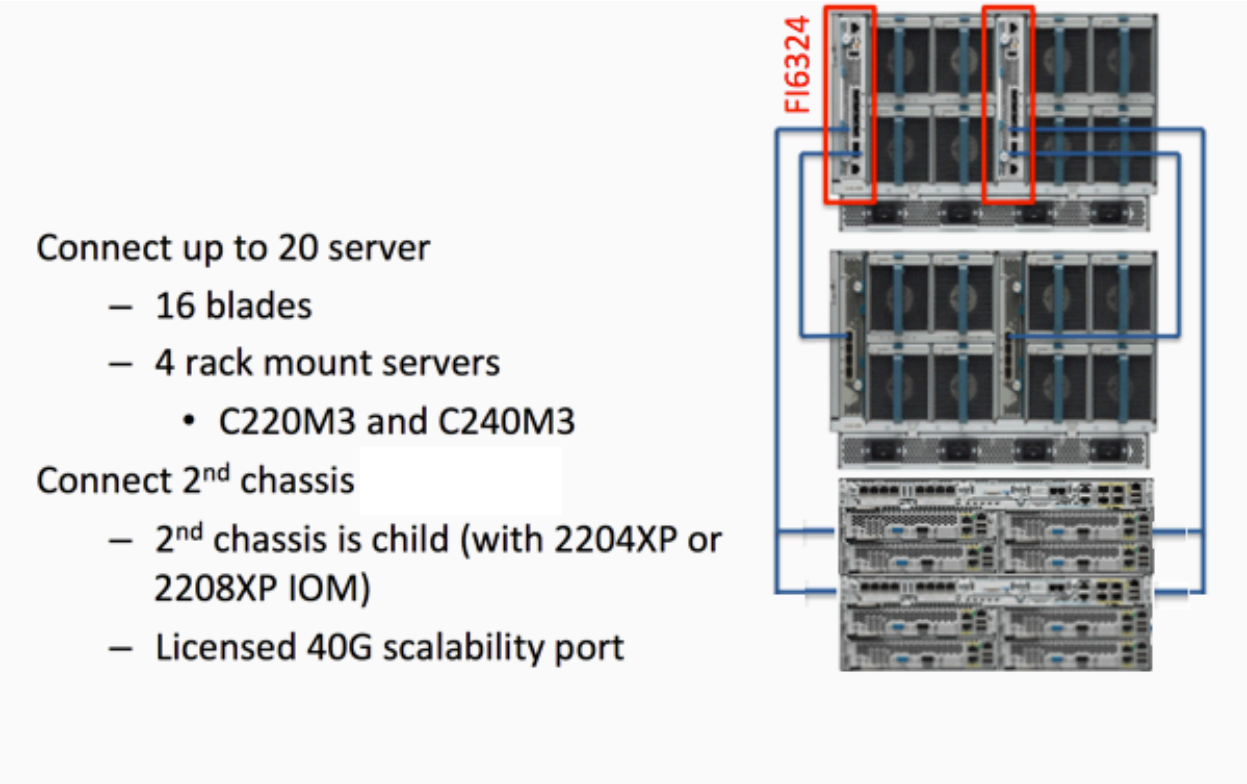


Figure 17. Cascaded UCS-B Mini with connected 3945 ISR G2 routers with UCS-E blades

The infrastructure in A-BLD-7 is as depicted in figure 17, but with the ISR routers on top.

The UCS-E servers will have the following specs:
Quad core E3-1100 Intel CPU
16GB RAM
2x 400GB SLC SSD
2 x SD cards (boot)

The UCS-B servers will have the following specs
Dual socket x 10core E5-2600v2 processors
256GB RAM
2 x 2TB SLC SSD
2 x SD cards (boot)

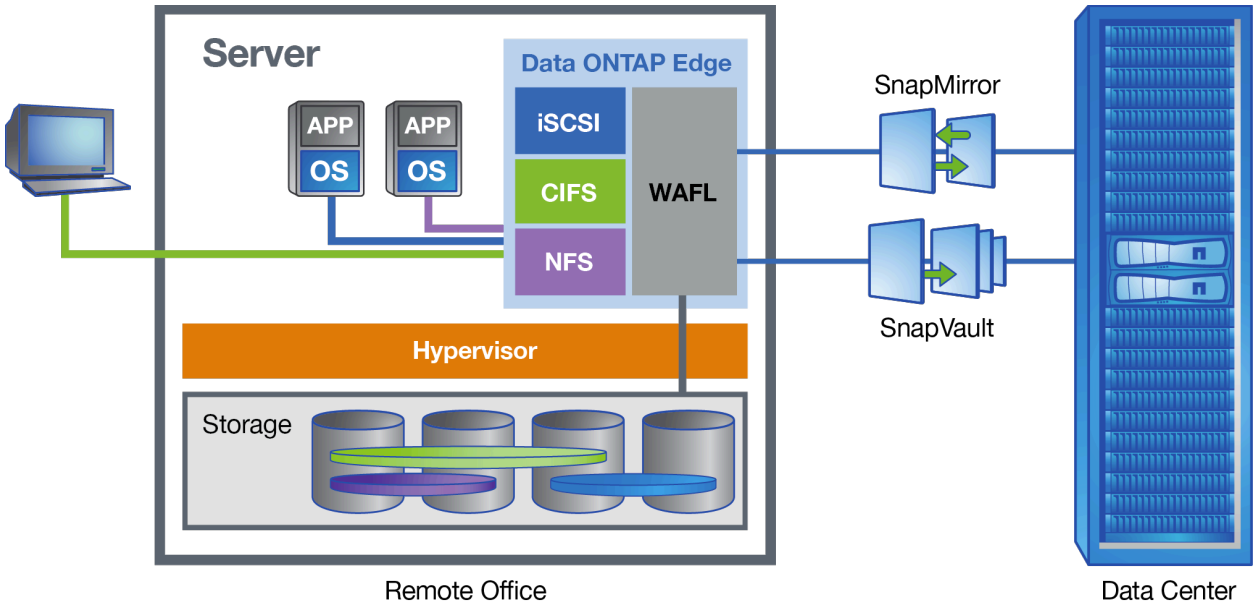
Cluster ID	Total Compute	Total Memory	Total RAW Storage
Cluster1-mgmt	6 hosts x 4cores x 1.8Ghz = 43.2Ghz	16 x 6 = 96GB	2x400x6=4.8TB
Cluster2-compute	16hosts x 20cores x 2.6Ghz = 832Ghz	256 x 16 = 4TB	2x400x16=12.8TB
Cluster3-dr	8 hosts x 4cores x 1.8Ghz = 57.6Ghz	16 x 8 = 128GB	2x400x8=6.4TB
			Usable is about 40% if RAID1 and overhead is calculated

Shared Storage Configuration

Shared storage is required to allow VM workloads to migrate from ESXi host to ESXi host without impacting performance levels. SLC SSDs are the only type of disk that will be used because of longevity constraints. The average lifespan of an enterprise grade SLC SSD is >20 years with a 50/50 r/w profile.

NetApp is another commonly known vendor that personnel may already know or can learn quickly.

One FAS8020 will be configured as primary storage and set to replicate to OnTap Edge VMs on the hosts via snapmirror. The OnTap Edge VMs will make use of the local storage on each one of the hosts. This allows for a distributed and redundant storage infrastructure.



A NetApp FAS8020 will be used for the shared storage. With 24 x 600GB SLC SSDs. With RAID-DP and 2 spares, the total space available is 6.57TB. The calculation is shown below.

NetApp Usable Space Calculator 2.1

- Murugappan Periyakaruppan

(murugappanp@gmail.com)

Input

Total Disks	24	Disk Size / Type	600GB SAS
Spare Disks	2	Disks Allocated for Root Volume	2
Raid Group Type	Raid-DP	Output Unit Format	TB
Volume Reserve	10 %	Raid Group Size	14
WAFL Reserve	10 %	Aggregate Reserve	5 %

Usable Space

Total Raw Capacity	14.06 TB	Total Base 2 Capacity	13.10 TB
Total Right Size Capacity	12.82 TB	**Usable Capacity**	6.57 TB

Raid Group Size Estimator

RG Output Based On: Optimal Capacity

The infrastructure with storage in A-BLD-7 will look as follows

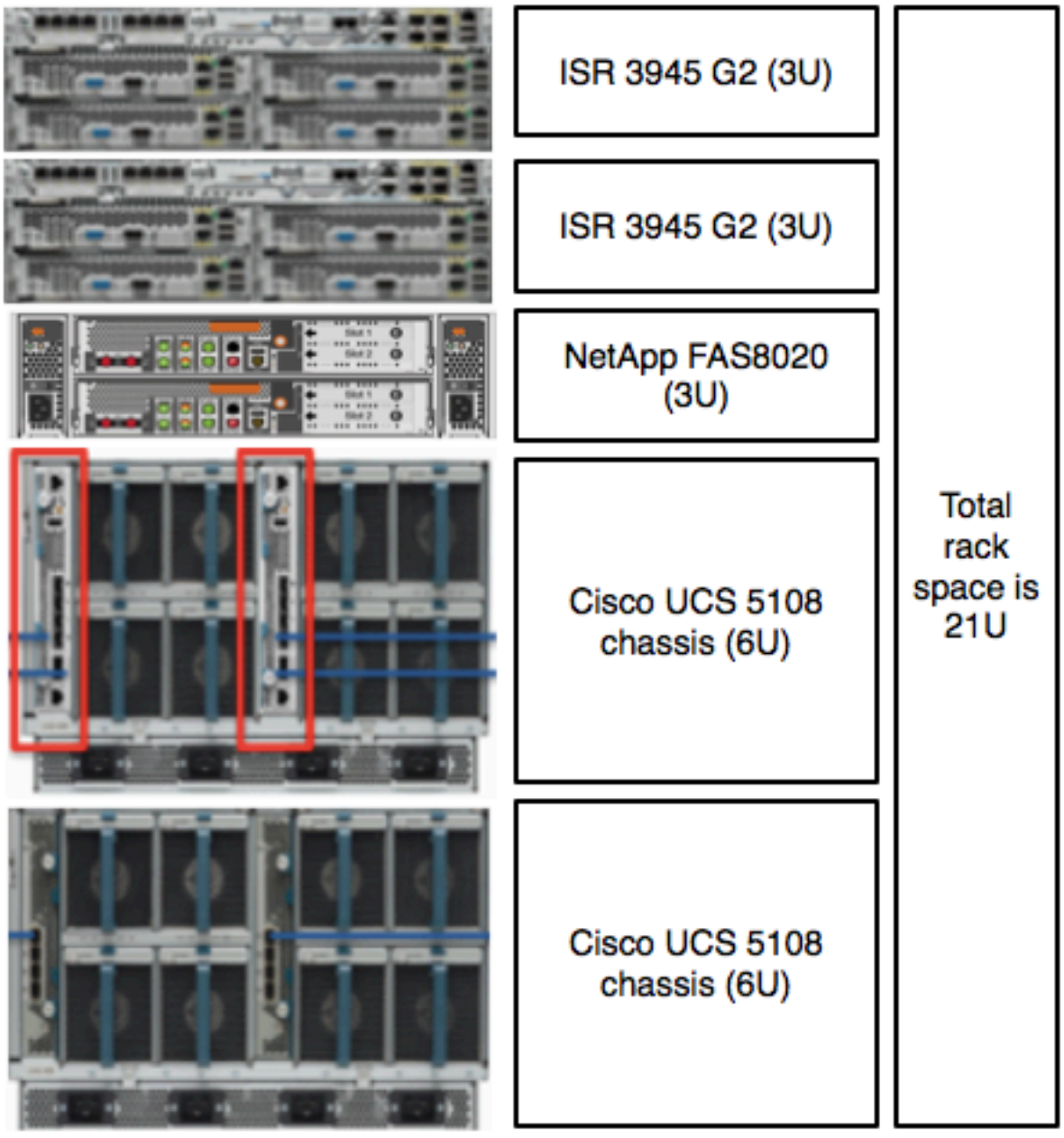


Figure 18. Infrastructure of A-BLD-7

The infrastructure of A-BLD-4 will look as follows

ISR 3945 G2 (3U)

ISR 3945 G2 (3U)

Total rack space is 6U

Power utilization for the 2 x 5108 chassis and the 14 UCS-E servers

Cisco Unified Computing System totals

Number of racks: 1

Available ports per Fabric Interconnect: 16

		Fabric Interconnect	Chassis	Rack-Mount Servers	Total
Number configured		2	2	0	
Idle power	watts	480	1768	0	2248
	BTU	1637	6030	0	7667
50% load power	watts	642	3282	0	3924
	BTU	2190	11190	0	13380
Max power	watts	804	5047	0	5851
	BTU	2742	17211	0	19953
Weight	kg	31	221	0	252

US/Metric

Tons cooling: 1.1

Total number blades configured: 16

Cisco Unified Computing System totals

Number of racks: 1

		Fabric Interconnect	Chassis	Rack-Mount Servers	Total
Number configured		0	0	14	
Idle power	watts	0	0	2239	2239
	BTU	0	0	7630	7630
50% load power	watts	0	0	3503	3503
	BTU	0	0	11942	11942
Max power	watts	0	0	4805	4805
	BTU	0	0	16380	16380
Weight	kg	0	0	171	171

US/Metric

Tons cooling: 1

Total number blades configured: 0