THE FOUNDATION

Challenge 2 – Adaptation & Constraints

By Rob Nelson (@rnelson0)

THE FOUNDATION

Executive Summary

The Foundation's Anacreon base on Luna will be the waypoint for humanity on our way to the new Mars colony. Anacreon has established significant constraints for the power, cooling, and space allocated to the manufacturing systems, as the majority of these resources are designated for life support and other critical services. Anacreon uses only IPv6 networking and the manufacturing system must be converted to IPv6 for integration. The system must be scaled down to meet these resource constraints and a new functional level defined.

THE FOUNDATION

Detailed Background

The launch facilities at Cape Canaveral are almost online. Additional facilities in the Netherlands, Australia, and New Zealand are now under construction. We are nearing the beginning of a new Space Age that will define the course of humanity in a much more significant manner than the Space Race of the 20th Century.

Mr. Seldon has separated the Foundation designers into two teams and assigned one of two Earth-based manufacturing designs to each team for analysis and integration with the existing Anacreon systems. Two winning designs will be chosen for diversity and must be both self-reliant and able to communicate with the other design.

Anacreon itself provides significant benefits to our system, but also some downsides. It is located in the underground chamber in the Oceanus Procellarum area of the moon [link]. The conversion of this natural vault to a human-ready base provides a stable temperature, protection from micro-meteor impacts and dust, and protection from interstellar radiation that could negatively affect computerized systems without expensive and bulky protection.

The size of Anacreon allows for a large number of colonists without overcrowding, but the projected numbers also require the computer systems to occupy little physical space. The majority of Anacreon will be dedicated to life support systems, including heating and ductwork to raise the temperature from -20C to a more comfortable level. Anacreon's existing infrastructure will be extended to the manufacturing demark and allow wide-area network access to the manufacturing system.

The two designs must be modified to integrate with these existing systems. The most significant resource limitation in Anacreon is space. Two half-racks of datacenter space, 21U each, have been made available to the two systems. Mr. Seldon has also tired of Earth's continual delays to implement IPv6 and has restricted Anacreon to be an IPv6-only network. The chosen designs must be modified to support IPv6 only. Where necessary, vendor products without viable IPv6 support can be replaced by IPv6-compatible systems from the same vendor.

These significant changes to the designs requires each designer to enumerate the configuration maximums and failure tolerance of the modified solution.

THE FOUNDATION

Overview

The manufacturing system designed by Daemon Behr has been chosen to integrate with Anacreon. This design supports the same 3-tiered manufacturing system and automation as the Foundation's Earth-bound design. Its physical design requires nearly 7 full racks (just under 294U) and must be reduced to a half rack (21U) of space or less.

This modified design will be connected to Anacreon's existing IPv6-only network. Vendor solutions must be re-evaluated for IPv6 support and may be replaced with alternative product lines from the same vendor as required.

The design must be entirely self-reliant, yet maintain the flexibility to integrate with the existing Anacreon facilities and the other chosen infrastructure design.

The completed re-design's new maximums and fault tolerance are described in detail.

THE FOUNDATION

Requirements

REQ1	The same 3-tiered (web/messaging/database) application is supported and used in the modified system.
REQ2	The solution must run in a micro-gravity environment with background radiation levels higher than Earth's.
REQ3	The system must be self-reliant and continue operating when external network connectivity is unavailable.
REQ4	The system must have the capability to interact securely with other systems when external network connectivity is available.
REQ5	Two support teams at Anacreon must run the system.
REQ6	The system must be maintainable if one full support team is lost.
REQ7	Manufacturing workloads will continue to scale up and down as needed.
REQ8	New configuration maximums (hosts, VMs, storage capacity) must be enumerated.
REQ9	Failure tolerance must be enumerated.
REQ10	The existing SLA (99.99% 24x7) must be maintained.

Constraints

C1	Only 21U of rack space is available for the full system. This includes sufficient power and cooling.
C2	The same vendors as the originating Earth-bound design must be used. Alternative product lines from the same vendor are acceptable.
C3	Anacreon's network is IPv6 only and new systems must also use IPv6 only.
C4	Active Directory services are already present on Anacreon and the system must integrate with the existing forest.
C5	A replica of the *Encyclopeda Galactica* database and software must be maintained in the system.

Assumptions

A1	The entire system manifest will arrive intact from Earth. No component failures will occur before implementation.
A2	The Foundation has acquired appropriate licensing for all vendor's products via government mandate.
A3	All specified hardware and software will operate in perpetuity without access to Earth's Internet (i.e. no "phone-home" for licensing).
A4	Computer facilities will have protection from interstellar radiation due to the underground location of Anacreon, preventing entropy in running systems and damage to hardware and stored data.
A5	Manufacturing iterations of 72 hour no longer apply.
A6	The primary support team for this system has and can maintain the skillsets required to implement and maintain the solution.
A7	The physical manufacturing plant will have the staff to be properly maintained and ensure the relevance of the virtualized control systems.
A8	Where IPv6-capable products require IPv4, this is acceptable if the IPv4 elements are local-network only and secured from traversing the network.
A9	If a product does not explicitly offer support for IPv6, through a vendor support statement or documentation referencing IPv6 configuration, the design assumes zero IPv6 support for the product.

A10	If a vendor supports IPv6 and does not explicitly state IPv4 configuration is required, the design assumes the product supports IPv6 only operation.
A11	SCADA/HMI systems are, or can be made, IPv6 compatible by our developers.

Risks

RISK1	If one support team is lost, the surviving support team may not have the familiarity with the system (possibly being the alternate system's support team) or the time to quickly familiarize themselves with it, as they will have two systems to support.
RISK2	There is no existing solution at this scale to measure anticipated resource requirements
RISK3	Power loss, environment de-pressurization, solar flares and radiation, and other variables well beyond the control of the manufacturing system may have negative effects upon its viability.
RISK4	Delays in the Mars colony's availability may force too many colonists upon Anacreon and could increase resource constraints upon the system, including further loss of space for the manufacturing system.
RISK5	A Zed infection on Anacreon could result in massive, and possibly total, loss of life, rendering the design moot.
RISK6	Incomplete or incorrect vendor documentation on IPv6 support and setup may result in modifications to the design during implementation.
RISK7	Administrator's and developer's lack of familiarity with IPv6 may stunt the system's growth and reduce ability to troubleshoot.

THE FOUNDATION

Architectural Design

Daemon Behr's original design decisions (DB#) will be revisited and compared to the increased constraints.

DB1	The system will have the fewest number of logical components possible
	Per C1, reducing the physical size from ~294U to 21U will require further reduction in logical components.
DB2	Automation will be done by scripting as opposed to using an automation engine
	Scripting requires more familiarity from the support teams and may violate RISK1. Automation will be used to mitigate the risk.
DB3	There will be Production, Development and QA environments
	This addresses REQ1, REQ7 and REQ10.
DB4	The facility will be fortified
	This relates to REQ2, REQ6 and A4, and reduces RISK1, RISK3, and RISK5.
DB5	Continues Integration and Continuous Delivery will be implemented
	This overall design decision continues to hold true. However, A5 and REQ5 will likely affect the rate of iterations such that hourly changes are unlikely to be feasible.
DB6	Technical teams will be made up of 9 groups of 8 people. 24 people in Prod, 24 in QA, and 24 in Dev.
	REQ5 provides us with two support teams of unspecified size. This design's team will support all three environments and all three shifts. There will be no 24 hour shift coverage initially. NASA research has shown significant impact on astronaut's alertness depending on sleep schedules [1]. The support teams will attempt to follow a "regular" Earth schedule, including an approx. 9-5 work schedule. There is no present literature on the effect of long-term sleep schedules on the moon, which experiences a ~29.5 Earth day-long Solar day. Foundation doctors on Earth will make further recommendations as research results are presented.

With these adjusted design decisions and the new constraints in mind, additional design decisions (RN#) are required.

RN1	All systems will be designed for IPv6 only.
	Per C3, all IPv4 configuration will be removed or effectively disabled. Some logical components may need replaced (C2, A8, A9, A10).
RN2	Hardware firewalls will segregate networks.
	While virtualized firewalls would reduce north-south traffic, there are no generally available and well known products to provide this functionality from the selected vendors (addresses A6, A9, A10, mitigates RISK6). The increased north-south traffic must have limited impact upon network traffic. This setup needs to remain simple to mitigate RISK7.
RN3	The manufacturing system will be enclosed and self-reliant
	Per REQ3, REQ4, the system will be able to function post-implementation without external network connectivity and will have a clear demark and security policy at the edge.
RN4	Designate configuration maximums and failure tolerance.
	RISK2 and RISK4 put severe pressure on the manufacturing system to be small, but powerful. To assist developers, administrators, and users, the configuration maximums and designed fault tolerance must be clearly available at all times (REQ8, REQ9) and easily understood.

THE FOUNDATION

IPv6 Design

Anacreon's IPv6-only constraint requires massive changes to the original system. All existing components were investigated for IPv6 support. Much vendor documentation does not mention IPv6 or does not specify its supported functionality. Given the severe consequences of failure in the closed-system of Anacreon, some assumptions were made to ensure only supported products were selected.

If vendor documentation (including installation/administration guides) does not reference IPv6, it is presumed to be unsupported in any way (A9). If the vendor documentation supports IPv6 capabilities and does not explicitly require IPv4 be present for operation, it is assumed that the product will work in an IPv6-only environment (A10). Due to the incompleteness and ambiguity in some vendor documentation, it is possible that not all products will properly operate with a lack of IPv4 networking, which would require revisiting some portions of the design (RISK6). Due to the well-known issues with IPv6 implementation and support at the time of the Zed outbreaks and the lack of ongoing development since, it is assumed that IPv4 or dual-stack configuration is permissible if the IPv4 network is local to the system and does not leave the LAN and the vendor does not offer an IPv6-only product line for the required system (A8).

By applying these rules against the list of products, the Foundation has obtained a matrix of original components, IPv6 capability (ready/only/both) and limitations, and replacements for non-compliant products. This has resulted in a number of product changes. Products in red have no viable replacement.

Vendor	Product	IPv6 Ready	IPv6 Only	Notes
VMware	VCSA	N	N	Replace with vCenter on Windows
VMware	Update Manager	Y	N	Uses IPv4 for remediation of virtual machines and virtual appliances, would have to leave connected network. (link)
VMware	Auto Deploy	N	N	PXE boot infra uses IPv4. (link)
VMware	vCOps	N	N	No explicit support statement (A9).
VMware	Log Insight	N	N	Confirmed via VMware Engineer. Replace with rsyslogd via NFS, for direct access if virtualization system is unavailable.
VMware	vFabric Hyperic	N	N	Merged into vCOps which is not IPv6 capable (above).
VMware	vFabric Application Director	N	N	Merged into vCAC which is not IPv6 capable (no explicit support statement, numerous community reports that it is not supported). Replaced with vCO.
PernixData	FVP	N	N	Contacted surviving PernixData SE for verification
Atlassian	Confluence	N	N	Outstanding feature request

After replacing products with IPv6-capable equivalents and removing those without equivalents, the system is left with the following logical components:

Vendor	Product	IPv6 Ready	Notes

THE FOUNDATION

Cisco	UCS C-Series	Y	C220 M3, 1U rack servers have a much smaller footprint than 8U B-Series chassis and reduce single points of failure.
Cisco	Nexus 5672UP	Y	1U, 72 port availability
Cisco	ASA 5555-x	Y	1U replacement for ASA 5585
NetApp	FAS-2252 w/Clustered Data OnTap 8.2	Partial	2U replacement for 3250s. Cluster and intercluster LIFS only support IPv4. Management uses IPv6 and storage uses FC.
VMware	Windows vCenter	Y	Requires additional config for vCenter Inventory Service.
VMware	vSphere (ESXi)	Y	Native IPv6 support.
VMware	vCO	Y	Native IPv6 support since v5.1. Replaces vFabric Application Director for automation.
CentOS	CentOS 6.5	Y	Rsyslogd server, other VMs
PuppetLabs	Puppet 3.6.2	Y	Puppet master will run on CentOS, clients run on Linux/Win VMs
Microsoft	Windows Server 2012R2 Data Center	Y	Allows unlimited VMs and use of previous OS versions. Active Directory domain controllers, vCenter server, etc.

General use of IPv6 requires working DNS and all components should reference FQDNs instead of addresses for all configuration. Many component's IPv6 compatibility is based on using FQDNs. Use of IPv6 addresses causes failures, especially with various vSphere components that register with vCenter.

Anacreon's WAN will provide the IPv6 prefixes to the manufacturing system. A general addressing scheme is proposed that includes the location '310b' (3rd rock from the sun, 1st moon, Team Beta) and the VLAN ID in decimal format. The use of decimal format leaves a large number of unused networks. These could later be claimed if necessary starting with 5000. Documentation will use the specified documentation prefix of 2001:DB8::/32, to be replaced with the correct Anacreon-provided prefix:

VLAN ID	Prefix	Description
-	2001:db8::/32	Documentation
1	2001:db8:310b:1::/64	VLAN 1
50	2001:db8:310b:50::/64	VLAN 50
4094	2001:db8:310b:4094::/64	VLAN 4096
-	2001:db8:310b:5000::/64	Future special-purpose network

All firewall interfaces will have a suffix of ::d00d, which is easily remembered by administrators (ex: 2001:db8:310b:50::d00d). DNS (usually Domain Controllers) will use the suffix ::53. Other static addresses will be noted in the appropriate design sections. Unless otherwise specified, all other hosts will use Stateless Address Autoconfiguration (SLAAC) for simple configuration. Hosts will register their address with ADS DNS. Non-SLAAC devices will be manually registered in ADS DNS. This ensures that FQDNs are resolvable at all times.

THE FOUNDATION

Hypervisor Design

ESXi v5.5 will be installed on rack mount servers. The hosts will have 24 physical CPUs, 128 GB of RAM, a 4 GB USB drive for hypervisor install, and access to shared storage. Management and VM traffic will be carried on redundant 1-Gigabit Ethernet interfaces. Storage and vMotion will be carried on redundant 10-Gigabit Ethernet interfaces. FCoE is the primary shared storage type. iSCSI is available for future workload needs, but only in the case of FCoE failure at iSCSI IPv6 support is experimental. Each host will be joined to a cluster and managed from vCenter. Local access is for last resort in case of emergencies.

USB drives will be created by administrators following KB2004784. If modified images are required, PowerCLI Image Builder cmdlets can generate new ISOs from an offline repository. Auto Deploy's reliance on IPv4 PXE prevents its use as a deployment method (C3).

All ESXi hosts will be joined to the system's Active Directory domain for proper DNS name resolution.

vSphere Management Layer

vCenter v5.5 with Enterprise Plus will provide centralized management of the ESXi hosts, VMs, and features. vCenter will be installed in a Windows Server 2012R2 VM to ensure availability via vSphere HA. The Windows version is required as VCSA does not support IPv6.

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.

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). Host remediation is IPv6 compatible, where VM/vAppliance remediation requires IPv4 that would leave the connected network. As no host patches are likely to be forthcoming, VUM is not useful for the design. If contact is re-established with the west coast and host updates are forthcoming this decision may be revisited.

vCenter Orchestrator (vCO) will provide automation for the system. This replaces vFabric Application Director and its replacement vCloud Automation Center, neither of which are IPv6 compliant. This change removes the self-service catalog and "X as a Service" functionality while retaining the automated workflows, REST APIs, and web views. Integration with other services, such as Puppet and Active Directory, is possible but requires more development effort.

VMware licenses with no expiration date have been acquired for all vSphere products listed here. Additional licenses are available via the Foundation network as required (A2, A3).

THE FOUNDATION

Server Hardware

The existing design called for large UCS B-Series chassis and blade servers. The minimum footprint for a single chassis with FIs is 8U, and 16U with redundant chassis. This is over the half the allocated 21U and allows no room for growth (3x = 24U > 21U).

To reduce the physical footprint, Cisco UCS C-Series rack mount systems have been chosen as the replacement server hardware. The UCS platform is popular, well known, and there exists copious training material (such as #vBrownBag videos), making it an ideal choice for the support teams to use and maintain. Rack systems maintain a smaller physical footprint and allow more granular flexibility to scale up and down within the space constraints. If a single system fails, service levels gracefully degrade on the remaining systems.

Based on hypervisor design specifications, the UCS C220 M3 server has been chosen. The system is configured with:

- 2x VIC-1225 (total pNICs: 2x1GB, 4x10GB, 1xOOB)
- 2x E5-2697 CPUs (2x12 core, 30MB cache)
- 512MB RAM (full)
- No storage
- 4GB USB Bootable Drive

The 1-Gigabit Ethernet interfaces are used for management and VM traffic and the 10-Gigabit Ethernet interfaces for FCoE and vMotion. All storage will be shared. The system will begin with 4 servers. The number can increase or decrease, and if stockpiles of C220s run low, any other UCS C-Series hardware the Foundation acquires can be inserted with a minimum of changes to the overall system. The hardware will be connected to the network as described in the Networking Configuration section.

THE FOUNDATION

Networking Configuration

Cisco Nexus switches work well with the Cisco UCS series and are offer a large number of Gigabit Ethernet ports, FCoE capability, and generous internal bandwidth. Like Cisco's UCS product line, the popularity of the Nexus switches ensures the systems are well known and training material exists for those unfamiliar with the product.

The network core will be a pair of Nexus 5672UP switches. The switches are capable for 72 10-Gigabit Ethernet interfaces with QSFP breakout cables, forty 1- and 10-Gigabit Ethernet interfaces, three fan modules and two power-supplies, providing redundancy within each unit. Each compute and storage 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 Cisco's best practices guidelines to ensure efficient performance.

The models were chosen to provide room for growth or component failure. The number of available ports are in excess of what the full 21U is designed to contain.

The system will include two ASA 5555-X firewalls (see Security Architecture) to connect to the Anacreon network, for instance to interface with the ADS Forest (C4) and access the Earth-Moon links.

THE FOUNDATION

Shared Storage Configuration

Shared storage is required to allow VM workloads to migrate from ESXi host to ESXi host without impacting performance levels.

NetApp is another commonly known vendor that personnel may already know or can learn quickly. Two FAS-2552s will be configured to replicate to each other for data redundancy and availability. Each unit will be fully loaded with 24 x 1.2TB SAS 10k drives for a total raw capacity of (24 * 1.2TB * 2) 57.6TB. Clustered OnTap 8.2 will be used to connect the two units and present the aggregate storage via Storage Virtual Machines (SVMs, vservers).

Three vservers will be created, one each for admin (management or mgmt. in older documentation), SAN, and NAS access. The SAN vserver will present storage for vSphere access and be managed by the support team. The NAS SVM will present storage for file sharing and will also be managed by the support team initially, but the segregation allows developers or any other group to administer the NAS vserver storage without affecting the SAN vserver. The ESXi hosts will use the SAN vserver storage as primary shared storage. The NAS vserver will store a copy of the *Encyclopedia Galactica* (C5) that all colonists may access.

Four 16GB FC and four 10GB iSCSI interfaces provide ample bandwidth and high redundancy/availability between the NetApp and the network. Shared storage will use Fibre Channel.

The requirement of IPv6 presents a complication for the NetApp storage. Clustered Data OnTap cannot be run in IPv6-only mode and require IPv4 use for cluster and intercluster LIFs. OnTap will be managed by IPv6. Only the cluster communication will be done with IPv4. The VLAN 4000 and network 192.168.4.0/28 will be used for this (12 remaining addresses allow for expansion). This contains the IPv4 service to the connected network (A8). The alternative of running non-clustered negatively impacts availability and manageability and should not be considered unless A8 is deemed invalid.

THE FOUNDATION

VM Design

Initial system VMs are described here. See the vCO section for additional VMs that may be provisioned during operation.

Microsoft Windows Server 2012R2 Datacenter Edition licenses have been acquired for an unlimited number of sockets (A2) and does not require activation (A3). Currently, 8 sockets are present in the system. Windows licensing allows the installation of 2012R2 or any previous Windows Server edition. CentOS 6.5 will be used for Linux VMs. All workloads are supported on these two platforms, which will be used throughout the design.

The previous design was of a significantly larger scale, over 1000% the physical size. The resource allocations for this smaller design are therefore estimations based on 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. Manufacturing encompasses the manufacturing application and its 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 (*anacreon.luna*). A new domain tree will be created (*beta.anacreon.luna)* and two Windows 2012R2 VMs will be provisioned as domain controllers for this tree. Each DC will be in its own network (ex: 2001:db8:310b:1001::53/64 and 2001:db8:310b:1002::53/64). Windows 2012R2 Datacenter licenses have been acquired and all additional Windows VMs will also run 2012R2 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

The inability to use Hyperic with IPv6 simplifies the vSphere systems supported. Only two additional vCO appliances are required.

vCO will provide the automation and orchestration of the private cloud. Unlike Hyperic Application Director, users do not have access to a self-service catalog or portal. Workflows will be developed by the vSphere administrators and development teams and users will be given access via web views.

No backup utility from the selected vendors provides IPv6 support (per A9). As the 21U system offers no true offsite backups, this is mostly relevant to data integrity and revision, as a loss of data in the manufacturing system would not be able to be restored. Still, these features are vital to the system. NetApp system snapshots will provide limited integrity and revision.

VMware's VDP(A) products should be tested for actual IPv6 support and if successful, should be integrated to the design. Other vendor products may be investigated, though do not fit the current system constraints (C2). This design decision should be re-visited monthly until a product with IPv6 support is acquired or IPv4 networking is approved for backups.

This table shows all vSphere systems and appliances and their initial resource allocations and quantities. VDPA has been included in red to show the estimated impact on provisioning the system.

Service	vCPUs	RAM (GB)	System disk (GB)	Data disk (GB)	Quantity
vCO Appliance	2	3	7	0	2
VDPA	4	4	3100	0	2

The original design calls for scripting rather than automation. Given the reduced staffing, scripting is being deprecated where possible to improve automation (DB2). CentOS VMs will provide this functionality. The vCO workflows can be complemented by the use of Kickstart, Puppet Master, Gitolite, and Jenkins CI VMs (see Puppet System) as initiators or targets of workflows. 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 was 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

Manufacturing System

The manufacturing system installed at Anacreon uses the same 3-tier architecture (nginx, RestMQ, MongoDB) as the Terminus City system (REQ1) and interfaces with SCADA/HMI systems, also developed and tested for IPv6 compatibility by the Foundation (A11). At Anacreon, the manufacturing system will be scaled down from starship construction to facilities and material goods construction.

THE FOUNDATION

Though the scale of construction is much smaller, the closed ecosystem of the moon increases the value of the manufacturing system. As the underground chamber Anacreon is housed in expands, excess excavation material and raw materials from Earth will be converted to finished goods. The system will rely on automation to maintain high accuracy and efficiency and keep manufacturing running during sleep cycles.

In front of the three tier system will be 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 manufacturing process relies on continuous integration and continuous deployment processes to improve the system outputs and correct errors through rapid code deployments. In order to ensure these rapid deployments do not have an adverse effect on the manufacturing process, Development adheres to a strict change control process that involves three environments: Development, QA, and Production (DB3). 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 manufacturing 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

Physical plant staffing and maintenance is beyond the scope of this document (A7).

Summary

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

vCPUs	RAM (GB)	Disk (GB)	VM Quantity
78	304	13044	41

vCO Workflow

The vCO Appliances are installed according to the VM Design section. The interfaces are the vCO client (admins only), REST APIs and Web Views. Developers will create workflows to provision/delete manufacturing components. vSphere admins will create workflows to provision/delete management components. Each group will assist the other and provide oversight as needed. vCO workflows will manipulate Puppet to determine and apply the correct application load out for each VM (see Puppet System. The Watchdog VMs will call the workflows via REST for scaling.

Management VMs will be restricted to vSphere admins only and can be deployed in any environment/network. Manufacturing VMs are restricted to developers and can be deployed in any of the manufacturing environments/networks (initially Development, QA, Production).

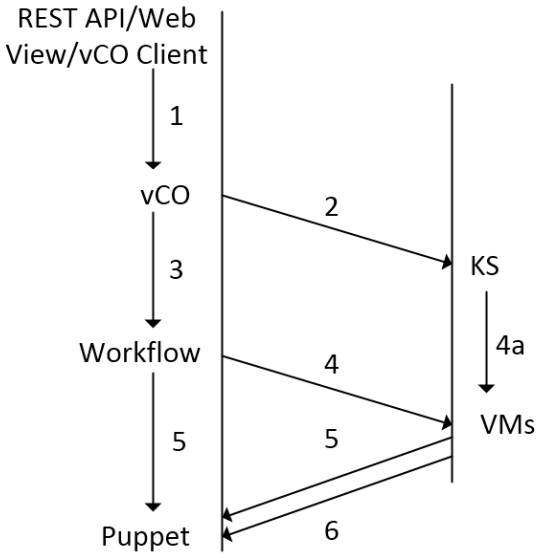
vCO will rely upon vCenter templates (Windows), kickstart processes (CentOS) and Puppet services to provision new VMs. Other objects will rely on various service specific to the object type, such as ADS for user services.

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 CentOS 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 normal provisioning requests (Windows or Linux) will be made through vCO. When the user selects a VM via the catalog (see vCO Portal for more information), vCO will initiate a workflow to provision the VM.



1. User initiates a workflow or automated systems submit a REST API query for a workflow and provide the required information to fulfill the request.
2. vCO creates the kickstart file for provisioning (Linux only) and deposits it on the Kickstart VM.
3. vCO initiates the VM provisioning workflow
4. The workflow builds the VM from template (Windows) or creates an new VM of the correct size (Linux)
    a. (Linux) The VM receives a kickstart file and downloads/installs the appropriate files.
5. The workflow forces the VM to register with puppet and signs the certificate (ref).
6. The vCO workflow forces the new VM to check in to Puppet and receive its initial setup.
7. vCO repeats steps 2-6 for any additional VMs required by the selected workflow.
8. vCO notifies the requestor the workflow has been completed.

THE FOUNDATION

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

The Kickstart server provides CentOS system contents for newly provisioned VMs, plus the kickstart file for the new VM. Each CentOS 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 vCO workflows 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 vCO workflows require that the environment information (VLAN, folder structure, etc.) be provided as input in the client or API call, and 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 considered important repositories. Jenkins CI is a Continuous Integration tool that is used to test changes to the manufacturing 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 manufacturing code, further validating code before deploying to any environment and assisting in maintaining the 99.99% SLA.

Due to the lack of IPv6 backup products, scheduled vCO workflows will back up Gitolite, Jenkins CI, and any RDP data that is significant to a dedicated LUN/datastore on the NetApp SAN vserver.

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 use of Dev->Production promotion ensures that all changes, even those rapidly required to adapt to evolving 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 (REQ10).

THE FOUNDATION

VMware Datacenter Design

The Foundation's vCenter server will define one datacenter for Anacreon Manufacturing Beta. A single cluster of 4 UCS ESXi hosts will be provisioned immediately. If additional ESXi host are required, the cluster can grow to 32 hosts, beyond the physical constraints of the space. The cluster cannot scale downward from 4 hosts without risking availability and the SLA (REQ10).

To meet the SLA, the cluster(s) will be configured with High Availability (HA) and Distributed Resource Scheduling (DRS). Due to the homogenous hardware stores acquired by the Foundation, Enhanced vMotion Capability (EVC) is not required at this time.

EVC cannot be changed to a lower CPU compatible mode on a running cluster and the Foundation is more likely to acquire older CPUs than newer, given humanity's current state, and EVC settings would need to be set very conservatively to future-proof the cluster against this likelihood. If this need did not arise, cluster performance and manufacturing capabilities would be impaired without justification. Both alternatives introduce risk. The probability that EVC is required is judged to be sufficiently low that the risk to manufacturing output is higher. If future iterations of the design require EVC, risk can be mitigated and manufacturing output conserved providing the current homogenous system to a colony ship and implementing a replacement heterogeneous system in the manufacturing facility.

HA will have an initial admission control policy of 25% of cluster resources to provide for 1 host failure (1/4 * 100) and will be revisited every 30 days as manufacturing capacity increases and if cluster size changes. 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 manufacturing system scales itself. DRS rules will help ensure availability of management VMs with multiple instances.

THE FOUNDATION

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	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 Policy - High	Automation - vCO App, Gitolite VM
VM Override VM Restart Policy - High	Manufacturing - Database and Watchdog VMs
VM Override VM Restart Policy - Low	Web Front End VMs

THE FOUNDATION

Security Architecture

The security of the manufacturing security is extremely vital. Any security compromises, accidental or purposeful, risk the Anacreon facility. 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.***

VMware vCloud Networking and Security does not have IPv6 support, forcing the system to rely on in-VM protection. 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).

The system's internal routing and edge, between the manufacturing network and Anacreon's LAN/WAN, will be protected with dual ASA 5555-X firewalls in an active/passive failover configuration. The ASAs are licensed for multi-context but will operate in single-context mode initially. Interfaces will correlate with VLANs, such that the firewall's address in any network will include the VLAN and the suffix ::d00d. I.e. in VLAN 2043, the address would be 2001:db8:310b:2043::d00d. Secondary addresses (passive firewall) end in ::d00f and are unused except for in-band emergency access.

Initial policy will allow unrestricted outbound common services and more restricted inbound services according to the table below.

Manufacturing to Anacreon LAN/WAN			
SRC	DST	SERVICE	ACTION
Internal Networks	External Networks	http, https, ssh, smb, dns	Permit
Anacreon LAN/WAN to Manufacturing			
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

Initial policies surrounding the manufacturing networks and environments are relatively simple and mostly ensure proper ingress/egress controls and wide network controls, as the host-layer firewall will assist with more granular controls. For instance, the following policy would protect the web/middleware edge:

Web Dev to Messaging Dev			
SRC	DST	SERVICE	ACTION
Web Dev Addresses	Messaging Dev Addresses	http	Permit

Messaging Dev to Web Dev			
SRC	DST	SERVICE	ACTION
Messaging Dev Addresses	Web Dev Addresses	http, https	PERMIT

Accurate policy definitions will be created by the developers as the manufacturing system is implemented. Python will be used to configure the ASA policy, generated by the Watchdog VMs.

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 support team lead and mission control in Terminus City 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 support team lead and mission control in Terminus City in case local DCUI/shell access is required.

Monitoring and Capacity Planning

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

vCenter Operations Management does not have IPv6 support (A9) and eliminates it from contention for both monitoring and capacity planning. vCenter is IPv6 capable and includes basic monitoring functions.

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 manufacturing plant's on-shift manager and a designated military liaison. Notification to three parties will ensure that alert responses are timely. After workloads are in place for five 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.

Syslog and snmp traps for all devices will be configured to go to a pair of CentOS VMs. The syslog server will run rsyslog, which supports IPv6 as a target in the format "@hostname.beta.anacreon.luna". SNMP traps will be collected with snmptrapd, which is a valid target in the format " trap2sink udp6:[2001:db8:310b:2099::1]:162". Commercial syslog/snmp analyzers are not available (C2) for further analysis, but the raw data is available for administrators to access for troubleshooting.

THE FOUNDATION

Configuration Maximums

The manufacturing system fits within the proscribed physical footprint with room to spare (C1):

Vendor	Product	U	Qty	Agg. U
Cisco	C220 M3	1	4	4
Cisco	Cisco Nexus 5672UP	1	2	2
Cisco	ASA 5555-X	1	2	2
NetApp	FAS-2552	2	2	4
		Total U		12

The Anacreon Beta system uses 12U of the 21U of rack space available. In addition, the Nexus switches provide 144 ports, in excess of the (4 * 8 + 2 * 3 + 2 * 4 =) 46 ports required. Each C220 requires 8 ports, easily allowing another 9 hosts and 9U of compute to be provisioned and still include excess capacity. Expansion of compute hosts would result in the following resources:

	4 Hosts (Initial)	13 Hosts (Max)
RAM (TB)	2	6.5
CPU Sockets	8	26
Cores	96	312

The dual NetApp storage systems are loaded with 1.2TB disks and occupy 2U each. Additional storage could be provisioned in a multitude of options. For simplicity and availability, we will consider capacity as additional FAS units are added, up to a total of 6. These numbers could change with the use of storage shelves:

	2 Units (Initial)	6 Units (Max)
Raw Capacity (TB)	57.6	172.8

The system can be expanded for full compute or full storage, but not both. Intelligent decisions based on capacity management must be used to maximize the aggregate resources available to the system to best fit the usage profile. If the usage varies over time, it may be possible to swap compute for storage, and vice versa, as long at least 4 C-series servers and 2 NetApp FAS units are present at all times. Such changes would not be rapid and should only be made to account for long-term changes, not short spikes of compute/storage usage.

THE FOUNDATION

Failure Tolerance

The base configuration described above strives to maximize availability and limit failures. All systems - compute, storage, network, security - have been implemented with a minimum of two devices for availability and redundancy. Initial capacity is below the capacity of a single device failure (75% for compute, 50% for all others). Each system offers its own method for ensuring continuity of service during failures and the ability to degrade with permanent failures.

Compute

The compute nodes run ESXi and have access to shared storage. vSphere HA will restart VMs hosted on a failed node on a responsive node. vCenter DRS will migrate workloads to keep utilization roughly consistent across nodes. The manufacturing Watchdog VMs will spin up additional VMs as required and spin down unutilized VMs.

In the case of a node failure, the remaining nodes are capable of handling the workload. In a long term failure, the system will continue to operate effectively. All DRS rules will still be enforceable. HA admission control should be set to 33% or disabled entirely. A second node failure would cause DRS rules to be violated and affect the long-term effectiveness of the manufacturing system. The system would operate properly, but degraded, if replacement hardware could not be acquired.

Storage

Clustered OnTap will have two members initially. The cluster will be able to migrate storage workloads to meet demand easily. If a FAS node fails, capacity will be cut in half but operational continuity is assured. Failure of a second unit would cause a complete loss of data and shut down the system as there is no local storage for the compute nodes. This is an acceptable risk as local storage prevents vSphere HA and does not enhance data integrity, vs. the low likelihood of a complete storage failure that does not include failure to compute as well. Replacement hardware should be obtained from Terminus City upon the first node failure.

Network

The dual Nexus 5672UP systems provide a great deal of redundancy. Each compute, storage, and security node is connected to both switches and the switches have redundant links between each other. Failure of a single switch will reduce port counts by half but have no operational impact on the system. Failure of the second switch would shut down the system as no node could communicate with another. This is a moderately acceptable risk given the space constraints (more below) and the low likelihood of failure that affects both switches and not the other devices. Replacement hardware should be obtained from Terminus City upon the first node failure.

Security

An active/passive ASA configuration provides redundancy to the security devices. Unlike an active/active configuration, there is no chance of pushing the combined load past what a single device could handle. During a node failure, there would be no operational impact, state would be picked up by the passive (now active) unit and sessions would continue to flow uninterrupted. Failure of the second node would have severe impact. As the ASAs protect the VLANs inside the edge as well as the edge, all traffic flow would be restricted to within each VLAN.

The risk could be shifted by using another firewall. Both Windows and Linux include firewall services that are capable, but were designed to be host firewalls, not network firewalls, and are not designed for proper manageability in such a situation. Popular free Linux firewalls include pfsense and untangle. These services could be installed in a VM, which would benefit from HA/DRS, or a compute node could be re-provisioned as a bare-metal OS plus firewall, which would benefit from direct hardware access and increased resources. Neither option are as flexible as the ASA but are reasonable compromises if replacement hardware could not be acquired.

Other Considerations

A significant risk to the design is the physical constraint (C1) itself. By placing all components in a single half rack, there is one single point of failure - the rack itself. Numerous accidents could occur that affect all components at once - a construction accident electrocutes/burns/smashes the rack, a colonists suffering from cabin fever takes an axe to it, pressure loss causes all systems to overheat, etc. - and these incidents cannot be predicted or mitigated. It would be extremely valuable to research acquiring an additional physical rack in a different location to spread risk out. Even two 10U locations with physical separations would be beneficial.

While whole-rack failures cannot be mitigated, some physical segregation of devices inside the rack could help mitigate certain failures. For instance, an overheating component could damage the device above and below it. A layout alternating devices types and use of empty space minimizes this impact.

ASA
-
Nexus
-
FAS
-
Compute
-
Compute
-
Compute
-
Compute
-
FAS
-
Nexus
-
ASA