VCAP-DCD

8/14/2011

Module 2: Design Process

Design Methodology: An iterative, high level process used to create a design:

4x Phases in vSphere Design Methodology:

1. Architectural Vision
2. Architectural Analysis
3. Technology Architecture
4. Migration Planning

① Architectural Vision: A high level vision for the project. Includes the following aspects

(1) Scope - Identify scope in detail!
- defined scope prevents unintended expansion. ie. this project includes the united states production servers only. No dev/test and no Canada servers.
- prevents the need to renegotiate cost of a project or work for free.

(2) Goals - set specific goals for a project.
- goals need to be specific, measurable, and actionable.
- without goals it is difficult to determine success and value of a project.
- ie. the organization wants to achieve a 50% reduction in production server equipment by the end of the year.

(3) Requirements: Identify key business and technical requirements for the project.
- example: SOX compliance, physical separation from production + dev/test, uptime requirements, etc.

(4) Assumptions: Design Components that are assumed valid without proof.
- ie. the organization has sufficient bandwidth between sites for replication

(5) Constraints: Constraints limit design choices - could be a policy, process, or technical constraint
- ie. Due to existing relationships, all hardware is DELL.

(6) Risks: Identify Risks that might prevent achieving project goals.
- ie. lack of core redundancy introduces risk of 99.99% uptime.
- discussing risks eliminates surprises.

VCAP-DCD

8/14/2011

Module 2: Design Process Continued..

② Architectural Analysis: Deeper dive into current infrastructure.

Perform 2x Activities:

① Current State Analysis:

- gather and analyze info on current servers, storage, network,
- gather and analyze applications and operating systems.
- gather resource consumption data.

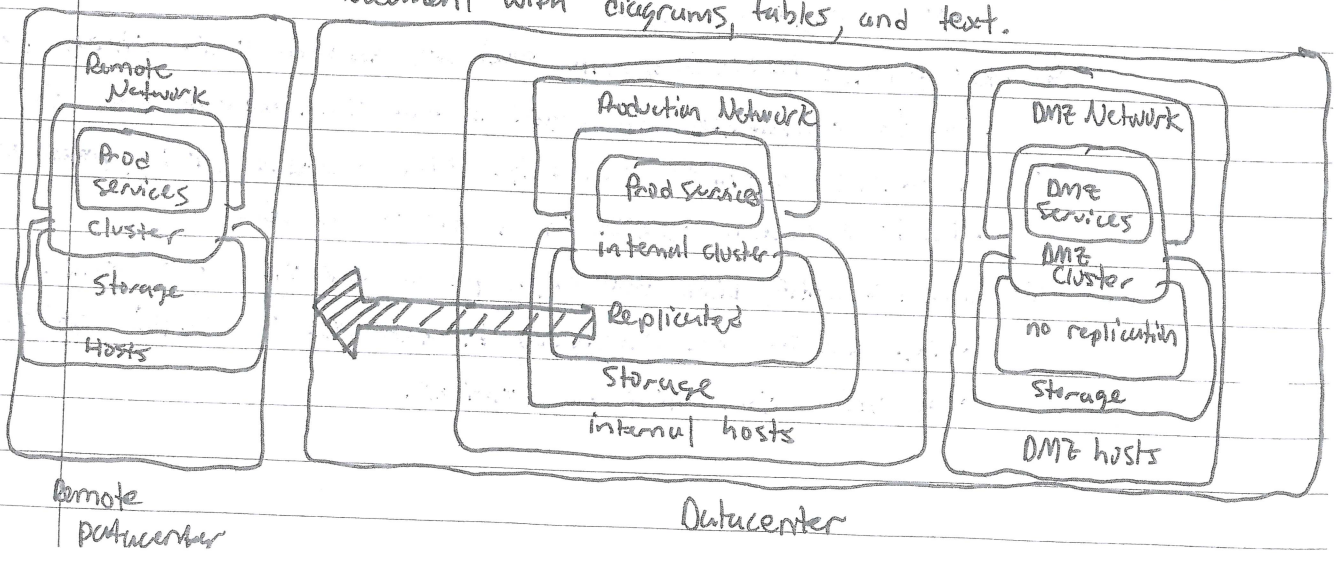
② ~~Define~~ Envision. Target state required to meet organizational goals while factoring in requirements, assumptions, and constraints.

③ Technology Architecture: Develop 3x designs:

① Conceptual Design: focus on Achieving goals + requirements.

- use info from SME/Stakeholder interviews (scope/goals/requirements/assumptions/constraints) and from info gathered in current state analysis
- Determine entities affected by project. (LOB, users, applications, process, physical machines, etc.).
- Determine how goals map to each entity.
- Design infrastructure that achieves each entity's goals + requirements, but stay within constraints. ie.. where do you need availability, scalability, performance, security, manageability, etc.
- Document with diagrams, tables, and text.

Example Conceptual design diagram.



VCAP-DCD

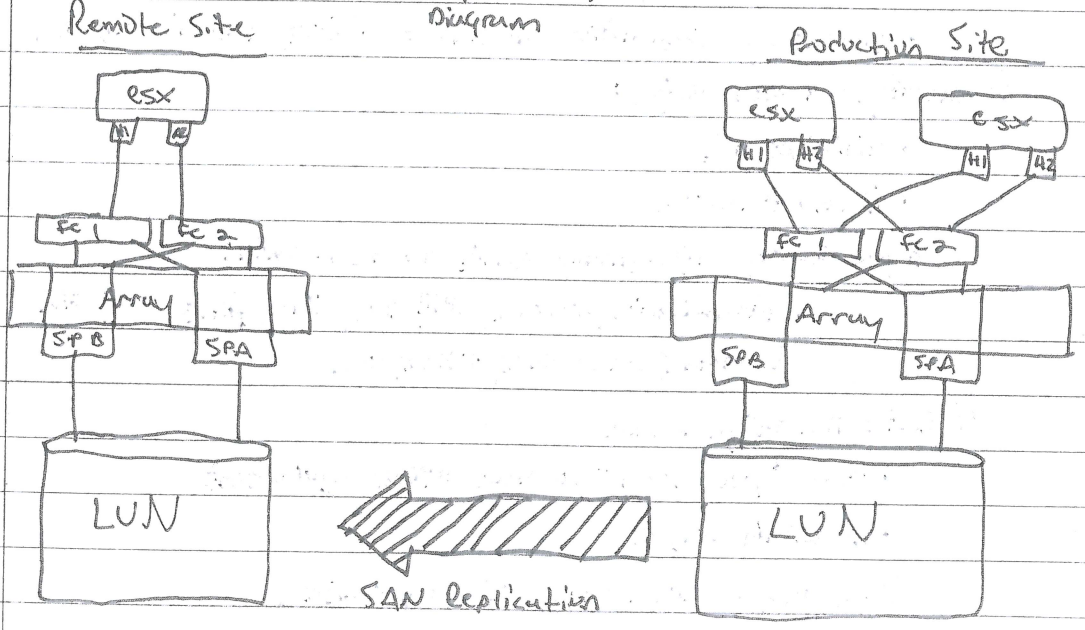
8/14/2011

Module 2: Design Process Continued...

② Logical Design: Design includes relationships between all major components of the infrastructure. Considers the conceptual design, constraints, and risks.

- Useful for understanding and evaluating the design of the infrastructure:
  - does it meet the Requirements but stay within constraints?
  - Does NOT include physical details like port assignments, hardware vendors, IPs,
- illustrate how to arrange infrastructure components
- dont get lost in configuration details.
- be aware of capacity analysis, but dont include things like LUN sizing, CPUs, etc.
- document in diagrams, tables, and text.

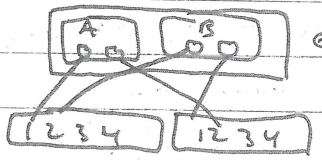
Example Logical Design Diagram



③ Physical Design: Use the logical design + include specific hardware details and implementation information.

- includes port assignments, pci slots, etc.

Example Physical design diagram



VCAP-DCD

8/14/2011

Design criteria

① Usability
1. Performance - may be measured by throughput, latency, transactions, etc.
2. Availability - defined as access to resources when needed.
achieved using redundancy
defined in SLA.
3. Scalability - Support future growth while maintaining acceptable performance.

② Manageability
- Easy to deploy.
- Easy to administer + maintain
- Easy to update + upgrade.

** Simplification is key. **
- unnecessary complexities can lead to failures, rises in costs, etc.

③ Security. - a good design:
- minimizes risks
- is easy to secure.

* good security design typically applies defence in depth:
- authentication, firewalls, DMZ, IPS, IDS, filters, VPN, etc.
biometrics, smart cards, tokens, etc.

④ Cost: the design needs:
- to be "good enough" to meet business requirements
- fit within budget.

** All designs are a series of compromises.

Design Approach: Sound Approaches include:
① USE A METHODOLOGY!
② Conduct interviews with SME's and knowledgeable stakeholders.
↳ discuss design alternatives + effects on performance, availability, scalability, manageability, compliance, etc.
③ Thoroughly DOCUMENT all aspects of a design. Two engineers should be able to build the same infrastructure from the design

VCAP-DCD

5/14/2011

End Goals: Good Designs

1. Involve SME's + Stakeholders.
2. Balances business requirements with technical considerations and best practices.
3. is not unnecessarily complex
4. is REPEATABLE
5. is not ambiguous.
6. has documented rationales + considerations.

- ✗ Clearly document all design decisions.
- ✗ Always involve SME's + stakeholders
- ✗ Create design documents like construction blueprints.
- ✗ Simple designs are easier to understand, perform better, cost less to manage, etc.

VCAP-DCD

8/20/2014

Module 3: ESX/ESXi Host Design.

CPU Capacity:

- Determine Requirements by using Capacity Planner, OS and app vendor guidelines/documentation, talking to SME + Stakeholders.
- Add needed growth to current usage for cpu capacity.
- DO NOT Plan to fully utilize CPU.
  - ↳ Plan for avg utilization of 60-70%
    - Leave headroom for → Short term util spikes. (business cycle
      - → Patching/Maintenance
      - → VMware HA.
      - → Datacenter failover
      - → VMK + SC Overhead.
      - → Growth.

Number of Hosts: CPU Capacity.

- No matter what, host config must meet CPU Req.
- How many CPUs per host?
  - → Cost of dense vs non-dense servers?
    - ↳ Consider not just the initial purchase, but the operational costs over time:
      - ie. Power, Cooling, man hour overhead for maintenance, service contracts, etc, licensing
    - ↳ Can enough memory be installed to keep the CPUs busy
    - ↳ Are there enough storage + network ports available to provide bandwidth to support VMs that could run on a dense host
  - ★ Consider VMware HA and the # of VMs per host limits. Isolation policies may drive more smaller hosts.

READ!!

★ CPU Scheduler Paper ★

Cores per CPU:

- ★ The # of cores per host must match or exceed the # of vCPUs in the largest VSMP VM.

VCAP-DCD

8/20/2011

Module 3 → ESX/ESXi Host Design.

# of vCPUs per Core:

Check vCPU per core maximum.

☆ VM Performance drives this limit ☆

Goal is to efficiently use CPU capacity without reducing performance

Maximum limit: Config max:

effective limit: based on effective acceptable performance.

More cores = more sharing of memory & data buses.

✗ 4-6 w/ dual core processors

✗ 3-5 w/ Quad core processors. ← lower bc of bus sharing.

1-2 vCPUs higher w/ VDI.

Features (CPU)

Buy the fastest CPUs and the most cache you can afford.

64-bit

Intel VT-x + EPT support

AMD-V + RVI. (Rapid Virtualization Index)

Licensing:

Must support # of sockets

Must support # of cores per pCPU.

Memory Capacity:

Same As CPU. See CPU.

DO NOT Plan to fully utilize memory Resources.

70-90% util.

Leave headroom for util, maintenance, HA, ← Same as CPU.

VCAP-DCD

8/20/11

Module 3 Continued...

Memory Overcommitment:

2 types: .total → does not affect performance
Active. → DOES

AVOID ← Add more memory to design
Reduce consolidation Ratio

Active Memory = Active Guest Memory.

↖ Amount of memory in use by guest OS and Applications.

if active memory is oversubscribed:

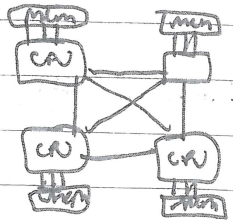
① Add memory to each host.
② Reduce # of VMs on each host.
③ Reduce vRAM allocation.

With large memory pages → Reduce memory overcommit.
Large pages are not shared.

Service Console Memory: (ESX only)

- if the host has more than 4 GB of memory, configure the SC with 800MB of memory
- Always configure MAX SC Allocation. (800MB)
- Configure SWAP w/ (1,600 MB) Partition
  SWAP always double SC Memory config.
  Cost of extra space is less than cost of reinstall to resize.

NUMA - Nonuniform Memory Access.



← NUMA.

- Manual processor affinity may not be manageable by NUMA.
- dense VMs that have more vCPUs than are per single socket cannot be managed

VCAP-DCD

8/20/2011

Module 3 Continued

NUMA

* Evenly Distribute memory among CPUs.
* Disable node interleaving in BIOS

* By default only enabled on hosts with a total of at least 4 CPU cores and with at least 2 CPU cores per NUMA Node.

Node interleaving is like RAID 1 disks. All data sent to physical memory is written across all physical host memory. = data written to both local AND Remote Memory. = BAD PERFORMANCE!

Host Hardware Type:

Blades or Pizza boxes?

* Choice is not always technical. *

Blades:

Good:
- Flexible
- Easy to install + Replace
- Simpler cabling.
  - ↳ fewer IO ports.
- ~~Less power~~ (overall)
- management. (Opex)
- more DC engineering.
- Smaller footprint.

BAD
- locks into vendor (proprietary)
- hot spots
- more power + cooling per rack
- Shared chassis + components could be single POF, or could lead to larger outages.
- not as much expansion.

Design A Consistent Platform.

Same servers - CPU, memory, Ports, PCI Assignments.

↳ Simplifies troubleshooting, planning, automated installation, unifying HA, etc.

VCAP-DCD

8/20/2011

Module 3 Continued.

ESX or ESXi:

ESXi Embedded Wherever possible.

Embedded or Installable

- less attack surface
- easier to patch
- no SC consuming resources

What would force ESX over ESXi?

1. Check Agents.
2. vSphere Web Access not included in ESXi.

BIOS Settings:

Enable VT.

Disable node interleaving.

Disable unused USB ports, serial ports, network ports, etc.

Disable power saving options for Power + CPU ← ESP w/ FT.

PCI Slot Design:

SAME CONFIG!!!

- Consistent HBA Numbering: Simplify troubleshooting, auto install + config, easier administration.

10gb cards should go in PCIe 8x slots to reduce bus limitation

Power + Cooling:

Design for Redundancy:

Check vendor limits (power) per Rack.

VCAP-DCD

8/20/2011

Before placing Hosts into Production:

1. Burn in all equipment.
2. Test memory.
3. Run HW Diags.

* Assign Static IP Addresses.

if IPs change, vCenter needs to know.

Configure hostnames in DNS!!!

Standardize naming conventions!!!

Host Security:

ESXi no SC = more secure.

ESXi has lockdown mode. (prevents remote access as root)

* Use Centralized Management!

→ vCenter

vMA.

PowerCLI.

— Limit local user accounts.

— Limit root access.

Limit installed agents.

install agents in vMA where possible.

VCAP-DCD

8/20/2011

Module 4: vSphere Virtual Datacenter Design.

vCenter w/ 300 hosts or 3000 VMs. Require 64-bit hosts. (200 of 64-bit)
Best Practice 200 hosts or 2000 VMs.

Requires 2CPU.
360 Memory. → increase if SQL runs there.
2 GB Storage.

Host	VMs	CPU	Memory
50 host	250 VMs	2	4
200	2000	4	4
300	3000	4	8

vpxa.cfg → IP of vCenter system on each host.

Virtual or Physical?
Best Practice: Use a virtual machine.
If the key stakeholders insist on a physical machine, list as a risk!

VM	Physical
• Protect with HA or VM HB.	• Protect w/ VM HB only.
• Dynamic Sizing	• Sizing Static
• easy image backups.	• Backups/Restore harder
• allows HW maintenance of vCenter host server.	

Manual Failover: initial setup cheap → operating expense expensive.
Auto failover: vCenter HB.

VCAP-DCD

8/20/2014

Module 4

vCenter Protection

① Manual — initial cheap, long term expensive. complex, high overhead. future downti
② Cluster w/ HB — expensive, but automatic + fast.
③ Backup/Restore — too long RTO for most.
④ VMware HA. — always use.
↳ disable DRS Migration.

MSCS supported in vCenter 4. NOT 4.1.

vCenter becoming Tier 1.

vCenter Database.

4.0 vCenter requires 32-bit DSN. 4.1?
Best Practice: use existing database server. if VM, better.
P2V existing db if archive.

only small infrastructures should colocate vCenter db + vCenter servers.
SQL express 5 hosts 50 VMs.

Highavailability of Database System: use vendor tools if possible. otherwise use VMware HA. ← does not protect db info.

Certificates:

- if concerned about man in the middle attacks from mgmt ~~clients~~ servers, use certs.
- FT and linked mode require vCenter to have certificate checking enabled.

VCAP-DCD

8/21/2011

Module 4: vSphere Virtual Datacenter Design.

Clusters and Resource Pools.

Critical workloads should be protected by VMware HA:

Scale up or scale out?

- Cost of hardware
- Cost of licensing.
- Cost of operational overhead.
  - ↳ maintenance
  - ↳ patching
  - ↳ power, cooling, rackspace (especially in Datacenter)

Scale up clusters	Scale out clusters
• easier to manage fewer hosts, reduces costs.	• host failure affects fewer VMs.
• Host failure affects more virtual machines	• failover is faster.
• HA failover takes longer.	• fewer resources reserved for failover.
• Certain failover policies might reserve more resources. Typically # of Host failures + specify failover host.	• more DRS migration choices • may affect size of virtual machines. • more redestribute Required. • more power + cooling. • higher operational overhead
• need to be mindful of virtual machine per host limits.	• expensive to manage
• DRS has fewer migration choices, so workloads could be less balanced.	• more licensing.

Mixed Clusters:

- if new features in VM hw version 7 are required, do not use mixed clusters.
- if they are not needed, mixed mode clusters are fine.

VCAP-DCD

8/21/2011

Module 4.

Cluster Failover Policies.

Smaller clusters w/ # of hosts failed tolerates or designated failover node policies configured, more resources will be reserved for failover.

ex. 5 hosts equally sized, if 2 hosts tolerated, 40% of Resources in CPU or Reser

↳ To Avoid above situation, configure percentage based policy.

↳ downside is that guarantees cannot be made on restarting all VMs.

explain Risks to SME + Stakeholders.

VMware Recommendation is to protect the workload.

Duncan Epping recommends % based.

Reservations!

- Powered on VM reservations determine the slot size used by Host failures cluster tolerates policy.
- Even a single VM w/ abnormally large reservations will reduce the # of slots available to run VMs, because the slots are bigger.
- Host failures cluster tolerates works best with large clusters where VMs have similarly sized reservation settings.

In clusters w/ VMs that have highly variable Reservation settings: 3 choices:

Best WL Protection: ← • Accept the large slot size + understand fewer VMs can power on.

- Configure das.slotCpuInMhz to manually set CPU slot size
- Configure das.slotMemInMB to manually set Memory slot size.

↳ Risk here is that any VM w/ reservations larger than the slot size may not power on in a failover.

- Configure a failover policy that is not based on slot sizes.

↳ either % based.

↳ Specify failover host policy: → Risk is if host fails while another is in maintenance mode, VMs may not have Resources to start.

VCAP-DCD

8/21/2011

Module 4: vSphere Virtual Datacenter Design

Cluster failover Policies:

Host failures Cluster Tolerates:

- Default.
- Always reserve enough failover capacity to Accomodate host failures while others are offline or in maintenance mode.
  - ↳ does increase costs, however should still be getting cost savings from consolidation
- Always add 1 to the desired host failure cluster tolerates to accomodate host failure during maintenance.

★ Always configure strict Admission control to protect the workload

Only turn it off in extreme circumstances.

make Admission control (strict) a part of change control policies.

↳ powering on unprotected vms should require change control pr

↳ makes this a management choice, instead of an admin choice.

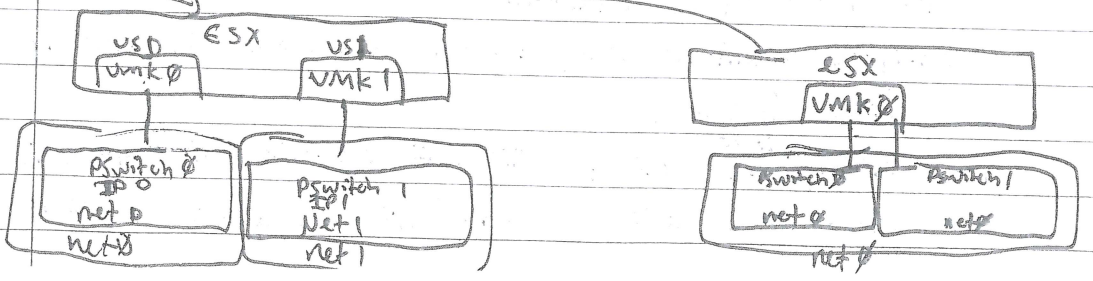
HA Network Redundancy:

★ Always configure it. ★

A switch failure could trigger datacenter wide isolation response.

2 Choices:

- 2x seperate networks, seperate subnets, etc.
- Single mgmt network w/ NIC teaming to multiple switches.



VCAP-DCD

8/21/2011

Module 4:

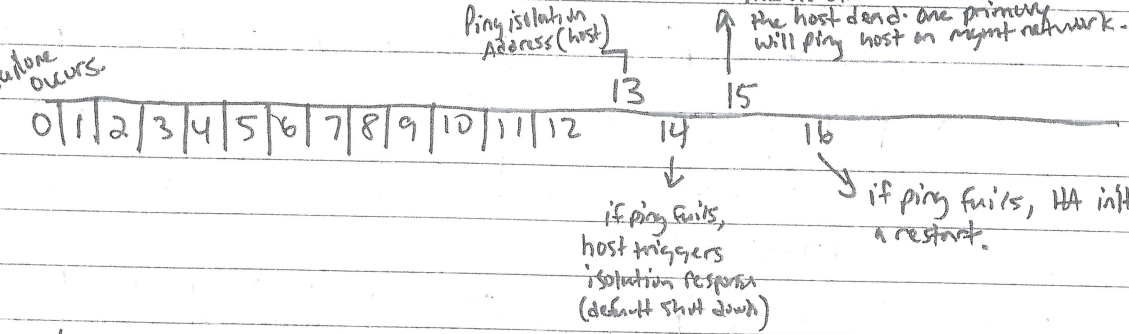
* HA will use all mgmt interfaces to send heartbeats.
  - ↳ in esxi on all vmkernel networks except for those marked as vMotion.

if there are 2x management networks, use das.isolationaddress to add an isolation test address for each additional management network.
- ↳ eliminates isolation test address as a single point of failure.

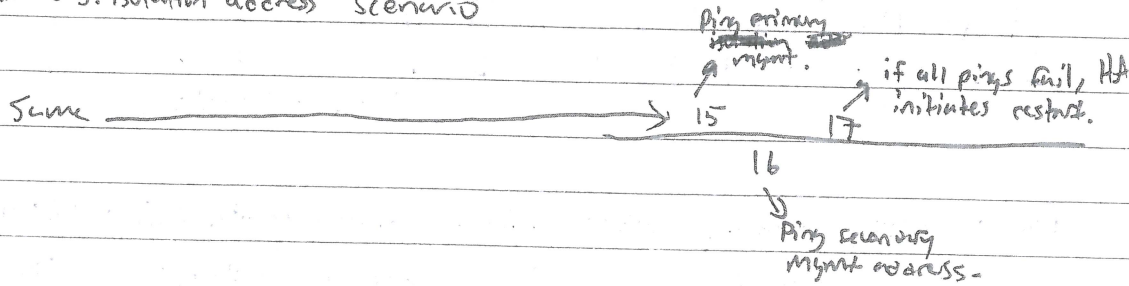
* When you specify an additional test address, increase the setting for das.failuredetectiontime to 20,000 milliseconds or greater.

[redacted]

Single das.isolationaddress Scenario



2x das.isolation address Scenario



2 (vs 3) second time limit where isolation response could initiate a power off at second 14. If the heartbeat comes back at second 16 or before, the VMs will already have been powered down but HA will do nothing because the heartbeat returned.

VCAP-DCD

8/24/2011

Module 4 vSphere Virtual Datacenter Design.

Isolation Response:

Default is shut down VM. - must have VMware tools. 300 second timeout until VMs are powered off.

If the mgmt network has redundancy, shut down VM is appropriate

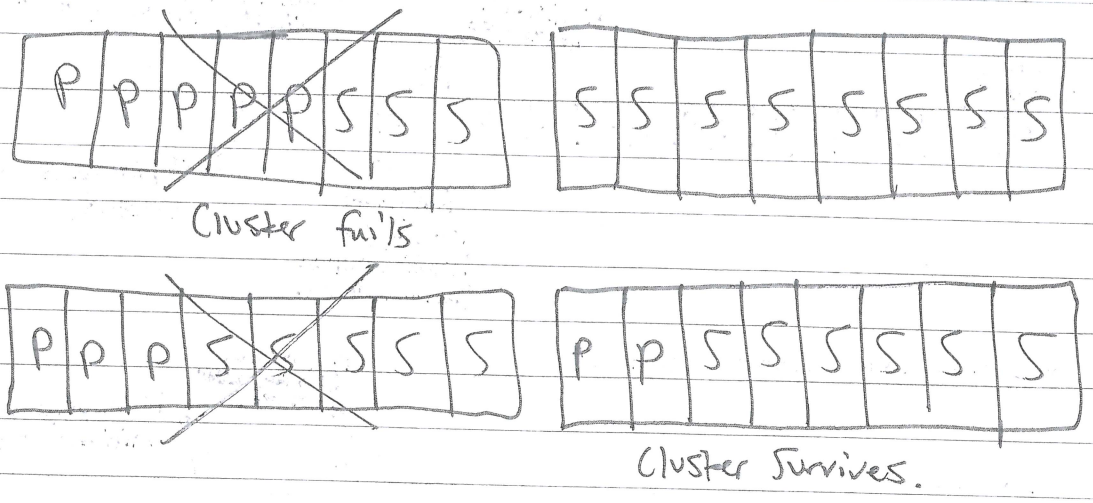
If not, select "leave powered on."

VM Monitoring:

use it.

ft and MSCS will respond faster than VM monitoring.

Blades:



Primary hosts maintain cluster consistency and handle failover events. If a primary is lost, a secondary is promoted to primary as long as there is at least 1 surviving primary.

- never add more than 4 hosts in the same blade enclosure to a single VMware HA cluster. ← guarantees that all primaries will not hit the same chassis.

VCAP-DCD

8/24/2011

Mod 4 → VMware HA Cont...

VMware FT.

- Only single CPU Workloads.
- FT VMs cannot have snapshots, DRS, or Storage vMotion.

☆ if using FT, clusters should not have both ESX/ESXi hosts. One or the other. ☆

FT VMs need to be in clusters with at least 3 hosts.
↳ ensures redundancy if a host fails.

FT must satisfy HA cluster Requirements. ← only in HA Cluster.

Use hosts with the same CPU speeds.

Disable CPU power saving in BIOS. on hosts running FT Workloads

Configure Minimum 1GB ethernet logging.

w/ 1GB enet, no more than 4x VMs protected by FT on that host

10GB enet is better.

Distribute primary VMs among hosts to reduce contention on the logging NIC

→ logging is asymmetric. Primary ⟹ secondary.

Create Redundant logging networks.

An FT VM automatically reserves all of its memory. Configure resource pools ~~memory~~ to have memory above configured memory for the VM for overhead.

VCAP-DCD

8/21/2014

Modules Continued.

DRS Clusters:

- vMotion Requirements must be met.
- Enable enhanced vMotion Compatibility to future proof cluster.
- Configure DRS for full Automation., w/ Default 3 star Rating
- DRS can balance workloads better in scale at scenarios.
  - ↳ gives the algorithms more choices.

Use Affinity + anti-Affinity as the exception, not the norm.

↳ many rules could limit migration choices.
hus negative effect on the scheduler.

affinity can be useful:

- if vms have a lot of network traffic
- if vms share similar memory pages - TPS can be more efficient.

anti-affinity is used to increase availability for service workloa

Multiple Clusters:

Reasons for multiple clusters:

- Cluster may hit host limit (32)
  - ↳ VMs per host limits.
  - ↳ Hosts per LUN limit.
- desktop cluster vs Server cluster.
- Organization requires hardware isolation.
  - ↳ security
  - ↳ dev/test/prod
  - ↳ departmental for cost accounting.

→

VCAP-DCD

8/21/2011

Module 4 Continued---

Cluster Size:

SMBs will typically deploy 1 cluster

Larger orgs with 20+ hosts may choose multiple clusters

↳ physical separation

↳ avoid limits.

In general, medium sized clusters perform better than large clusters ← cluster operations will be faster.

↳ Consider 4x 8 host clusters instead of 2x 16 host clusters because of sluggish vCenter performance.

Avoid Extremes.

Resource Pools.

Deploy resource pools with limit and reservation settings ONLY for virtual machine workloads that require DEDICATED and ISOLATED Resources.

Easier to allocate reservations and limits at the RP level, rather than at the individual VM level.

!!! Resource Pool CPU and Memory share settings do NOT get pushed down to the VM level. If VMs in a specific RP need higher shares, they must be configured manually.

If common VMs are created + need share settings, create a VM template for it.

VCAP-DCD

8/21/2011

Module 4 → vSphere virtual Datacenter Design.

VMware DPM. – Distributed Power Management.

- Enable DPM in environments where workloads vary significantly over time.
- Buy hosts with iLO or. IPMI (intelligent Platform mgmt); (interf
- if hosts don't support above, buy WOL NICs.

- Configure DPM for Automatic. Use the default threshold.
  - ↳ decreases power/cooling costs.
  - ↳ decreases mgmt overhead.

Naming Conventions:

Use Standards.

Pick meaningful names.

VCAP-DCD

8/2#/2011

Module 5: vSphere Network Design.

Design Infrastructure that Reduces costs, boosts performance, improves availability, provides security, and enhances functionality.

Configuring 2x 10Gb NICs could be more cost effective than 12x 1Gbics on separate Physical networks.

High level Design Requirements:

Connectivity, Bandwidth, Latency, Availability, cost.

Build a modular network solution. → helps control future costs.

Components that ~~be~~ should be on separate networks:

- VMs
- VMware FT
- VMware HA
- VMware vMotion
- IP Storage
- Management

2 Reasons to separate Networks:

- Reduce latency & improve performance (esp IP Storage) ← Reduce Contention
- Improve Security → FT, vMotion, etc.

Architect enough Physical NIC ports for bandwidth.

- Use current state analysis (capacity) to determine Requirements.
- Consider failure. Will there still be enough BW left after device failure?

Segmentation:

VLANS or physical Separation?

Depends on # of NICs, # of switchports,

- the preferred choice is VLANs, however some orgs may have policies against them (constraint)

24.

VCAP-DCD

8/24/2011

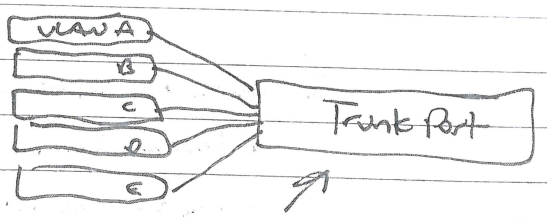
Module 5: vSphere Network Design.

Physical segmentation will likely require bigger hosts with greater I/O expansion. More NICs, more switch ports, more cables, more rack space, etc..

Other factors.

- 1Gb ethernet will not be able to handle as many VLANs than 10Gb.
- Physical infrastructure may not support VLANs.
- Security policies may require physical separation.
- Organization may not currently use VLANs.

VLAN Benefits	VLAN Risks
• Reduced physical ports.	• port failures can be more serious
• reduced cabling.	• easy to oversubscribe & create contention.
• flexibility.	



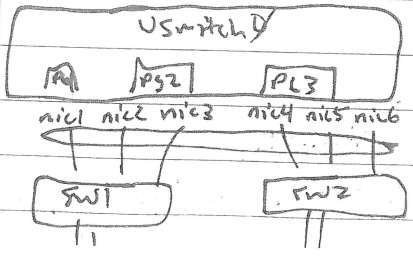
enough bandwidth?

Configure teaming!

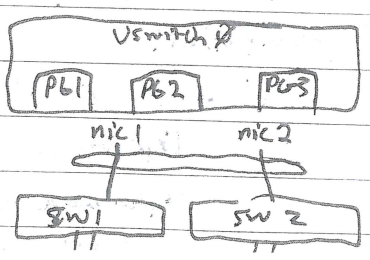
Implementing VLANS:

- 802.1Q switches.
- Configure switch ports as trunks.
- Configure vSwitch trunk ports
- Configure STP PortFast on all ESX facing ports!

1Gb Ethernet



10Gb Ethernet



VCAP-DCD

8/24/2011

Module 5, VSphere Network Design.

Private VLANS (PVLANs)

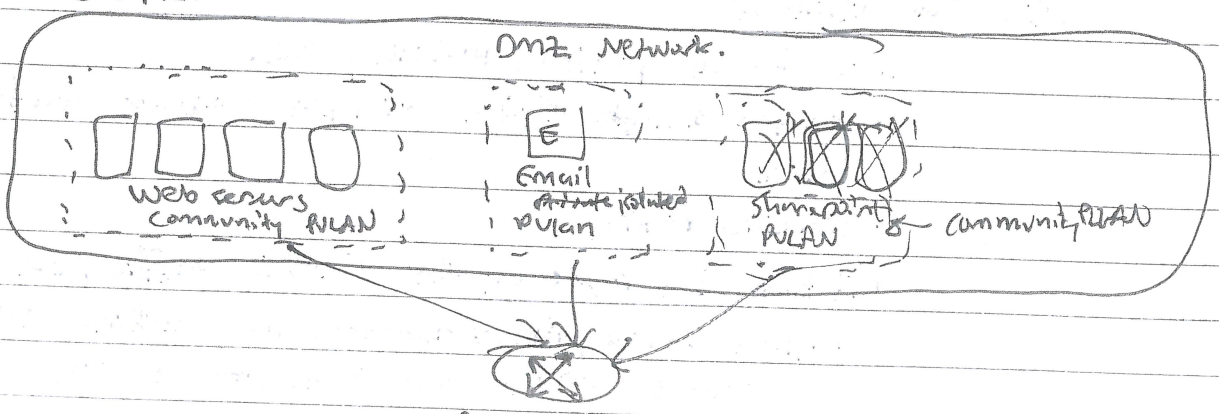
Provide layer 2 isolation between servers on the same layer 2 network without needing to do MAC ACLs.

↖ This can reduce the # of VLANs required and still achieve isolation

- Configure PVLANS to protect the servers from one another if a machine becomes compromised.
- Community PVLANS allow servers in the same community PVLAN to communicate with one another, but are isolated from the rest of the machines on the same layer 2 subnet.
- Isolated PVLANS mean each server is isolated from other servers on the same subnet.

Important for DMZ / public networks.

Example:



VCAP-DCD

8/24/2014

Module 5: vSphere Network Design.

Type of vSwitches:

Distributed vSwitch	Standard vSwitch
• Centralize management. • Consistent configuration. • Offer PVLANs, Load based teaming, Persistent network state, • bi-directional traffic shaping. • vCloud + Lab mgr cross host fencing require dvSwitches. • not manageable when vCenter is down. • Requires ent + licenses.	• required if blending with v3. •

Nexus 1000v

- consistent mgmt interface
- extends virtual network feature set to a level consistent w/ Cisco switches.
- satisfies mgmt requirements by allowing the network team to manage the virtual infrastructure
- Ent + and nexus license.

Q. How Many vSwitches?

A. As few as possible. Preferably one.

- Configure a single vSwitch with port groups for each type of traffic.
- Simplifies config.
- If there is a requirement that VM to VM traffic passes through a physical firewall, the infrastructure needs multiple vSwitches.
  → could use vShield Zones if acceptable by org.

NIC Teaming:

Increases BW, Redundancy.

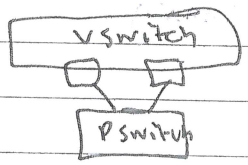
Try to team Across onboard + Riser NICs.

VCAP-DCD

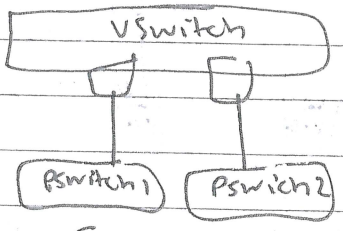
8/24/2011

Mod 5: Network Design.

- Create a single virtual switch with teamed Nics across separate physical switches:



OK.
Source portID
Source MacAddr
IP hash.



Source portID
Source MAC Addr
ip hash (with stacked switches)

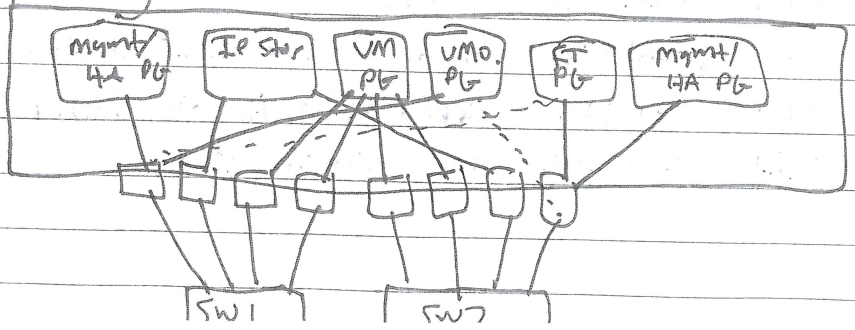
use originating port ID as default, IP hash where hardware supports it.

use Active/Standby port configuration to load balance between port groups.

* Management Network needs to be redundant for HA isolation testing + heartbeats.
* FT logging network should be redundant.

- Beacon probing only works w/ teams of at least 3 ports, connected to separate physical switches.
- Beacon probing is unable to detect upstream network failures, that requires link state tracking or dependency ports in DELL.

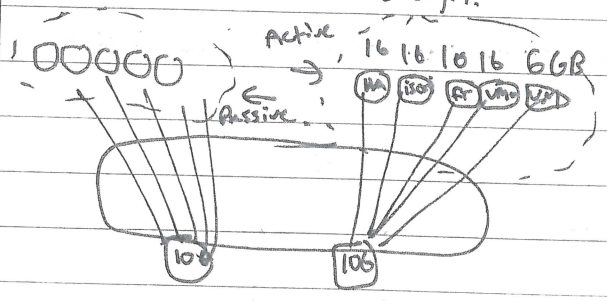
16B Sample Config



VCAP-DCD

8/24/2011

Mod 5 : Network Design.



10Gb example Active/Passive w/ traffic shaping.

← ingress traffic shaping

put 10Gb Adapters in PCIe 8x Slots.
dual port 10Gb adapters in PCIe 16x slots.

Backplanes can be a bottleneck (esp. riser cards)

Design book says to put SC/VMK nics in dvSwitch. Real world suggests otherwise.

vSwitch Security Design:

- Change default settings for Forged transmits and MAC address changes to REJECT, unless otherwise required not to.
  - ↳ prevents compromised VM from using MAC spoofing to impersonate another ~~VM~~ server on the network.
- ✱ Need to allow → NLB.
  Testing IOS,
  When App licensing depends on MAC Addresses.
- If MAC changes + Forged transmits are needed, enable them only on a specific port group.
- Leave promiscuous mode disabled unless needed. Enable only on required port group. Apps needing it are IDS, packet capture utilities, and performance monitoring tools.

VCAP-DCD

8/24/2011

Mod 5 Network Designs.

Physical NICs.

- Configure at least 2x pNICs for each network.
  - Checksum offload
  - TCP segmentation Offload (TSO)
  - 64-bit direct DMA.
  - Jumbo Frames.
- leave Auto negotiation enabled.
- use 10G Adapters that support NetQueue ← multiple Queues, each queue associated w MAC Addr, managed by the hypervisor
-

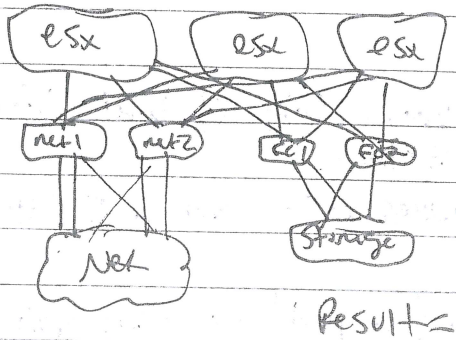
FCoE

- if the infrastructure supports it, use FCoE to converge FC/IP [redacted] traffic, reduces # of HBAs.
  → CNA not supported w/ iSCSI or NFS in VSH.0.

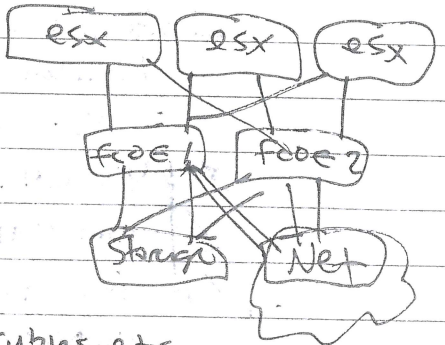
CNA Appears to ESX as 10Gb Enet NIC + FC HBA: use priority group bandwidth reservations to slice traffic up. ← tool on the CNA.

Converged Design Comparison

Traditional:



Converged:



Result ← Fewer ID Ports, Switches, Cables, etc.

VCAP-DCD

8/24/2014

Mod 5: Network Design:

Physical Switches:

- Always Redundant.
- manually configure trunk ports if using VLANS.
- STP Portfast enabled.

Jumbo Frames:

- Book says use them. Community Disagrees.
  - for IP Storage
- for VMs, leave Standard MTU unless specific App Requires it

VMDirectPathIO.

Dont use unless there is a clear use case.

- are these examples of Apps with high transactional workloads that require the lowest possible latency?
  - esp effective on high speed devices like 10Gb enet

Whatist? → provides access to a network device while bypassing the VM layer.

Limitations:

Limited Hardware Support: only w/ Intel VT-d, only on a few 10Gb network controllers.

- You cannot assign devices in the same DCI slot to both VMdirect and VMkernel.
  - ie. on a 10Gb NIC (Dual port) you cant split it so 1 goes to VMDPIO and the other to another Port group.
- NOT compatible with vMotion, HA, DRS, hotAdd,

VCAP-DCD

8/24/2011

Module 5: Network Design.

IPV6: Enable it only if Required.

IPV6 increases mgmt overhead.

NO compelling reason to enable.

Limitations:

IP Storage is Experimental on IPV6.

HA/FT Not supported

Requires endnets.

DNS:

Resolve long ANY Short Names.

Records for the following, both forward & Reverse:

- ESX/ESXi hosts
- vCenter
- vCenter Module systems
- IP Storage.

Always Highly Available.

Configure each host w/ Redundant DNS Servers., Same for VMs.

* Configure Common port group names.

VCAP-DCD

8/21/2011

Mod 6: Storage Design.

Designing Storage Tiers is cost efficient.

Tiers:

Smallest data

Largest data.

- Mission Critical Data — Highest Performance, 5 9's, Highest cost, near 0 downtime
- Business Critical Data — Good Perf + 5 9's, less than 8 hr recovery, moderate cost
- accessible-online data — Online performance, 4 9's, less 8 hr recovery, lower cost
- near-line data — Large volumes, protected data 3.9's, less than 1 hr auto retrieval, cost-sensitive
- [redacted]: Offline data — Archive data for backup or compliance, 72 hr retrieval, very cost sensitive.

Determine what tier data falls into:

- IOPS
- Capacity
- latency.
- Throughput
- availability

Consider SLA's.

Data may move b/n tiers during information lifecycle.

Tiering Example:

Tier 1	FC	15k	R10	8 Disks	10 VMs per datastore
Tier 2	SAS	10k	R5	6 Disks	15 VMs / DS.
Tier 3	SATA	7.2k	R6	6 Disks	15 VMs / DS.

VCAP-DCD

8/24/11

Mod 6 Storage.

What Protocol?

iSCSI, NFS, FC, FCoE?

- Factor in existing expertise.
- cost
- vendor relationships
- existing infrastructure.
- Performance (throughput + latency) requirements.

IP Storage Design.

- Use a Dedicated storage network.
  → increases security.
  If you can, physical isolation is preferred.
- limit the # of switches to traverse. more switches = more latency
- routing iscsi/NFS is not supported.
- Configure Jumbo frames.
- use TSO Nics.
- Storage Design should be modular + scalable.
  ↳ be able to add spindles
  front-end ports
  network infrastructure.

Calculating Capacity:

use CP or vendor tools.

Factor in vSwap files + snaps. ← 20-30% for this.

Factor in growth. ← Store sizes can be impacted by RTO limits.

VCAP-DCD

8/24/2011

Module 5: vStorage Design.

12 to 15 VMs per VMFS Volume.
20-30% for overhead.

VMFS Block Size:	MAX Disk Size
1 MB	256 GB
2 MB	512 GB
4 MB	1024 GB
8 MB	2048 GB

Determine block size by largest required virtual disk.
Block size also used to extend thin LUNs. important for design.

* Recommend always using MAX block size.

Reducing Latency:
- Dedicated Networks
- reduce queuing.
- jumbo frames

Command Queuing on Host:
- VMs on the same host and same LUN share a LUN Queue depth
- If more IOs are generated than the Queue depth, IOs are queued in the kernel and latency goes up.

Queuing @ the Array.
- IO can bottle up @ the LUN Queue or at the individual disk Queue, depending on the array.

Queuing = BAD

VCAP-DCD

8/21/2011

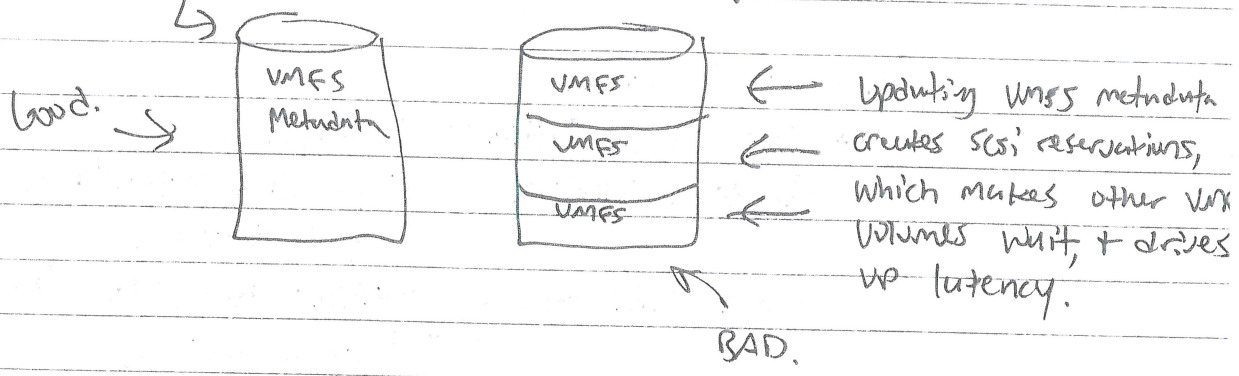
Module 5 vStorage Design.

- ensure IOs are not consistently hitting LUN Queues.
  - ↳ either increase the LUN Queue depth or
  - ↳ move VMs to another (less congested LUN)

Use esxtop to see if queue depths are being stressed.

VMFS volumes per LUN:

Single Partition w/ Single VMFS per LUN.



Latency:

Below 15MS Good (Green)
B/N 15MS + 30MS (Yellow)
Above 30MS Bad. (red)

Improve latency? Reduce oversubscription!

Control Access to NFS:

1. Network Segmentation
2. Dont mount datastore
3. Datastore Permissions.

ISCSI Access:

- configure static IPs.
- physical + VLAN separation
- CHAP.
- Datastore permissions.

VCAP-DCD

8/24/2011

Module 5: vStorage Design.

Fiber Channel Access Control:

1. Zoning
2. Lun Masking
3. Datastore Permissions.

Zoning is done for Performance + Access control.
↳ controls which HBA's have access to which Storage Ports.

Zoning for Performance:

* Use Single initiator Zoning!!

Single initiator: → single HBA port w/ one or more target devices.

Multi-initiator: → multiple HBA ports w/ one or more target devices.

no need for initiators to talk to each other. also risks ~~RSCN~~ RSCN (Registered state Change Notification messages from interrupting normal IO.

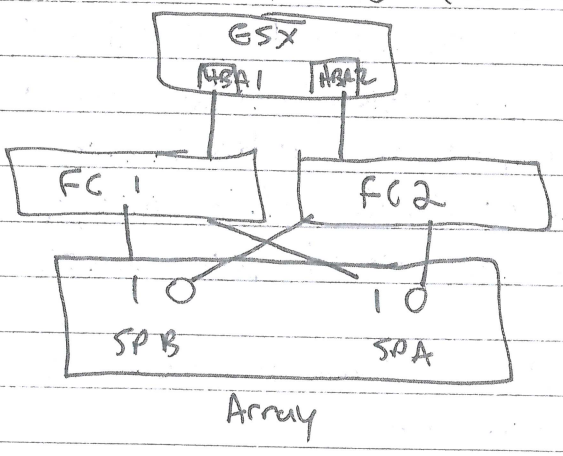
RSCN Messages are sent when an initiator enters or leaves a fabric. This initiates a Loop Initialization Primitive (LIP), which can interrupt IO from other initiators. Single initiator zones cannot interrupt other initiators w/ LIP's, due to RSCN because the initiators cannot talk to each other.

VCAP-DCD

8/24/2011

Mod 5: vStorage Design.

Zone for performance: Make sure HBA's are zoned to SP ports on different controllers for availability & multi-pathing.



HBA1 to SPAØ
HBA1 to SPBØ
HBA2 to SPA1
HBA2 to SPB1

Zoning for Access control:
Security.
Performance. Limit LUNs per host
Cluster members need access to the same LUNs.
Use WWN zoning rather than port zoning
↳ can move cables to different ports w/o affecting zoning

- LUN Masking: Can be used when Access Zoning is not enough.
  - Masking on the Array is preferred, although it is possible on the host.

Use LUN Masking for Boot from SAN Requirements.

Use the same LUN #s across hosts, esp for arrays that do not support NAA names.

Multi-Pathing: Recommends 4 paths per LUN for max performance + availability.

VCAP-DCD

8/24/2011

Mod 5: vStorage Design.

Use 2 single port HBA over 1 dual port HBA.

Verify that the correct MP Policy is selected for the array
- MRU for Active/Passive → avoids path thrashing.
- fixed, RR for Active/Active
- MRU or RR for ALUA arrays.
- MRU for virtual port arrays.

Use the vSphere client to create volumes. Auto Aligned on 64k boundries.
- VMFS on boot disk cannot be Aligned.

VMDK or RDM?
- VMDK where possible. Performance is not a reason to use RDM.

RDM Required for:
- MS clustering.
- N_PIV
- SAN Mgmt inside VM.
- Anything needing access to HW specific SCSI commands

Use virtual compatibility where possible.
- Allows snapshots.
- W2K8 clustering requires physical RDM Access.

Thick or Thin?

Thick	Thin
• Support for FT.	• decreases cost.
• Simpler config + monitoring	• pay as you grow.
• less efficient	• high risk.
• more expensive.	• complex.

VCAP-DCD

8/24/2011

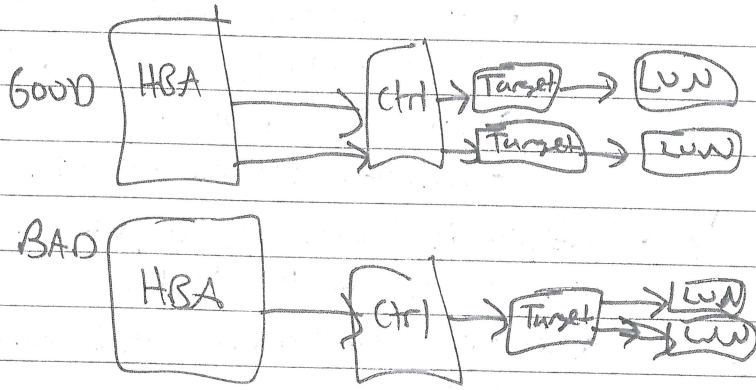
Mod 5 vStorage:

Software vs Hardware initiators.

- Software is free.
- do not mix both types in the same host.
- Hw iscsi uses less host CPU. About same performance.

One LUN per target is best for performance.

initiator establishes a single connection to a target.
One target per LUN.



Configure ports consistently across hosts, switches, and arrays.
↳ consistent tuning + easier trouble shooting.

☆ Dont mix HBAs from different vendors. Different model from same vendor is ok.

Boot from SAN:

FC + HW iscsi only.

- Can increase complexity, adds add'l dependency on SAN.
  HCL smaller.

Servers can be denser. Run cooler, NO disks to configure or install. Hosts less expensive.

VCAP-DCD

8/24/2011.

Boot from SAN.

- Exclusively mask boot LUN to host.
- use single initiator zoning.
- create single 1600MB Diag partition

BFS supported in ESX 4.0, ESXi 4.1.

NPIV:

NPIV assigns a WWN to a VM. Gives a VM an identity on the SAN.

- track storage traffic per VM.
- Zone & Mask LUNs per VM.
- SAN QoS per VM.
- Improve ID to VM via array caching.
- Configure if there is a mgmt requirement to monitor SAN LUN usage @ VM level.
- a security Requirement that mandates zoning a LUN direct to a VM.

Requirements:

- VMs use RDMs.
- FC HBA must support NPIV.
- FC Switches must support NPIV.
- ESX/ESXi must have access to all LUNs used by their VI

NPIV cannot be used on VMs protected w/ FT.

VCAP-DCD

8/24/2011

Module 7: Virtual Machine Design

Virtual Machine CPUs.

How many? Default to 1. Scale up if there is a need.

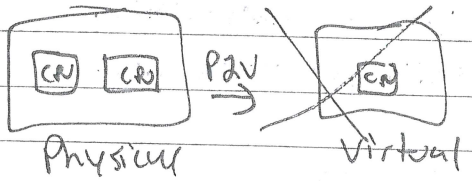
if using SMP VMs, make sure apps are multi-threaded so they can make use of 2x or 4x or 8x CPUs.

If the App is not multi-threaded, use multiple single VCPU VMs.

Use as few vCPUs as possible.

★ # of vCPUs in a VM cannot exceed the core/hyperthread count in a host it must be possible for all vCPUs to be scheduled.

for P2V's, create VM hardware that matches the physical machines hardware abstraction layer.



← BAD.

Virtual Machine Memory:

- keep all guest active memory in physical RAM.
  - → limit host memory overcommitment, or set reservations, or both.
  - → Reservations should be set slightly higher than any active memory size (for overhead)

Better to design around not overcommiting memory

★ Always have Transparent Page Sharing Enabled.

Always load VMware tools and always enable ballooning.

VCAP-DCD

8/24/2014

Module 7 : Virtual Machine Design.

Shares, Reservations, Limits.

- ☆ Deploy VMs w/ default settings unless clear reason to do otherwise.
  - ? Are there Apps that NEED Resources even during contention?
    - ↳ Use reservations.
- increases complexity + admin overhead.

VM Disks:

Deploy a system and data disk.

☆ Dont place ALL system disks on one datastore and data disks on another.

☆ Put system + data disks on the same datastore, unless they have widely varying IO characteristics.

☆ Configure one partition per virtual disk.

Simplifies SRM + snapshots.

↳ exception is if one VDisk is very large.

Swap location.

1st Default, shared storage w/VM files.

↳ con is that vswap files are replicated w/SRM.

2nd

Local storage: Reduces replication time. Local swap should not affect vm performance. Slows vMotion migration.

3rd

Dedicated VSwap SAN storage. Good vMotion performance, non-replicated LUN, bad is admin overhead.

VCAP-DCD

8/24/2011

Module 7 : Virtual Machine Design.

Disk Adapters

MS Clustering Requires SAS Devices for shared & Quorum disks.

Use Paravirtual Scsi adapters.
Cannot be used on boot disks in 4.0.
PVSCSI not suited for local storage.
Snapshots can negate performance of pvscsi.
Not supported w/ FT in 4.0.

NICs.
use vmxnet3 where possible
use standard sizing for VMs
ex

item	small	medium	large	ex large
CPU	1	2	3	4
Memory	1	2	4	8
disk	50	100	200	300

Use vApps to guarantee power on order.

Secure virtual machines as you would physical machines.
↳ AV, AS, ID, Firewalls.
Use vCenter server roles to limit access to VM console views.
Create VMs from templates that have been secured.
Add security products like vShield Zones to the design.

VCAP-DCD

8/24/2011

Module 8: Management and Monitoring Design.

Design management around ESXi, even if using ESX.
Limit the use of service console CLI and agents.
Deploy vCLI, PowerCLI, vMA
↳ install into VMs protected by HA.

Use ESXi embedded where possible.

Always automate! Use host profiles!

Create directory service users.

Create specific windows groups. Remove windows Administrators group from the vCenter Administrator role and create a specific vCenter Admins group.

Grant roles to groups, not users.
Use folders to assign roles to objects that require similar acce

In large environments, consider a small management cluster.

Use thinapp to package the vSphere client w/ All necessary plugins + put that thinapp on a network share.

Use NTP for All hosts + Mgmt tools. Native time sync in VMs. Disable VMware tools time sync.

VCAP-DCD

8/24/2011

Module 8. Mgmt + Monitoring Design.

Create snapshot management policies that govern their use.

Snaps are not backup solutions. they are back-OUT solutions.

Stat level 1 for longest term Retention.
Level 2 for long term performance monitoring when device stats are not required but you want more than basics.
level 3 for short-term monitoring after encountering a problem w/ a device.
Level 4 same as 3.
Levels 3 and 4 should only be used for a short time.

Use best practices + SLAs to determine what to monitor.

Logging → if org has its own logging (syslog) server that can handle vSphere, use it.
Otherwise, use VMA + vilogger.

VCAP-DCD

4.0

Config maximums for HA.

VMs per host (~~test~~ 8 or fewer hosts in cluster) in 4.0 = 100
4.0 U1 = 160

VMs per host for clusters w/ greater than 8 hosts = 40

Max failover hosts = 4

vCenter, powered on VMs = 2000 (32-bit), 3000 inventory.
3000 (64-bit), 4500 inventory.

100 hosts per datacenter.

320 VMs per host. 256 LUN/Server.

4.1

→ 10,000.

→ 320 VMs per host.

4.0 GA. Knodes or smaller - 100 VMs / host
4.0 Update 1 160 VMs.

All 4.0 9 nodes or higher - 40 VMs/host

4.1 → 320 VMs per host.