

Rahul Stephen

VCAP5 - DCD resources

<http://www.techexams.net/forums/virtualization/95566-vcap5-dcd-resources.html#post804369>

<http://networksandservers.blogspot.in/2011/02/high-availability-terminology-ii.html>

<https://communities.vmware.com/docs/DOC-9279>

<http://kendrickcoleman.com/index.php/Tech-Blog/vmware-vsphere-5-host-nic-network-design-layout-and-vswitch-configuration-major-update.html>

<http://www.youtube.com/watch?v=cTTU5f-9fNc>

Most Imp

<https://communities.vmware.com/thread/489031>

<http://www.youtube.com/watch?v=dEZQVFtsmU>

<https://communities.vmware.com/thread/489935>

<http://www.youtube.com/watch?v=yUftSnx39N0>

My Notes

When to Virtualize:

- DataCenter Consolidation
- Reduced OPEX (Power \ Cooling \ Cabling \ DC Rentals \ Support Cost)
- IT Agility (Provisioning \ upgrades \ VDI)
- Increased Stability and Security

When not to virtualize

- Un-Supported OS
- Licensing \ No-Vendor Support
- Extremely latency sensitive application
- No-inhouse knowledge of virtualization

Vision → Scope → Goals → Requirements → Risks → Assumptions → Constraints

5 Step Design Process – Initial Design Meeting → Current state analysis → Stakeholder \ SME Training

→ Design Sessions (Stakeholders \ SMEs) → Design Deliverables

Requirements (Functional and non-functional)

Functional - what the system must be able to do.

- Business Rules
- Transaction corrections \ Adjustments \ Cancellations
- Admin Functions
- Authentication
- Audit Tracking
- External Interfaces
- Reporting requirement
- Legal and regulatory requirements

Non-Functional – How the system performs \ Behavior of the system \ Quality Traits.

- Availability – Uptime \ Downtime \ Redundancy \ Resilience
- Manageability – Maintainability \ scalability \ interoperability \ serviceability
- Performance – TPS \ IOPS \ Latency
- Recoverability – Backups \ RTO \ RPO
- Security – Access points \ Regulatory

Traceability – Bi-directional and is used to manage change and for test planning. It tracks the allocation of requirement to :

- Design document and build components
- UAT
- Test cases and results
- Impact of change

<http://professionalvmware.com/2011/09/brownbag-follow-up-vcap-dcd-objective-1-jason-boche/>

Functional testing is concerned with the functional requirements and covers how well the system executes its functions. These include user commands, data manipulation, searches and business processes, user screens, and integrations.

Non-functional testing is concerned with the non-functional requirements and is designed to evaluate the readiness of a system according to several criteria not covered by functional testing. It enables the measurement and comparison of the testing of non-functional attributes of software systems.

Performance Testing \ Security Testing \ Usability Testing \ Dependability Testing \ Reliability, Maintainability, Availability, Recoverability \ Miscellaneous Testing \ Interoperability, Compatibility, Portability, Configuration, Installability.

Conceptual \ Logical \ Physical Design:

- Conceptual – Business requirements \ Logical representations of the business
- Logical – Architects view \ Models \ ERD
- Physical – Implementers view \ Details information and product - solution names.

VMWare Value Journey – Future ROI is expected from OPEX savings and improved operational and business agility.

- Test \ Dev → IT production → Business Production → IT as a Service

Enterprise Java application on vSphere:

- Load Balancer → Web Servers → Java Application Layer → Database Servers
- Host-Affinity Rules - This is very useful in honoring ISV Licensing requirements. Rules can be created so that VMs run on ESX hosts in different blades for higher availability. Conversely, limit the ESX host to one blade in case network traffic between the VMs needs to be optimized by keeping them in one chassis location

Time Keeping in Virtual Environment:

- Tick Counting – causes CPU consumption via interrupts
- Tickless Timekeeping – No CPU utilization
- When Windows user DC based time sync, time from VMTools must be disabled. When this is disabled then VMtools still do time sync when tools startup \ snapshot \ resuming from suspend state or vMotion

Establish business objectives → Identify critical success factors → Identify Constraints and Risks → Capture assumptions

Reviewing the current design (WAN \ LAN \ SAN Config \ Server Architectures \ CMDB \ DR Planning \ AD \ License agreements \ SLAs \ SOPs \ dependencies \ inter-relationships)

vShield Suite :

- vShield Zones – Virtual Firewall
- vShield Edge – Network edge security and gateway services \ DHCP \ NAT \ VPN

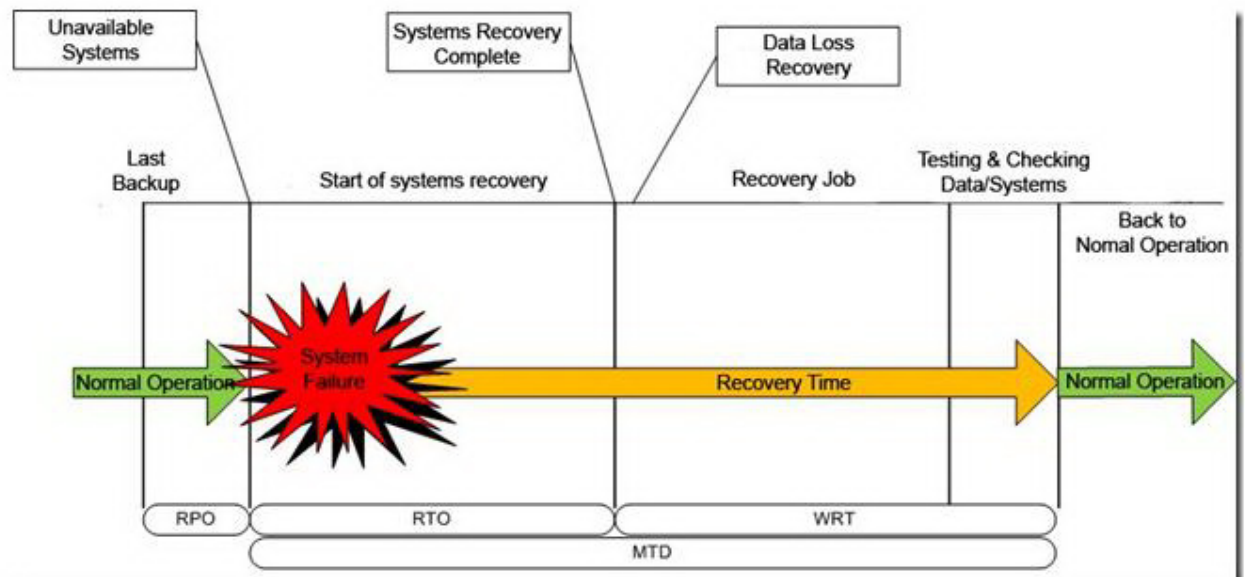
- vShield App – virtual NIC level firewall
- vShield EndPoint – Anti-virus

Application Discovery and dependency mapping

- Active Discovery : This method uses common network protocols to remotely query servers in the managed network and obtain supplementary CI data about network hosts. However, using just active discovery can place an unnecessary burden on the network. In addition, large segments of CI data don't change all that often, making repeated realtime active discovery unnecessary for many. Furthermore, although active discovery uncovers detailed CI data about hosts and services, it doesn't easily or directly provide information about how they relate to others. But active discovery doesn't require agents, and delivers a wealth of solid CI data.
- Passive Discovery : This method provides more of that relationship data. By connecting to core span or mirror ports on network switches and sampling network traffic, passive discovery can identify network hosts and servers, their communications and connections, and what services and protocols are being exchanged at what time. Although another rich source of data, you need some additional capabilities to assemble this raw data into actionable information

The VMware vCenter™ Application Discovery Manager (ADM) is an enterprise datacenter management solution that uses agentless discovery and provides continuous dependency mapping of applications. ADM helps you gain an understanding of your service dependencies. ADM also provides automated and real-time application discovery capability across physical and virtual environments.

Traditional BC \ DR solutions are Costly \ Unreliable \ Complex and difficult to Test and Deploy



Availability – HA \ FT \ Redundancy at different layers (Storage – PSA \ Network – teaming \ Servers – HA and FT \ Guest OS and Application – HA VM Monitoring)

- HA – Master and Slave relationship. Works when vCenter is down responds to Servers failure and network isolation. Elected happens and hosts with max datastores is selected as master, communicates with vcenter about hosts and VM status. Host network isolation occurs when a host is still running, but it can no longer observe traffic from vSphere HA agents on the management network. If a host stops observing this traffic, it attempts to ping the cluster isolation addresses. If this also fails, the host declares itself as isolated from the network. vSphere HA uses TCP and UDP port 8182 for agent-to-agent communication.

http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2033250

vSphere HA Advanced Attributes

You can set advanced attributes that affect the behavior of your vSphere HA cluster.

Table 2-2. vSphere HA Advanced Attributes

Attribute	Description
das.isolationaddress[...]	Sets the address to ping to determine if a host is isolated from the network. This address is pinged only when heartbeats are not received from any other host in the cluster. If not specified, the default gateway of the management network is used. This default gateway has to be a reliable address that is available, so that the host can determine if it is isolated from the network. You can specify multiple isolation addresses (up to 10) for the cluster: das.isolationaddressX, where X = 1-10. Typically you should specify one per management network. Specifying too many addresses makes isolation detection take too long.
das.usedefaultisolationaddress	By default, vSphere HA uses the default gateway of the console network as an isolation address. This attribute specifies whether or not this default is used (true false).
das.isolationshutdowntimeout	The period of time the system waits for a virtual machine to shut down before powering it off. This only applies if the host's isolation response is Shut down VM. Default value is 300 seconds.
das.slotmeminmb	Defines the maximum bound on the memory slot size. If this option is used, the slot size is the smaller of this value or the maximum memory reservation plus memory overhead of any powered-on virtual machine in the cluster.
das.slotcpuinmhz	Defines the maximum bound on the CPU slot size. If this option is used, the slot size is the smaller of this value or the maximum CPU reservation of any powered-on virtual machine in the cluster.
das.vmmemoryminmb	Defines the default memory resource value assigned to a virtual machine if its memory reservation is not specified or zero. This is used for the Host Failures Cluster Tolerates admission control policy. If no value is specified, the default is 0 MB.
das.vmcputminmhz	Defines the default CPU resource value assigned to a virtual machine if its CPU reservation is not specified or zero. This is used for the Host Failures Cluster Tolerates admission control policy. If no value is specified, the default is 256MHz.
das.iostatsinterval	Changes the default I/O stats interval for VM Monitoring sensitivity. The default is 120 (seconds). Can be set to any value greater than, or equal to 0. Setting to 0 disables the check.

Attribute	Description
das.ignoreinsufficienthbdastore	Disables configuration issues created if the host does not have sufficient heartbeat datastores for vSphere HA. Default value is false.
das.heartbeatdsperhost	Changes the number of heartbeat datastores required. Valid values can range from 2-5 and the default is 2.

In environments that use only network-based storage protocols, such as iSCSI and NFS, and those that share physical network components between the management and storage traffic, the recommended isolation response is Power Off.

<http://www.yellow-bricks.com/vmware-high-availability-deepdiv/>

Primary is responsible for: •restarting failed virtual machines \ •exchanging state with vCenter \ •monitor the state of slaves.

The host that is participating in the election with the greatest number of connected datastores will be elected master. If two or more hosts have the same number of datastores connected, the one with the highest Managed Object Id will be chosen. Geo-Dispersed cluster, when the cluster is split in two sites due to a link failure each “partition” will get its own master.

- Isolated - Is not receiving heartbeats from the master \ Is not receiving any election traffic \ Cannot ping the isolation address
- Partitioned - Is not receiving heartbeats from the master \ Is receiving election traffic \ (at some point a new master will be elected at which the state will be reported to vCenter)
- T0 – Isolation of the host (slave)
- T10s – Slave enters “election state”
- T25s – Slave elects itself as master
- T25s – Slave pings “isolation addresses”
- T30s – Slave declares itself isolated
- T60s – Slave “triggers” isolation response

Ensure port 8182 is open as that’s used for interhost communication.

Slot size is comprised of two components, CPU and memory.

- vSphere HA calculates the CPU component by obtaining the CPU reservation of each powered-on virtual machine and selecting the largest value. If you have not specified a CPU reservation for a virtual machine, it is assigned a default value of 32 MHz. You can change this value by using the `das.vmcputminmhz` advanced attribute.)
- vSphere HA calculates the memory component by obtaining the memory reservation, plus memory overhead, of each powered-on virtual machine and selecting the largest value. There is no default value for the memory reservation.
- Slot Calculations: `das.slotCpuInMHz` or `das.slotMemInMB`. The advanced setting `das.slotCpuInMHz` and `das.slotMemInMB` will allow you to specify an upper boundary for your slot size
- Used \ Available and Failover Slots
- The Host Failures Cluster Tolerates policy avoids resource fragmentation by defining a slot as the maximum virtual machine reservation.
- On ESXi hosts in the cluster, vSphere HA communications, by default, travel over VMkernel networks, except those marked for use with vMotion. If there is only one VMkernel network, vSphere HA shares it with vMotion, if necessary. With ESXi 4.x and ESXi, you must also explicitly enable the Management Network checkbox for vSphere HA to use this network.
- You can use vSphere Fault Tolerance with vSphere Distributed Resource Scheduler (DRS) when the Enhanced vMotion Compatibility (EVC) feature is enabled. This process allows fault tolerant virtual machines to benefit from better initial placement and also to be included in the cluster's load balancing calculations. FT needs thick VMDK files or RDMS in Virtual mode. No support for SMP \ Snapshots \ sVmotion and Linked Clones.
- A PDL is a condition that is communicated by the array to ESXi via an SCSI sense code. This condition indicates that a device (LUN) is unavailable and likely permanently unavailable.

Manageability – Effective VM cost Model \ Visibility into performance and utilization \ CMDB \ Scheduled maintenance plans

Performance:

PCPU USED(%)	<p>A PCPU refers to a physical hardware execution context. This can be a physical CPU core if hyperthreading is unavailable or disabled, or a logical CPU (LCPU or SMT thread) if hyperthreading is enabled.</p> <p>PCPU USED(%) displays the following percentages:</p> <ul style="list-style-type: none"> ■ percentage of CPU usage per PCPU ■ percentage of CPU usage averaged over all PCPUs <p>CPU Usage (%USED) is the percentage of PCPU nominal frequency that was used since the last screen update. It equals the total sum of %USED for Worlds that ran on this PCPU.</p> <p>NOTE If a PCPU is running at frequency that is higher than its nominal (rated) frequency, then PCPU USED(%) can be greater than 100%.</p> <p>If a PCPU and its partner are busy when hyperthreading is enabled, each PCPU accounts for half of the CPU usage.</p>
--------------	--

%USED	Percentage of physical CPU core cycles used by the resource pool, virtual machine, or world. %USED might depend on the frequency with which the CPU core is running. When running with lower CPU core frequency, %USED can be smaller than %RUN. On CPUs which support turbo mode, CPU frequency can also be higher than the nominal (rated) frequency, and %USED can be larger than %RUN.
%SYS	Percentage of time spent in the ESXi VMkernel on behalf of the resource pool, virtual machine, or world to process interrupts and to perform other system activities. This time is part of the time used to calculate %USED.
%WAIT	Percentage of time the resource pool, virtual machine, or world spent in the blocked or busy wait state. This percentage includes the percentage of time the resource pool, virtual machine, or world was idle.
%VMWAIT	The total percentage of time the Resource Pool/World spent in a blocked state waiting for events.
%RDY	Percentage of time the resource pool, virtual machine, or world was ready to run, but was not provided CPU resources on which to execute.

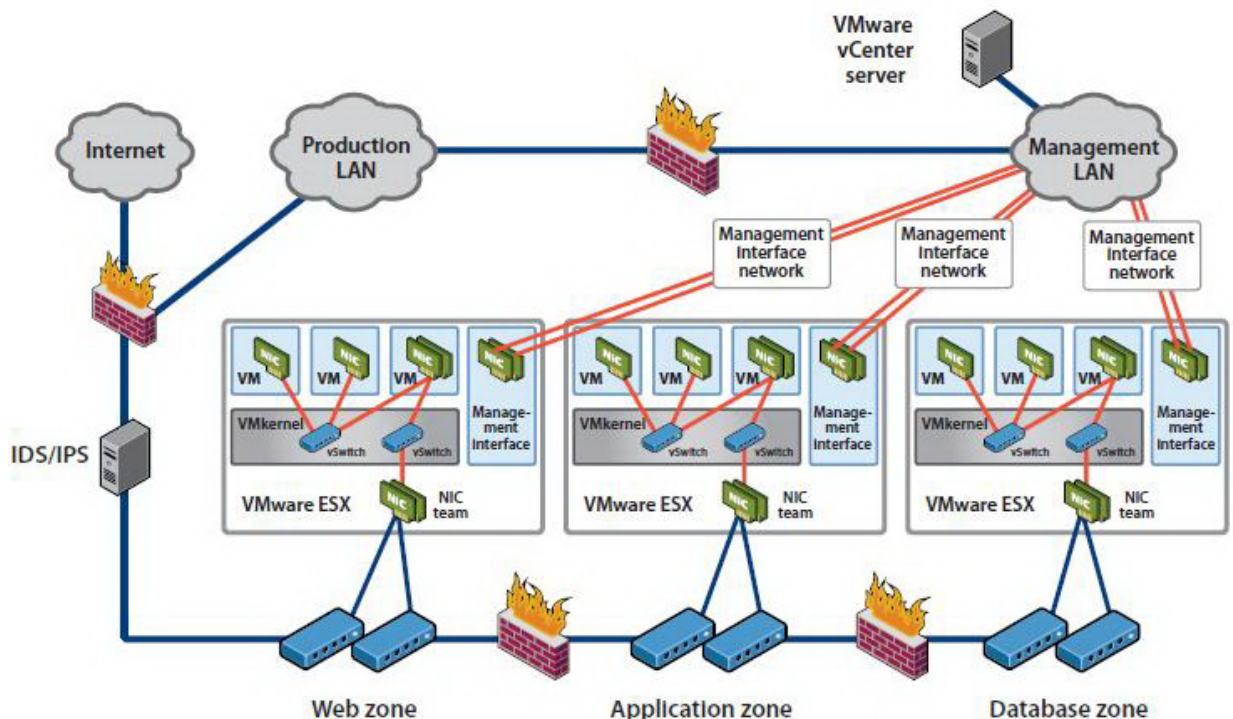
Storage:

DAVG/cmd	Average device latency per command, in milliseconds.
KAVG/cmd	Average ESXi VMkernel latency per command, in milliseconds.
GAVG/cmd	Average virtual machine operating system latency per command, in milliseconds.
QAVG/cmd	Average queue latency per command, in milliseconds.

Cloud Computing Benefits - On-Demand expansion \ Par per usage \ shared infra \ Reduced Cpx \ Access to Level-4 DC \ Increased security and adherence to best practices

Cloud Computing Challenges – Lack of interoperability between Cloud vendors \ Security \ adherence to Change management of vendors \ Regulatory requirements.

Partially Collapsed with Separate Physical Trust Zones



Advantages

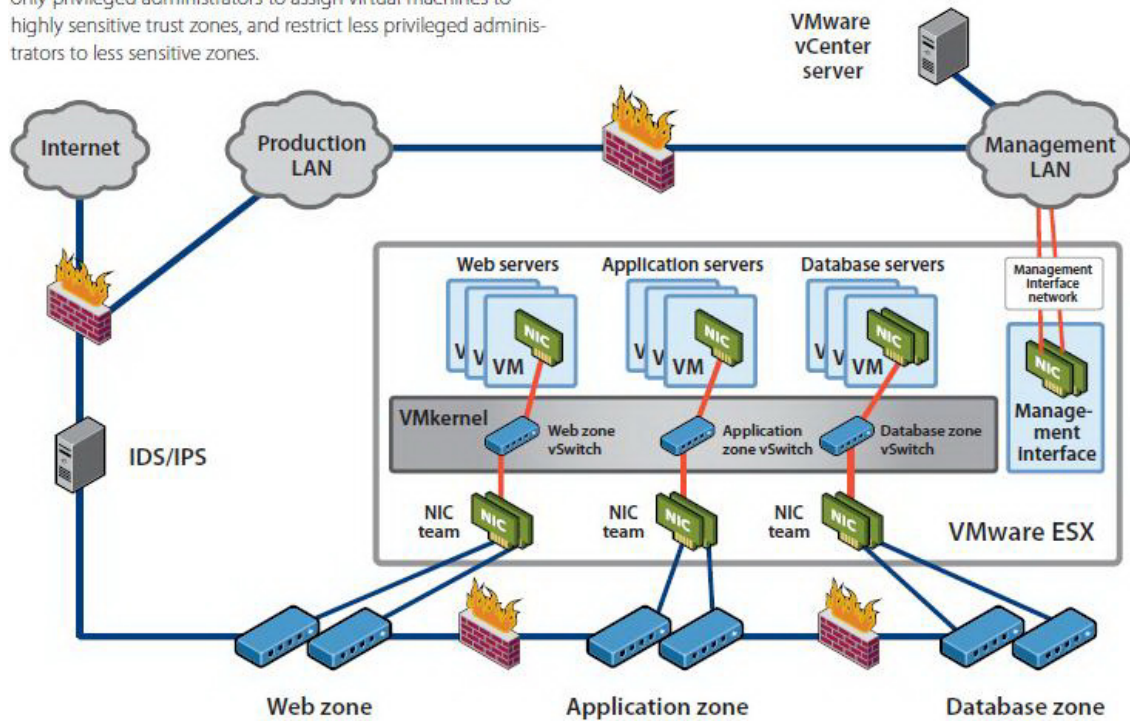
- Simpler, less complex configuration
- Less change to physical environment, and thus less change to separation of duties and less change in staff knowledge requirements
- Less chance for misconfiguration because of lower complexity

Disadvantages

- Lower consolidation and utilization of resources
- Higher costs because of need for more ESX hosts and additional cooling and power
- Incomplete utilization of the operational efficiencies virtualization can provide

Partially Collapsed with Virtual Separation of Trust Zones

only privileged administrators to assign virtual machines to highly sensitive trust zones, and restrict less privileged administrators to less sensitive zones.



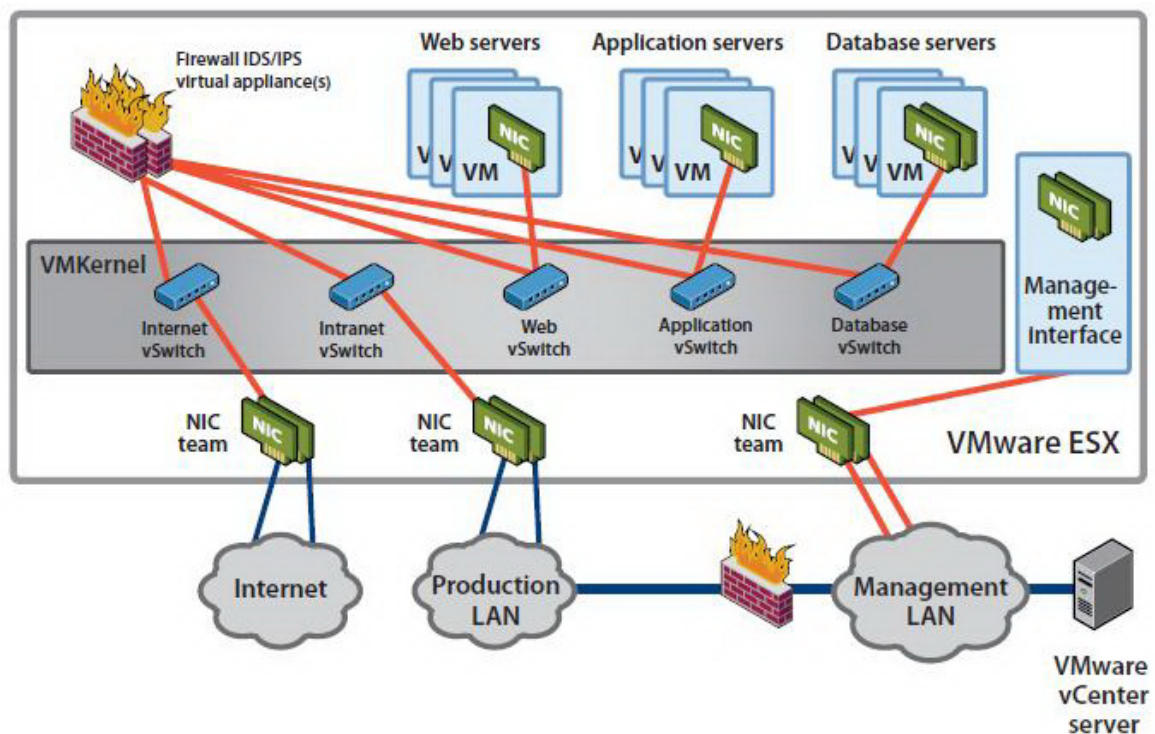
Advantages

- Full utilization of resources
- Full utilization of the advantages of virtualization
- Lower cost

Disadvantages

- More complexity
- Greater chance of misconfiguration requires explicit configuration of separation of duties to help mitigate risk of misconfiguration; also requires regular audits of configurations

Fully Collapsed Trust Zones



Advantages

- Full utilization of resources, replacing physical security devices with virtual
- Lowest-cost option
- Management of entire network from a single management workstation

Disadvantages

- Greatest complexity, which in turn creates highest chance of misconfiguration
- Requirement for explicit configuration of separation of duties to help mitigate risk of misconfiguration; also requires regular audits of configurations
- Loss of certain functionality, such as VMotion, if current virtual

VMotion

<http://frankdenneman.nl/2012/09/07/vsphere-5-1-vmotion-deepdive/>

The execution state primarily consists of three components:

1. The virtual device state, including the state of the CPU, network and disk adaptors, VSGA, and so on.
2. External connections with devices, including networking and SCSI devices
3. The virtual machine's physical memory

MTD – Maximum tolerable downtime – How long before major damage to business and significant loss of revenue occurs.

RPO – Recovery Point Objective – Point in time to which data must be restored after an outage \ disaster.

RTO – Recovery Time Objective – Amount of time that it takes to bring the services online after an outage \ disaster.

BIA – Business Impact Analysis

Event: A change of state which might have an influence for the management of a service or system

Incident: An event which is not part of the standard operation. It might cause a service disruption or reduce the productivity.

Problem: The cause of one or more incidents. Problems are usually identified because of multiple incidents.

Metro Cluster-

- Use case is aggressive RTO \ RPO
- ESXi management round trip time latency is 10ms
- Sync Storage replication RTT is 5ms
- Have 4 Network isolation address (2 on each site)
- Have 4 datastores for DS heartbeating (2 on each site)
- vSphere 5 introduces a new latency-aware Metro vMotion feature that not only provides better performance over long latency networks but also increases the round-trip latency limit for vMotion networks from 5 milliseconds to 10 milliseconds.

With the hardware-initiator iSCSI implementation, the iSCSI HBA provides the translation from SCSI commands to an encapsulated format that can be sent over the network. A TCP offload engine (TOE) does this translation on the adapter.

The software-initiator iSCSI implementation leverages the VMkernel to perform the SCSI to IP translation and requires extra CPU cycles to perform this work. As mentioned previously, most enterprise-level networking chip sets offer TCP offload or checksum offloads, which vastly improve CPU overhead.

MSCS on VMWare:

- In a cluster of virtual machines across physical hosts, the shared disk must be on a Fibre Channel (FC) SAN.
- LSI Logic Parallel for Windows Server 2003
- LSI Logic SAS for Windows Server 2008
- Disk format Select Thick Provision to create disks in eagerzeroedthick format.
- NO - iSCSI \ FCoE \ NFS disks \ vMotion \ FT \ NPIV \ Round Robin
- For a cluster of virtual machines on one physical host, all MSCS virtual machines must be in the same virtual machine DRS group, linked to the same host DRS group with the affinity rule "Must run on hosts in group."
- For a cluster of virtual machines across physical hosts, each MSCS virtual machine must be in a different virtual machine DRS group, linked to a different host DRS group with the affinity rule "Must run on hosts in group."

Networking:

TRAFFIC TYPE	BANDWIDTH USAGE	OTHER TRAFFIC REQUIREMENTS
MANAGEMENT	Low	Highly reliable and secure channel
VMOTION	High	Isolated channel
FT	Medium to high	Highly reliable, low-latency channel
ISCSI	High	Reliable, high-speed channel
VIRTUAL MACHINE	Depends on application	Depends on application

vDS supports static link aggregation but does not support LACP

NIOC configures on DVSwitch and the settings get applied on all dvUplinks (vmknics)

Users defined network resource pools are defined on DVSwitch, DVPortgroups are then mapped to network resource pools

<http://www.youtube.com/watch?v=7utuL4uAsdc>

NIC scenario with 8x1Gb Nics

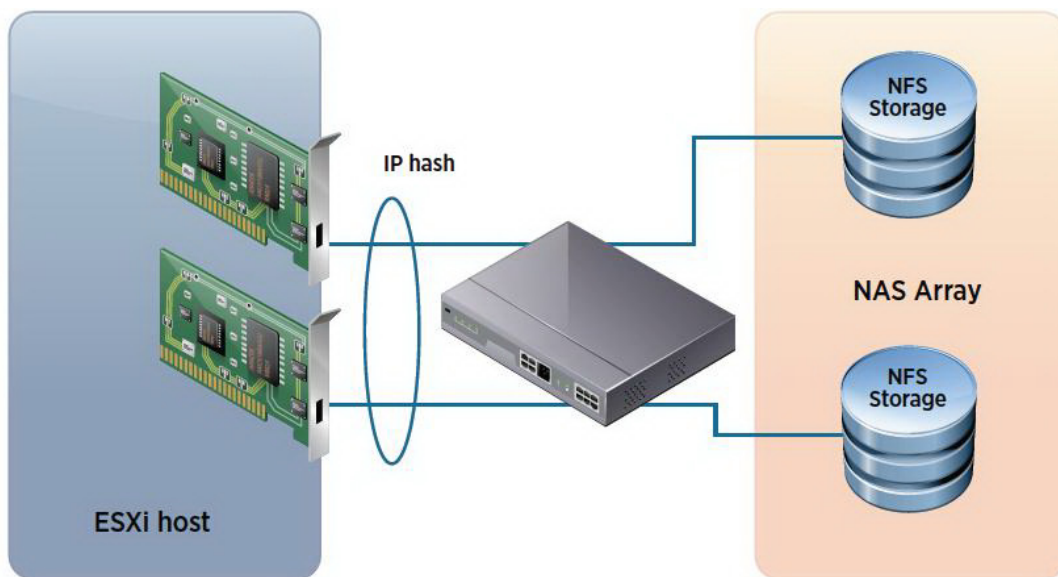
TRAFFIC TYPE	PORT GROUP	TEAMING OPTION	ACTIVE UPLINK	STANDBY UPLINK	NIOC SHARES	NIOC LIMITS
MANAGEMENT	PG-A	LBT	1, 2, 3, 4	None	5	–
VMOTION	PG-B1	None	5	6	–	–
VMOTION	PG-B2	None	6	5	–	–
FT	PG-C	LBT	1, 2, 3, 4	None	10	–
ISCSI	PG-D1	None	7	None	–	–
ISCSI	PG-D2	None	8	None	–	–
VIRTUAL MACHINE	PG-E	LBT	1, 2, 3, 4	None	20	–

2x10GB NIC

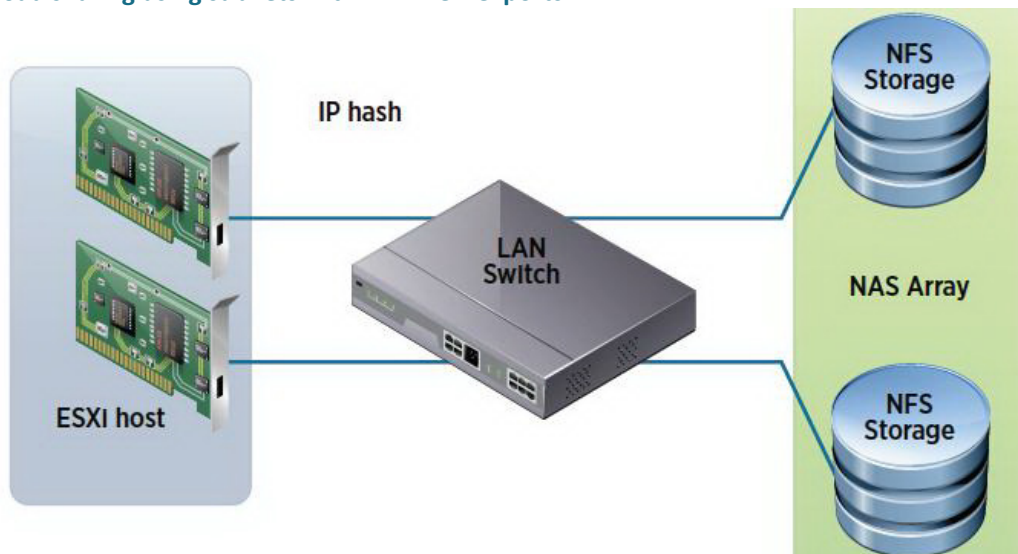
TRAFFIC TYPE	PORT GROUP	TEAMING OPTION	ACTIVE UPLINK	STANDBY UPLINK	NIOC SHARES	NIOC LIMITS
MANAGEMENT	PG-A	LBT	dvuplink1, 2	None	5	–
VMOTION	PG-B	LBT	dvuplink1, 2	None	20	–
FT	PG-C	LBT	dvuplink1, 2	None	10	–
ISCSI	PG-D	LBT	dvuplink1, 2	None	20	–
VIRTUAL MACHINE	PG-E	LBT	dvuplink1, 2	None	20	–

Storage NFS:

Load sharing using IP hash (Same Subnet) with LACP \ EtherChannel



All port groups using the same set of uplinks should have the IP hash load-balancing policy set
Load sharing using subnets with 2 VMKernel ports.



NFS.MaxVolumes for increasing # of NAS volumes from 8 – this is done on per host.