VCP Tips Corner

Know Your VMware Storage

Part1

Author: Ricky El-Qasem: 2009

This document points you at some information that might be useful for the VCP exam. It's worth noting that I may not cover everything about storage in this one document and you may see storage related info crop up in other documents. I may also have to revise the document as I remember extra's at a later date.

The following table describes what can be achieved between the different storage platforms.

Feature/Function	Local SCSI	NFS	iSCSI	FC	Comments
Boot ESX	Yes	No	Partial	Yes	To boot from iSCSI you need a hardware initiator
VMFS	Yes	No	Yes	Yes	with NFS the file system is dictated by the NFS OS
RDM	Yes	No	Yes	Yes	with NFS you don't see a physical disk but rather a network share
MSCS Cluster	No	No	Yes	Yes	A physical shared disk is required
VMotion	No	Yes	Yes	Yes	Shared storage is required
HA	No	Yes	Yes	Yes	Shared storage is required
DRS	No	Yes	Yes	Yes	Shared storage is required
DPM	No	Yes	Yes	Yes	Shared storage is required
Storage VMotion	Yes	Yes	Yes	Yes	

Yes	(green)
Partial	(orange)
No	(red)

Reference:
http://www.vmware.com/pdf/vsphere4/r40_u1/vsp_40_u1_esx_server_config.pdf#page=78

You need to also understand what is required in the way of networking components to attached storage:

Version/Storage	Service Console Port	Vmkernel Port
ESX - FC	No	No
ESX - Hardware iSCSI	No	No
ESX - Software iSCSI	No	Yes
ESX - NFS	No	Yes
ESXi - FC	No	No
ESXi - Hardware iSCSI	No	No
ESXi - Software iSCSI	No	Yes
ESXi - NFS	No	Yes

Reference :
http://www.vmware.com/pdf/vsphere4/r40 u1/vsp 40 u1 esx server config.pdf#page=83

In ESX 4 the installation of the service console filesystem is performed in a virtual disk now opposed to how it was installed in 3.x. The service console's partitions are stored in a .vmdk file. These partitions include /, swap, /var/ log, and all the optional partitions. The name of this file is esxconsole-<system-uuid>/esxconsole.vmdk. All .vmdk files, including the esxconsole.vmdk, are stored in VMFS volumes. This virtual disk has to be located on a local VMFS or a non-shared external SAN disk. It has to be non-shared to avoid deletion or being written over by another host.

Reference:
http://www.vmware.com/pdf/vsphere4/r40 u1/vsp 40 u1 esx vc installation guide.pdf#page=20

If you plan to upgrade from 3.x to 4.0 you will need:

At least 8.4GB of storage for the service console .vmdk

Reference:
http://www.vmware.com/pdf/vsphere4/r40_u1/vsp_40_u1_upgrade_guide.pdf#page=78

The /boot partition needs to be at least 100mb

Reference:
http://www.vmware.com/pdf/vsphere4/r40_u1/vsp_40_u1_upgrade_guide.pdf#page=75

The actual partition sizes for the installation are as follows:

/boot	=	1.25GB
/	=	5GB
Swap	=	600mb

You also have the option of creating optional partitions:

/home	=	512mb
/tmp	=	1024mb
/usr	=	no recommendation
/var/log	=	2000mb

Creating the optional partitions is advantages as it could help you avoid filling up the / partition. Filling the / partition has undesired effects and could lead to instability.

Reference:
http://www.vmware.com/pdf/vsphere4/r40_u1/vsp_40_u1_esx_vc_installation_guide.pdf#page=62

Know the difference between *Thin Provisioned* vdisks and *Thick* vdisks.
In ESX 4.0 you now have the option to choose between using a pre-allocated vdisk (*Thick*) or a vdisk that will grow on-demand to a predefined limit (*Thin Provisione*d)

Reference:
http://www.vmware.com/pdf/vsphere4/r40_u1/vsp_40_u1_admin_guide.pdf#page=118

You can convert between formats but converting from Thick to thin requires the use of Storage VMotion.

Reference:
http://www.vmware.com/pdf/vsphere4/r40_u1/vsp_40_u1_admin_guide.pdf#page165

It also worth noting that the largest VMFS volume size is that can exist on a single physical disk/LUN is 2TB minus a few bits for metadata. You can extend multiple volumes, 32 of them to make a 64TB volume, again minus a few bits for metadata. The largest file size you can have (vdisk) is 2TB but the block size of the VMFS volume has to reflect the size of vdisk you plan to use. Following is a simple rule of thumb to follow:

256GB file =   1MB block size
512GB file =   2MB block size
1TB file =   4mb block size
2TB file =   8mb block size

Notice how each selection doubles up from the previous:

Reference:
http://www.vmware.com/pdf/vsphere4/r40_u1/vsp_40_u1_esx_server_config.pdf#page=75

ESX 4 has the ability to communicate to *ISCSI storage* arrays in 2 different ways:

1. Through a *software initiator* which is code that resides within ESX that is able to talk directly to ISCSI storage through a standard physical NIC.
2. Through a *hardware initiator* which is a hardware HBA dedicated for the use of talking to ISCSI storage and tends to increase performance by 30% opposed to using a software initiator.

Reference:
http://www.vmware.com/pdf/vsphere4/r40_u1/vsp_40_u1_iscsi_san_cfg.pdf#page=10

http://www.vmware.com/pdf/vsphere4/r40_u1/vsp_40_u1_esx_server_config.pdf#page=85

ISCSI nodes which include ESX servers and ISCSI storage communicate to each other over IP on default port *3620* and address each using a unique identifier known as the **IQN address**.

Reference:
http://www.vmware.com/pdf/vsphere4/r40_u1/vsp_40_u1_esx_server_config.pdf#page=73

http://www.vmware.com/pdf/vsphere4/r40_u1/vsp_40_u1_iscsi_san_cfg.pdf#page=12

http://www.snia.org/tech_activities/standards/curr_standards/smi/SMI-S_Technical_Position_v1.1.1r1.pdf#page=103

To maximise performance VMware decided to go for *CHAP* authentication to secure the ISCSI communication. Not very secure I know but VMware's recommendation is to isolate ISCSI traffic on a segregated LAN or VLAN thus eliminating possible security breaches. Isolation is also useful to ensure dedicated bandwidth for storage IO for performance reasons. Of course from the storage array side of things you can tie down communications by restricting access to storage by IP or IQN as well as using CHAP.

Reference
http://www.vmware.com/pdf/vsphere4/r40_u1/vsp_40_u1_iscsi_san_cfg.pdf#page=36

ESX is also apt at using Fibre Chanel SANs as a storage platform too. There is far to much information to cover on FC to include in this document so I'm going to focus in on the important things that may crop up in the exam. You must have a good understanding of FC SANs whilst the purpose of this document is not to teach about SANs

First thing worth talking about is how FC nodes communicate with each other. FC devices which include servers and storage arrays use common protocol (FCP) developed to communicate over fibre optic cable like FC2, FC4 and FC8 which you guessed it refers to the possible speeds e.g. FC2 = 2Gbs. As far as anything you need to know for the exam I doubt VMware would go into this kind of depth. What is important is how these devices address each other. For the purpose of the exam you need to know that a unique address is given to each FC device known as the *WWN* number (*World Wide Name*) which is similar to a Ethernet MAC address if you are familiar to that concept. Under the hood there is a bit more to it than that but that's all you will need for the exam.

Reference:
http://en.wikipedia.org/wiki/Fibre_Channel

We mentioned the concept of isolation for security & performance for ISCSI while using VLAN etc, well in the FC world there is also a mechanisms available to provide isolation of FC devices and is known as *"Zoning"*. With Zoning you create a virtual segmented storage network where members of a Zone can talk to each other while devices outside of the Zone are unable to communicate with those inside. This function is configured through the FC Switch (the device used to connect FC devices together). There are 2 types of Zones:

1. Hard Zone – this is where the physical FC port on the switch is the member of the Zone. If a device is plugged into a port that is not a member it will not be able to communicate with other devices plugged into member ports.
2. Soft Zone – this is where the Zone is made of members using their WWN number so it matters not where the device is plugged in membership is restricted by WWN.

LUN Masking – In a scenario where you have just 1 storage array and many servers accessing that storage array how do you stop disks that are associated with one server but should not be associated with other servers from being visible to all servers? The answer is **LUN masking**. LUN masking is a feature of 99.9% of storage arrays that enables the presentation of LUNs (disks) to some servers whilst hiding them away from other servers that maybe still using the same storage array. You'd want do this to stop alien servers from wanting to claim/destroy disks that belong to a different platform. So for example in **most** cases you would not present a LUN with a VMFS datastore to a Windows server in hear of saying good bye to your VMFS partition.

NOTE: There is a misconception that LUN Masking is done at the FC Switch which is incorrect. LUN Masking as generally achieved through the management software of the storage array. The reason this myth came about was due to a VMware trainer had picked up this idea from one of the old VMware books and it filtered its way down. Now the reason it was mentioned in this manner is because some FC Switches (**some**) also have the ability to perform LUN masking but you should never take it for granted that it works in this way. Out of the 20+ FC Switches I had the fortune to play with I only ever saw this feature once.

There is somewhere else though where you can achieve LUN Masking... VMware includes the ability to perform LUN masking at a software level through the advanced settings of an ESX server.

Reference:
http://www.vmware.com/pdf/vsphere4/r40_u1/vsp_40_u1_san_cfg.pdf#page=81

Some Max Configs:

LUNs per Host	=	256
VMs per Host	=	256
NFS Datastores	=	64
FC HBAs per Host	=	8
FC HBA ports	=	16
FC paths to LUN	=	32
ISCSI HW initiators	=	4
ISCSI paths to LUN	=	8

Reference:
http://www.vmware.com/pdf/vsphere4/r40/vsp_40_config_max.pdf#page=3

Have a good understanding of SAN multipathing . ESX 4 maintains its own multipathing function and is driven by a choice of 3 driven policies:

Most Recently Used (MRU)
If a failover occurs and the path is reconnected again ESX will use the most recently path to access the given LUN and not fall back to the original path.
Fixed
Uses the designated preferred path. If ESX cannot find the preferred path, it selects a random alternative available path. ESX will automatically reverts back to the preferred path as soon as that path becomes available again.
Round Robin (RR)
A path selection algorithm that rotates through all available paths thus load balancing across the paths.

Reference:
http://www.vmware.com/pdf/vsphere4/r40_u1/vsp_40_u1_san_cfg.pdf#page=20

Once the LUN is presented to an ESX server and is formatted with VMFS it is given a human friendly name but underneath the hood the ESX refers to VMFS volumes using a unique address known as the vmhba address. Vmhba addresses are appear and are formatted like this:

vmhbaC:T:L

C = Channel (path and FC HBA used)
T = Target (which storage processor is used on the target)
L = LUN Id

So for example vmhba1:1:0 would refer to using HBA1, Storage Processor and LUN 0.

Note: storage processor is a term used by one storage vendor and some other vendors don't use the same terminology. Whilst the term storage processor is often referred to by VMware it's worth knowing that other vendors may use terms like storage connector.

Reference:
http://www.vmware.com/pdf/vsphere4/r40_u1/vsp_40_u1_san_cfg.pdf#page=51

Disclaimer:

Whilst this document is useful to consolidate the amount of revision needed for the VCP exam it's not intended to help you cheat your way through the exam. To find this information useful you will needed to have been on a certified VMware course at some point. Also the information in this document by itself isn't enough and you need to follow the reference links. The intention of this document is to kick start you and point you in the right direction.