## Questions based on HA

### Q1: Explain how restart of VM's is handled by HA in case of a master Esxi Host failure.

Ans: HA restarts VM's after failure of an Esxi host. But the time taken by HA to restart VM's is different in case of a slave Esxi failure and master Esxi failure. We will discuss here the case when master Esxi has failed.

In case of a failure of a master Esxi, restart of VM's are delayed till the time a new master is elected because only a master can perform VM restart. The timeline is explained as follows:

- T0 – Master failure.
- T10s – Master election process initiated.
- T25s – New master elected and reads the protected list.
- T35s – New master initiates restarts for all virtual machines on the protected list which are not running.

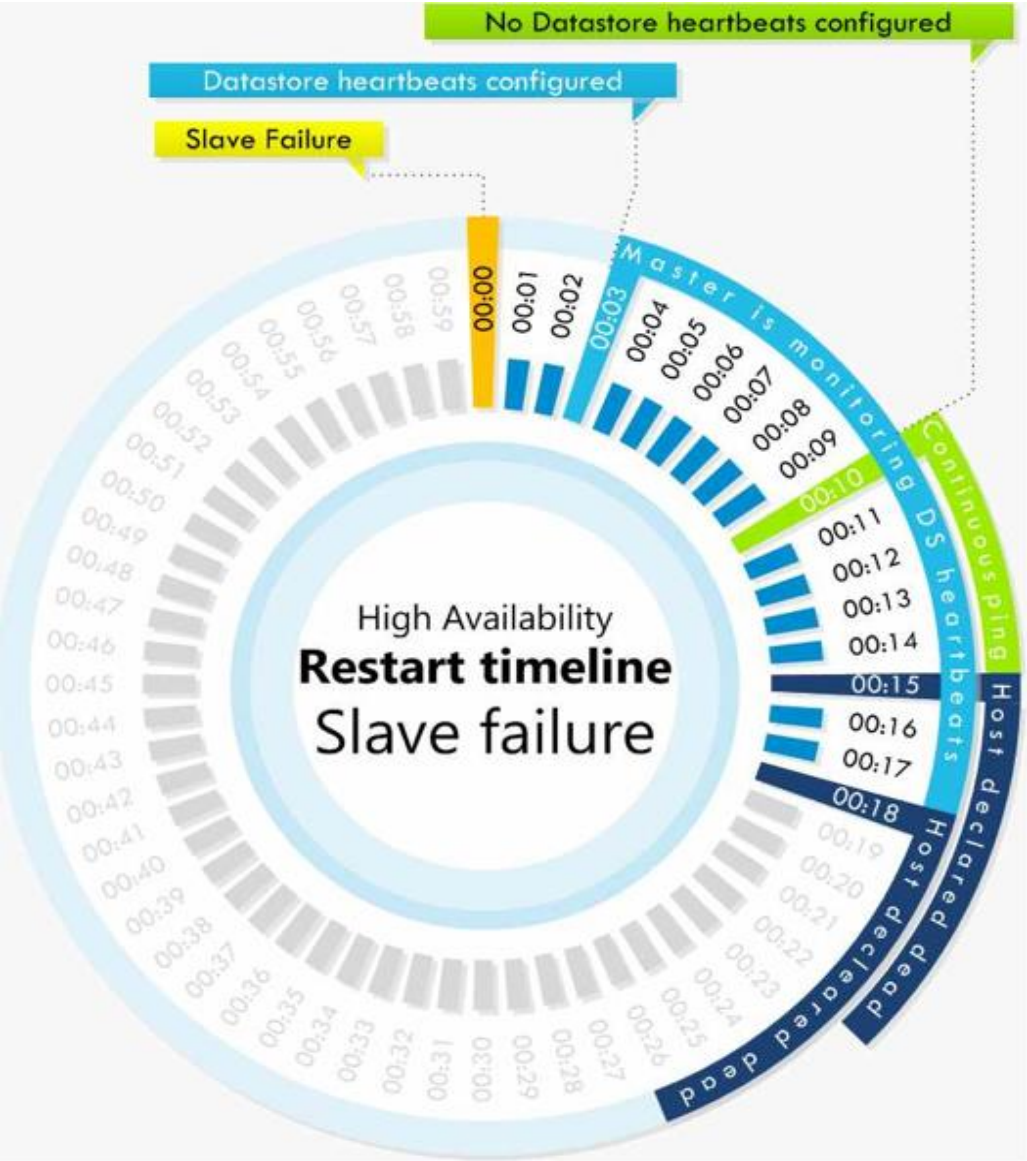
At T0 seconds master Esxi has failed, the election process is initiated by slave Esxi hosts after 10 seconds at T10. At T25 the newly elected master first reads the protected list file to find out which VM were protected by HA and are currently not running. At T35 seconds the master Esxi initiates the VM restart.


**Q2: Explain how restart of VM's is handled by HA in case of a slave Esxi Host failure.**

Ans: There are two different scenarios for restarting VM's in case of slave Esxi failure: one where heartbeat datastores are configured and one where heartbeat datastores are not configured. The timeline is as follows:

- T0 – Slave failure
- T3s – Master begins monitoring datastore heartbeats for 15 seconds
- T10s – The host is declared unreachable and the master will ping the management network of the failed host. This is a continuous ping for 5 seconds

- T15s – If no heartbeat datastores are configured, the host will be declared dead
- T18s – If heartbeat datastores are configured, the host will be declared dead
- The master monitors the network heartbeats of a slave. When the slave fails, these heartbeats will no longer be received by the master. We have defined this as T0. After 3 seconds (T3s), the master will start monitoring for datastore heartbeats and it will do this for 15 seconds. On the 10th second (T10s), when no network or datastore heartbeats have been detected, the host will be declared as "unreachable".
- The master will also start pinging the management network of the failed host at the 10th second and it will do so for 5 seconds. If no heartbeat datastores were configured, the host will be declared "dead" at the 15th second (T15s) and VM restarts will be initiated by the master.
- If heartbeat datastores have been configured, the host will be declared dead at the 18th second (T18s) and restarts will be initiated.




**Q3: Explain the VM restart retries timeline**

Ans: HA will respond when the state of a host has changed, or when the state of one or more virtual machines has changed. There are multiple scenarios in which HA will attempt to restart a virtual machine of which we have listed the most common below:

- Failed host
- Isolated host
- Failed guest Operating System

Prior to vSphere 5, the actual number of restart attempts was 6, as it excluded the initial attempt. With vSphere 5.0 the default is 5. There are specific times associated with each

of these attempts. The following bullet list will clarify this concept. The 'm' stands for "minutes" in this list.

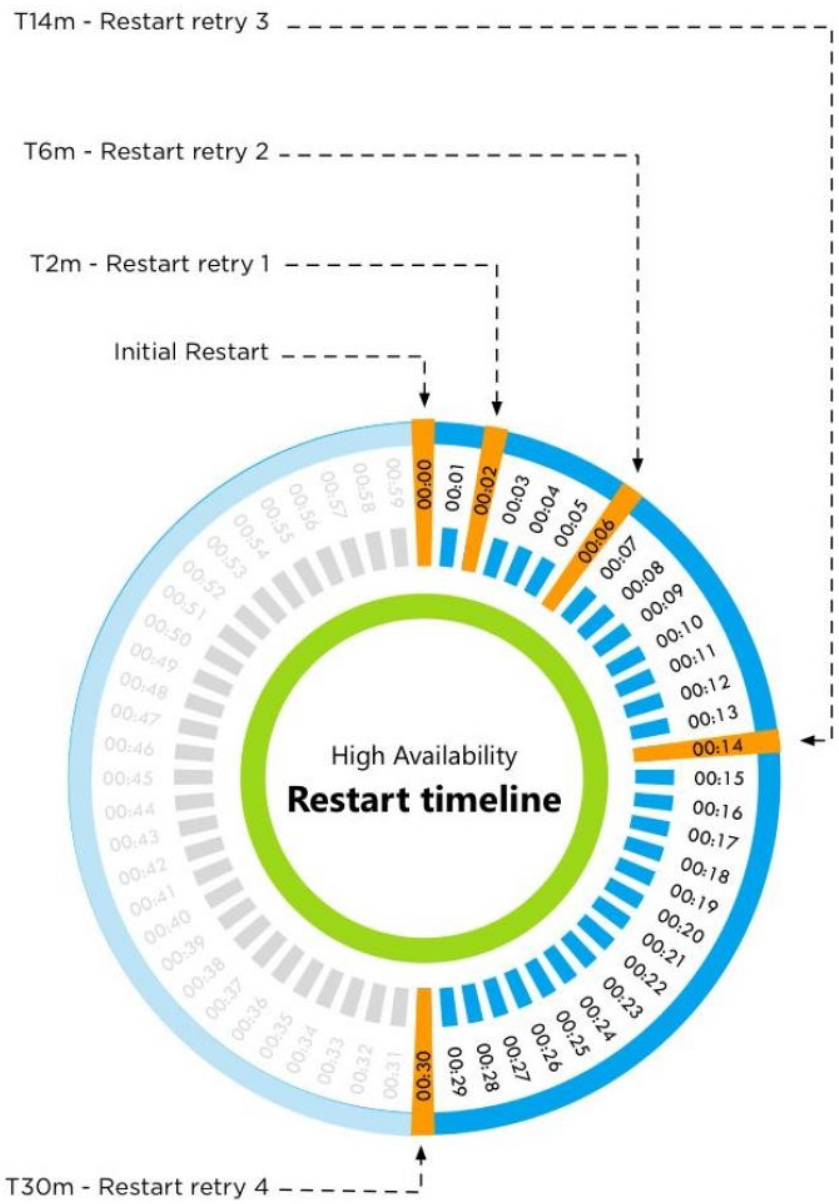
- T0 – Initial Restart
- T2m – Restart retry 1
- T6m – Restart retry 2
- T14m – Restart retry 3
- T30m – Restart retry 4

In case of a host failure, HA will try to restart the virtual machine on other hosts in the affected cluster; while performing the restart if this is unsuccessful on that host, the restart count will be increased by 1.

Let's say first restart attempt is made at T0 minutes when the host failure has occurred (In actual restart is not performed as soon as host has failed because HA takes some time before declaring host failure; read above the 2 scenarios which I have mentioned)

If the first restart attempt is failed, then the restart counter is increased by one and the next restart is attempted after 2 minutes (T2). In the same fashion HA keep trying restarting the VM until issued power on attempt is reported as "completed".

A successful restart might never occur if the restart count is reached and all five restart attempts were unsuccessful.




**Q4: Explain does HA declares or determines that slave Esxi has isolated.**

Ans: Isolation of Esxi hosts are validated on the basis of heartbeats. The timeline for declaring isolation of slave and master Esxi is different. In this case we will discuss isolation of slave Esxi.

HA triggers a master election process before it will declare a slave Esxi host is isolated. In this timeline, "s" refers to seconds:

- 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

When an Esxi host is isolated, the value in “poweron” file is raised to 1, HA reads this file and validates that Esxi host has been isolated. There is one Poweron file per Esxi host and this file contains entries of all those VM’s which are currently powered on an Esxi host.

**Q5: Explain does HA declares or determines that master Esxi has isolated.**

Ans: In the case of the isolation of a master, this timeline is a bit less complicated because there is no need to go through an election process. In this timeline, “s” refers to seconds.

- T0 – Isolation of the host (master)
- T0 – Master pings “isolation addresses”
- T5s – Master declares itself isolated and “triggers” isolation response

**Q6: Is admission control policy is dependent on vCenter server and will admission control policy will work if your vCenter is not available.**

Ans: Yes admission control policy is dependent on vCenter Server although it is part of HA and we all knows HA works independently of vCenter Server. Admission control policies doesn’t work when at the time of failure of an Esxi host, vCenter server is not available. This doesn’t mean VM that were running on failed host will not be restarted, but whatever policy you have chosen that policy will not work.

For E.g.: You have chosen “Specify failover host” policy and dedicated one Esxi host for handling the failover. Now in normal scenario, if a host failure has occurred then HA will failover the failed VM’s on only this dedicated host and not on any other hosts in cluster. But if vCenter is not available and this happens then HA might restarts your VM’s on other hosts also if there are not sufficient resources available on your specified failover host.

**Q7: How does HA determines that Esxi host is network partitioned.**

Ans: There is a slight difference between Esxi host isolation and network partitioned. When multiple slave Esxi hosts has isolated together but they can ping each other than this condition is known as network partitioned.

For e.g.: Subnet mask of 5 Esxi has been changed then they will be unable to talk to master (being on different subnets) but they can communicate to each other (being on same subnet).

When network partitioned happens in a cluster then election happens between the isolated slaves Esxi and a new master is elected among them. In this case there will be 2 master in a cluster.

### Q8: How does HA determines which VM's it need to restart which were powered off or shutdown due to triggered isolation response?

Ans: If isolation response is set to "shut down" or "power off' then when an Esxi host is isolated, VM's are powered off or shutdown as a result of trigger of isolation response. Now the question is how HA keep tracks of which VM were powered off/shutdown due to this trigger.

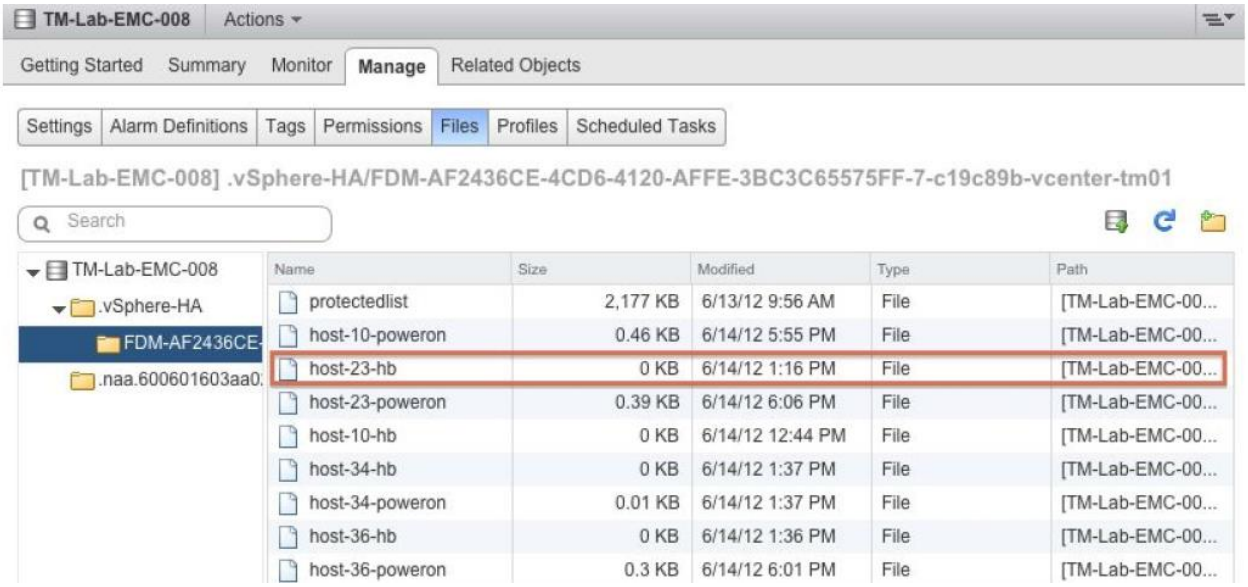
The answer to this question is as follows: When a VM is shutdown/powered off due to triggering of isolation response than the host that has isolated remove entries of those VM's from poweron file and creates a per virtual machine file inside a directory called "powered off". HA reads these files to identify the state change of the VM's and based on that it takes decision to restart those VM's.

This is necessary because, suppose when a host is isolated and at the same time if someone has manually issued a shutdown/powered off command to a VM, then HA will not restart that VM. There will be no file created for that VM by isolated host because it has been manually shut down.

### Q10: What are datastore heartbeats and how it is communicating or providing info to FDM that an Esxi host is alive or dead?

Ans: Datastore Heartbeat is nothing but just a file which is maintained in a reserved area called "Heartbeat Region" on every Esxi host and this file is updated every 5 seconds by the Esxi hosts. The master HA agent checks the timestamp of this file to check the host liveliness. If HA agent file find that this file is not updated in last 5 seconds then it comes to find out that there is some problem with that Esxi host. The naming convention of this file is as follows:

host-<number>-hb



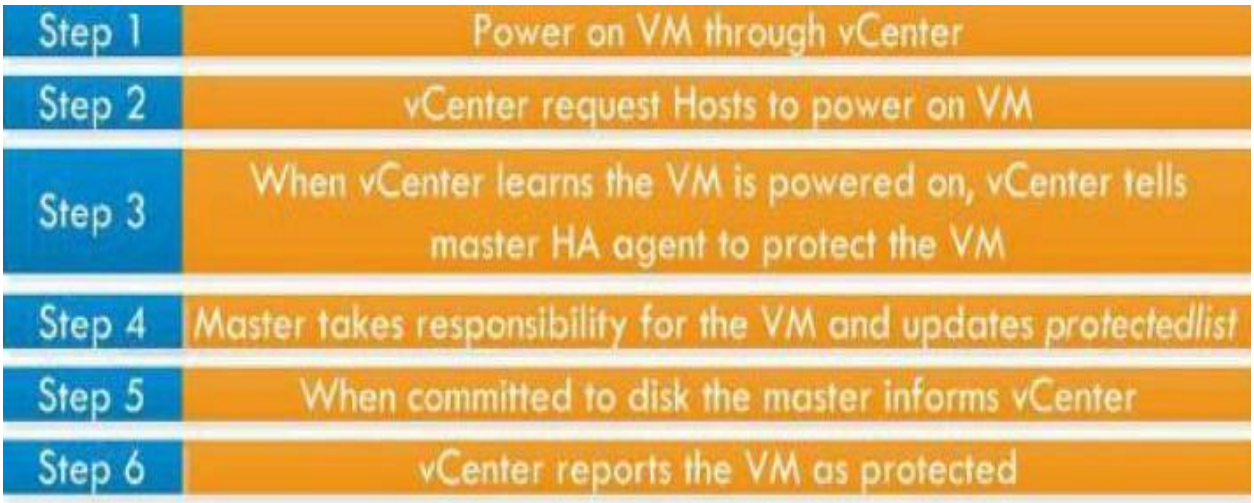

**Q11: How protection or unprotection of virtual machines is done by HA.**

Ans: When the state of a virtual machine changes, vCenter will direct the master to enable or disable HA protection for that virtual machine. Protection, however, is only guaranteed when the master has committed the change of state to disk.

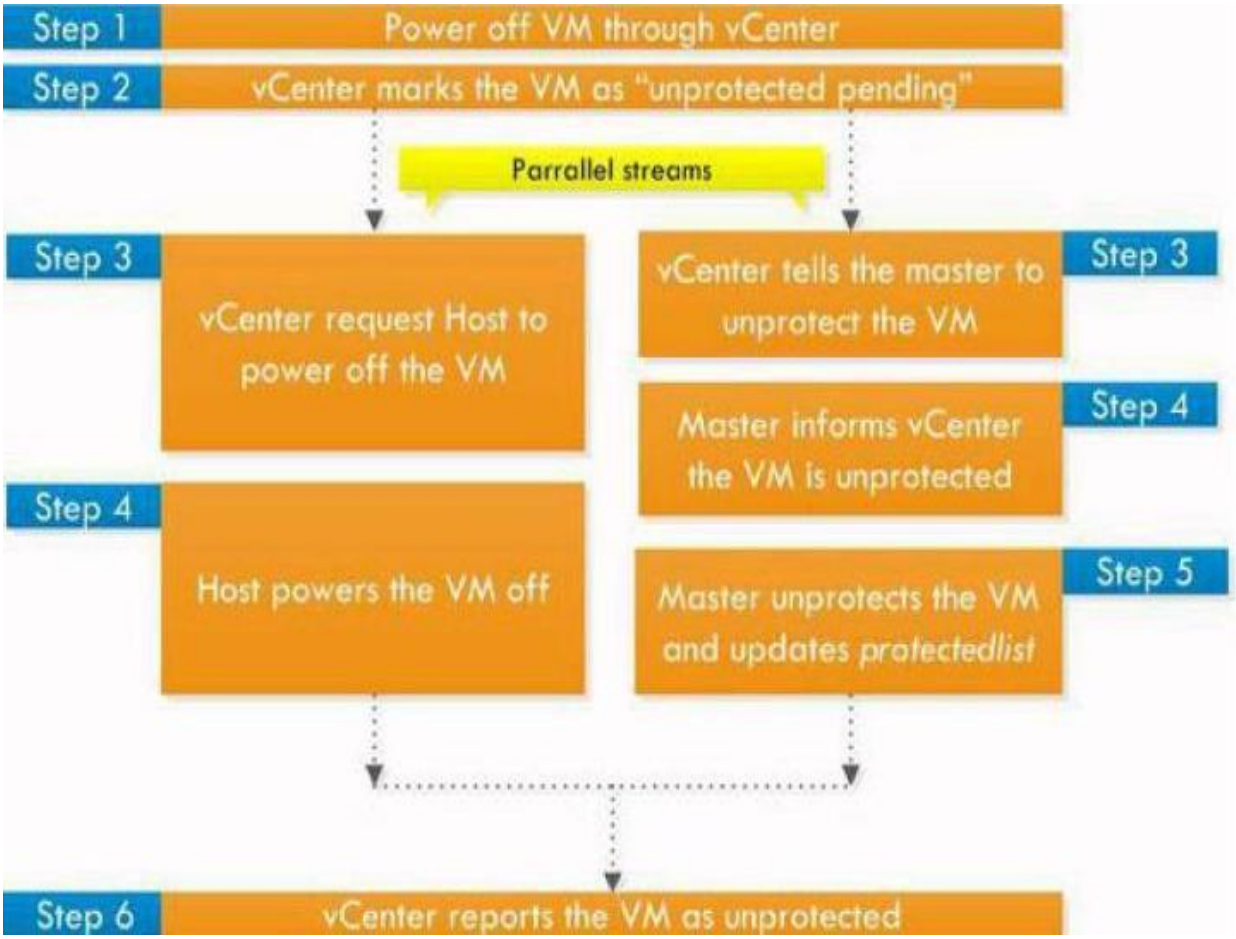
This state is distributed across the datastores and stored in the "Protectedlist" file. When the power state change of a virtual machine has been committed to disk, the master will inform vCenter Server so that the change in status is visible in vCenter.

When a VM is powered off it is removed from the "Protectedlist" file.

**Protection Overflow**



**Unprotection workflow**




**Q12: How does HA keep track of which VM are needed to be restart in case of an Esxi host failure?**

Ans: When an Esxi host fails, the VM's which were running on that Esxi are restarted on remaining nodes in the cluster. But how HA knows that how many VM's were running on the host before it has failed. The answer is:

HA takes help of 2 files namely "poweron" and "Protectedlist". The "poweron file is maintained by each Esxi host individually and it contains entries of those VM's which are currently running on that Esxi. The "Protectedlist" file is maintained at datastore level and tells HA that what were the VM's which were protected before the failure. On the basis of contents of these 2 files HA takes decision of restarting VM's.

When a VM is powered off manually then entry of that VM is removed from "Protectedlist" file so that HA do not accidently restart that VM also.

**Q13: Which parameter need to configure to increase the response time for isolation detection.**

Ans: You can configure a parameter called “das.isolationShutdown.Timeout”. The value of this parameter is specified in minutes and it is time which will be taken by HA to gracefully shutdown a VM when isolation response is set to “Shutdown VM” and it is triggered.

**Q14: What are the conditions when election of master takes place in a cluster?**

Ans: A master is elected by a set of HA agents whenever the agents are not in network contact with a master. A master election thus occurs when HA is first enabled on a cluster and when the host on which the master is running:

- fails,
- becomes network partitioned or isolated,
- is disconnected from vCenter Server,
- is put into maintenance or standby mode,
- or when HA is reconfigured on the host.

Note: Removing slave Esxi from a cluster doesn’t have any effect on election process i.e. if any slave Esxi is removed or shutdown or put into maintenance mode, election will not happen.

**Q15: What will happen when election of master is going on in a cluster and at the same time one of the slave Esxi host also failed? How this failure will be handled since at the time of failure there is no master Esxi host.**

Ans: It is mandatory that for restarting VM's master should be present in cluster. Now when election is happening in a cluster, it takes 15 seconds to complete the election process. Now during that time if a slave Esxi also fails then restart of VM has to wait until election process is completed.

The newly elected master will first read the "Protected List" file to find out the VM's whose power state has been changed. After reading that file it will decide that how many vm's were there which failed during election time and then will perform restart of those VM's.

**Q16: What are the things which HA takes into account before restarting VM’s?**

Ans: HA has to take many things into considerations before restarting VM’s in case of Esxi failure. These includes:

1. CPU and memory reservation including memory overhead.
2. Unreserved capacity of host in cluster
3. Restart priority of VM

4. VM to host compatibility
5. Number of dvPorts required by VM and number of dvPorts that are available
6. Max no vCPU & VM that can be run on a given host.
7. Restart latency

**Q17: What will happen if a VM fails when SvMotion was going on that VM and has not been completed yet? How this failure will be handled by HA?**

Ans: If a virtual machine needs to be restarted by HA and the virtual machine is in the process of being Storage vMotioned and the virtual machine fails, the restart process is not started until vCenter informs the master that the Storage vMotion task has completed or has been rolled back.

**Q18: Will master election happen if a new Esxi that has visibility to more datastores than existing master is introduced in a cluster?**

Ans: No election will not happen even if the newly introduced Esxi has visibility to more datastores than master Esxi host. But if you reconfigure HA on the cluster then the newly added Esxi will become master because it is connected to more number of datastores.

**Q20: If a slave Esxi has been removed from a cluster then will election be triggered again?**

Ans: No removal of slave Esxi from cluster doesn't has any impact on master. No election will be happening in this case.

**Q19: Does HA seeks assistance from DRS before starting failover of failed VM's?**

Ans: Yes HA do takes assistance from DRS sometimes before starting the failover of failed VM's. If a cluster is configured with admission control policies and either "specify number of host failure cluster tolerates" or "percentage" based policy is used then sometimes it may happen that resources are not fully available on single host and is scattered throughout the cluster. In that case HA will wait before performing failover of VM's and ask assistance of HA to defragment the resources.

## Scenario Based Questions?

**Q1: Suppose that one of the slave Esxi host has been failed and HA is trying to restart the VM's that were on the failed host. For one particular VM 3 restart attempt has been already made and during the 4th restart attempt master fails itself. Now how the restart of this VM will be handled.**

Ans: The restart count will be reset to zero if master fails when it is in process of attempting restarts of failed VM's. This means again 5 attempts of restart can be made on VM.

**Q2: If the admission control policy is set to specify failover host and vCenter is not available at the time of one of the Esxi failure. What will HA do now? Will it still restarts vm on the specified failover host or it will distribute restarting of vm among all the Esxi hosts in the cluster.**

Ans: HA will restart VM's on designated failover host. If designated failover host is incapable of accommodating all VM's then HA will start restart remaining VM's on other nodes also.

**Q3: When "Network Partition" situation occurs in a cluster then there will be more than 2 masters in cluster. Now when this partitioning is aligned then what will happen? Again election process will be started or old master will continue to govern the cluster?**

Ans: When "Network Partition" problem is resolved then all the Esxi host will again come in contact with each other. But master election will not happen. Old master will be continue governing the cluster.

Diclaimer: Concepts and graphics has been taken from "Clustering Deep Dive" book written by Frank and Duncan Epping. Million thanks to them for such a wonderful book.