Interview Questions on Red Hat Cluster

How can you define a cluster and what are its basic types?

A cluster is two or more computers (called nodes or members) that work together to perform a task. There are four major types of clusters:

- Storage
- High availability
- Load balancing
- High performance

What is Storage Cluster?

- Storage clusters provide a consistent file system image across servers in a cluster, allowing the servers to simultaneously read and write to a single shared file system.
- A storage cluster simplifies storage administration by limiting the installation and patching of applications to one file system.
- The High Availability Add-On provides storage clustering in conjunction with Red Hat GFS2

What is High Availability Cluster?

- High availability clusters provide highly available services by eliminating single points of failure and by failing over services from one cluster node to another in case a node becomes inoperative.
- Typically, services in a high availability cluster read and write data (via read-write mounted file systems).
- A high availability cluster must maintain data integrity as one cluster node takes over control of a service from another cluster node.
- Node failures in a high availability cluster are not visible from clients outside the cluster.
- High availability clusters are sometimes referred to as failover clusters.

What is Load Balancing Cluster?

- Load-balancing clusters dispatch network service requests to multiple cluster nodes to balance the request load among the cluster nodes.
- Load balancing provides cost-effective scalability because you can match the number of nodes according to load requirements. If a node in a load-balancing cluster becomes inoperative, the load-balancing software detects the failure and redirects requests to other cluster nodes.
- Node failures in a load-balancing cluster are not visible from clients outside the cluster.
- Load balancing is available with the Load Balancer Add-On.

What is a High Performance Cluster?

- High-performance clusters use cluster nodes to perform concurrent calculations.
- A high-performance cluster allows applications to work in parallel, therefore enhancing the performance of the applications.
- High performance clusters are also referred to as computational clusters or grid computing.

How many nodes are supported in Red Hat 6 Cluster?

A cluster configured with qdiskd supports a maximum of 16 nodes. The reason for the limit is because of scalability; increasing the node count increases the amount of synchronous I/O contention on the shared quorum disk device.

What is the minimum size of the Quorum Disk?

The minimum size of the block device is 10 Megabytes.

What is the order in which you will start the Red Hat Cluster services?

In Red Hat 4

service ccsd start

service cman start

service fenced start

service clvmd start (If CLVM has been used to create clustered volumes)

service gfs start

service rgmanager start

In RedHat 5

service cman start

service clvmd start

service gfs start

service rgmanager start

In Red Hat 6

service cman start

service clvmd start

service gfs2 start

service rgmanager start

What is the order to stop the Red Hat Cluster services?

In Red Hat 4

service rgmanager stop

service gfs stop

service clvmd stop

service fenced stop

service cman stop

service ccsd stop

In Red Hat 5

service rgmanager stop

service gfs stop

service clvmd stop

service cman stop

In Red Hat 6

service rgmanager stop

service gfs2 stop

service clvmd stop

service cman stop

What are the performance enhancements in GFS2 as compared to GFS?

- Better performance for heavy usage in a single directory
- Faster synchronous I/O operations
- Faster cached reads (no locking overhead)
- Faster direct I/O with preallocated files (provided I/O size is reasonably large, such as 4M blocks)
- Faster I/O operations in general
- Faster Execution of the df command, because of faster statfs calls
- Improved atime mode to reduce the number of write I/O operations generated by atime when compared with GFS
- GFS2 supports the following features.
- extended file attributes (xattr)
- the lsattr() and chattr() attribute settings via standard ioctl() calls
- nanosecond timestamps
- GFS2 uses less kernel memory.
- GFS2 requires no metadata generation numbers.

- Allocating GFS2 metadata does not require reads. Copies of metadata blocks in multiple journals are managed by revoking blocks from the journal before lock release.
- GFS2 includes a much simpler log manager that knows nothing about unlinked inodes or quota changes.
- The gfs2_grow and gfs2_jadd commands use locking to prevent multiple instances running at the same time.
- The ACL code has been simplified for calls like creat() and mkdir().
- Unlinked inodes, quota changes, and statfs changes are recovered without remounting the journal.

What is the maximum file system support size for GFS2?

- GFS2 is based on 64 bit architecture, which can theoretically accommodate an 8 EB file system.
- However, the current supported maximum size of a GFS2 file system for 64-bit hardware is 100 TB.
- The current supported maximum size of a GFS2 file system for 32-bit hardware for Red Hat Enterprise Linux Release 5.3 and later is 16 TB.
- **NOTE:** It is better to have 10 1TB file systems than one 10TB file system.

What is the journaling filesystem?

- A journaling filesystem is a filesystem that maintains a special file called a journal that is used to repair any inconsistencies that occur as the result of an improper shutdown of a computer.
- In journaling file systems, every time GFS2 writes metadata, the metadata is committed to the journal before it is put into place.
- This ensures that if the system crashes or loses power, you will recover all of the metadata when the journal is automatically replayed at mount time.
- GFS2 requires one journal for each node in the cluster that needs to mount the file system. For example, if you have a 16-node cluster but need to mount only the file system from two nodes, you need only two journals. If you need to mount from a third node, you can always add a journal with the gfs2_jadd command.

What is the default size of journals in GFS?

When you run mkfs.gfs2 without the size attribut for journal to create a GFS2 partition, by default a 128MB size journal is created which is enough for most of the applications

In case you plan on reducing the size of the journal, it can severely affect the performance. Suppose you reduce the size of the journal to 32MB it does not take much file system activity to fill an 32MB journal, and when the journal is full, performance slows because GFS2 has to wait for writes to the storage.

What is a Quorum Disk?

- Quorum Disk is a disk-based quorum daemon, qdiskd, that provides supplemental heuristics to determine node fitness.
- With heuristics you can determine factors that are important to the operation of the node in the event of a network partition

For a 3 node cluster a quorum state is present untill 2 of the 3 nodes are active i.e. more than half. But what if due to some reasons the 2nd node also stops communicating with the the 3rd node? In that case under a normal architecture the cluster would dissolve and stop working. But for mission critical environments and such scenarios we use quorum disk in which an additional disk is configured which is mounted on all the nodes with qdiskd service running and a vote value is assigned to it.

So suppose in above case I have assigned 1 vote to qdisk so even after 2 nodes stops communicating with 3rd node, the cluster would have 2 votes (1 qdisk + 1 from 3rd node) which is still more than half of vote count for a 3 node cluster. Now both the inactive nodes would be fenced and your 3rd node would be still up and running being a part of the cluster.

What is rgmanager in Red Hat Cluster and its use?

- This is a service termed as Resource Group Manager
- RGManager manages and provides failover capabilities for collections of cluster resources called services, resource groups, or resource trees
- it allows administrators to define, configure, and monitor cluster services. In the event of a node failure, rgmanager will relocate the clustered service to another node with minimal service disruption

What is luci and ricci in Red Hat Cluster?

- **luci** is the server component of the Conga administration utility
- Conga is an integrated set of software components that provides centralized configuration and management of Red Hat clusters and storage
- luci is a server that runs on one computer and communicates with multiple clusters and computers via ricci
-
- **ricci** is the client component of the Conga administration utility
- ricci is an agent that runs on each computer (either a cluster member or a standalone computer) managed by Conga
- This service needs to be running on all the client nodes of the cluster.

What is cman in Red Hat Cluster?

- This is an abbreviation used for Cluster Manager.
- CMAN is a distributed cluster manager and runs in each cluster node.
- It is responsible for monitoring, heartbeat, quorum, voting and communication between cluster nodes.
- CMAN keeps track of cluster quorum by monitoring the count of cluster nodes.

What are the different port no. used in Red Hat Cluster?

IP Port no.	Protocol	Component
5404,5405	UDP	corosync/cman
11111	TCP	ricci
21064	TCP	dlm (Distributed Lock Manager)
16851	TCP	Modclustered
8084	TCP	luci
4196,4197	TCP	rgmanager

How does NetworkManager service affects Red Hat Cluster?

- The use of NetworkManager is not supported on cluster nodes. If you have installed NetworkManager on your cluster nodes, you should either remove it or disable it.
- # service NetworkManager stop
- # chkconfig NetworkManager off
- The cman service will not start if NetworkManager is either running or has been configured to run with the chkconfig command

What is the command used to relocate a service to another node?

clusvcadm -r service_name -m node_name

What is split-brain condition in Red Hat Cluster?

- We say a cluster has quorum if a majority of nodes are alive, communicating, and agree on the active cluster members. For example, in a thirteen-node cluster, quorum is only reached if seven or more nodes are communicating. If the seventh node dies, the cluster loses quorum and can no longer function.
- A cluster must maintain quorum to prevent split-brain issues.
- If quorum was not enforced, quorum, a communication error on that same thirteen-node cluster may cause a situation where six nodes are operating on the shared storage, while another six nodes are also operating on it, independently. Because of the communication error, the two partial-clusters would overwrite areas of the disk and corrupt the file system.
- With quorum rules enforced, only one of the partial clusters can use the shared storage, thus protecting data integrity.
- Quorum doesn't prevent split-brain situations, but it does decide who is dominant and allowed to function in the cluster.
- quorum can be determined by a combination of communicating messages via Ethernet and through a quorum disk.

What are Tie-breakers in Red Hat Cluster?

- Tie-breakers are additional heuristics that allow a cluster partition to decide whether or not it is quorate in the event of an even-split - prior to fencing.
- With such a tie-breaker, nodes not only monitor each other, but also an upstream router that is on the same path as cluster communications. If the two nodes lose contact with each other, the one that wins is the one that can still ping the upstream router.That is why, even when using tie-breakers, it is important to ensure that fencing is configured correctly.
- CMAN has no internal tie-breakers for various reasons. However, tie-breakers can be implemented using the API.

What is fencing in Red Hat Cluster?

- Fencing is the disconnection of a node from the cluster's shared storage.
- Fencing cuts off I/O from shared storage, thus ensuring data integrity.
- The cluster infrastructure performs fencing through the fence daemon, **fenced**.
- When CMAN determines that a node has failed, it communicates to other cluster-infrastructure components that the node has failed.
- **fenced**, when notified of the failure, fences the failed node.

What are the various types of fencing supported by High Availability Add On?

- **Power fencing** — A fencing method that uses a power controller to power off an inoperable node.
- **storage fencing** — A fencing method that disables the Fibre Channel port that connects storage to an inoperable node.
- **Other fencing** — Several other fencing methods that disable I/O or power of an inoperable node, including IBM Bladecenters, PAP, DRAC/MC, HP ILO, IPMI, IBM RSA II, and others.

What are the lock states in Red Hat Cluster?

A lock state indicates the current status of a lock request. A lock is always in one of three states:

Granted — The lock request succeeded and attained the requested mode.

Converting — A client attempted to change the lock mode and the new mode is incompatible with an existing lock.

Blocked — The request for a new lock could not be granted because conflicting locks exist.

A lock's state is determined by its requested mode and the modes of the other locks on the same resource.

What is DLM lock model?

- DLM is a short abbreviation for Distributed Lock Manager.
- A lock manager is a traffic cop who controls access to resources in the cluster, such as access to a GFS file system.
- GFS2 uses locks from the lock manager to synchronize access to file system metadata (on shared storage)
- CLVM uses locks from the lock manager to synchronize updates to LVM volumes and volume groups (also on shared storage)
- In addition, rgmanager uses DLM to synchronize service states.
- without a lock manager, there would be no control over access to your shared storage, and the nodes in the cluster would corrupt each other's data.