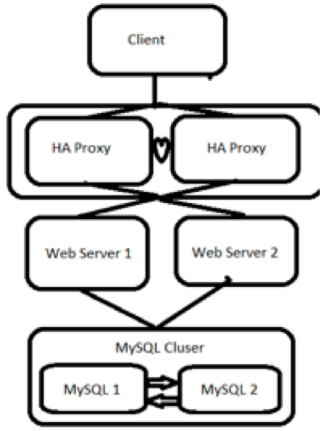
High Availability Clusters

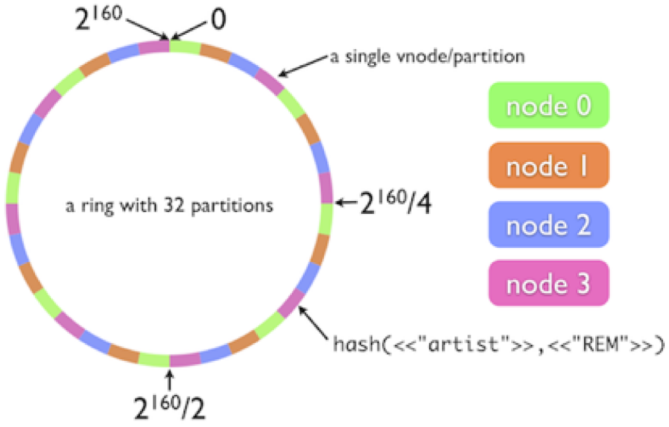


Question:
If a load balancer is used - isn't that the single point of failure?

Answer:
Yes, it will become a single point of failure if it is a single device.
So to avoid this, we need to use more than one balancer,
and we need to run them as a HA (High Availability) cluster.
- https://en.wikipedia.org/wiki/High-availability_cluster

There are multiple ways to achieve High Availability.
For example, consider simple fail-over mechanism with just two servers (master/slave).
They receive the same input, do the same calculations, constantly in sync, but only the master provides the output.
The slave server monitors the heartbeat of the master. If the master server dies (heartbeat stops), the slave server becomes the new master.

In the above simple master/slave architecture, the slave is mostly doing nothing.
In real life you may have many servers receiving same inputs and separating their responsibilities to achieve higher performance. Then if one of the servers dies, others can take over its responsibilities.
Common method of separating responsibilities (and doing fail-over) is called "**consistent hashing**".
- https://en.wikipedia.org/wiki/Consistent_hashing
Good explanation:
- https://dzone.com/articles/simple-magic-consistent
Original paper (1997):
- Consistent Hashing and Random Trees: Distributed Caching Protocols for Relieving Hot Spots on the World Wide Web – by David Karger, Eric Lehman, Tom Leighton, Matthew Levine, Daniel Lewin, Rina Panigrahy.



Consistent Hashing