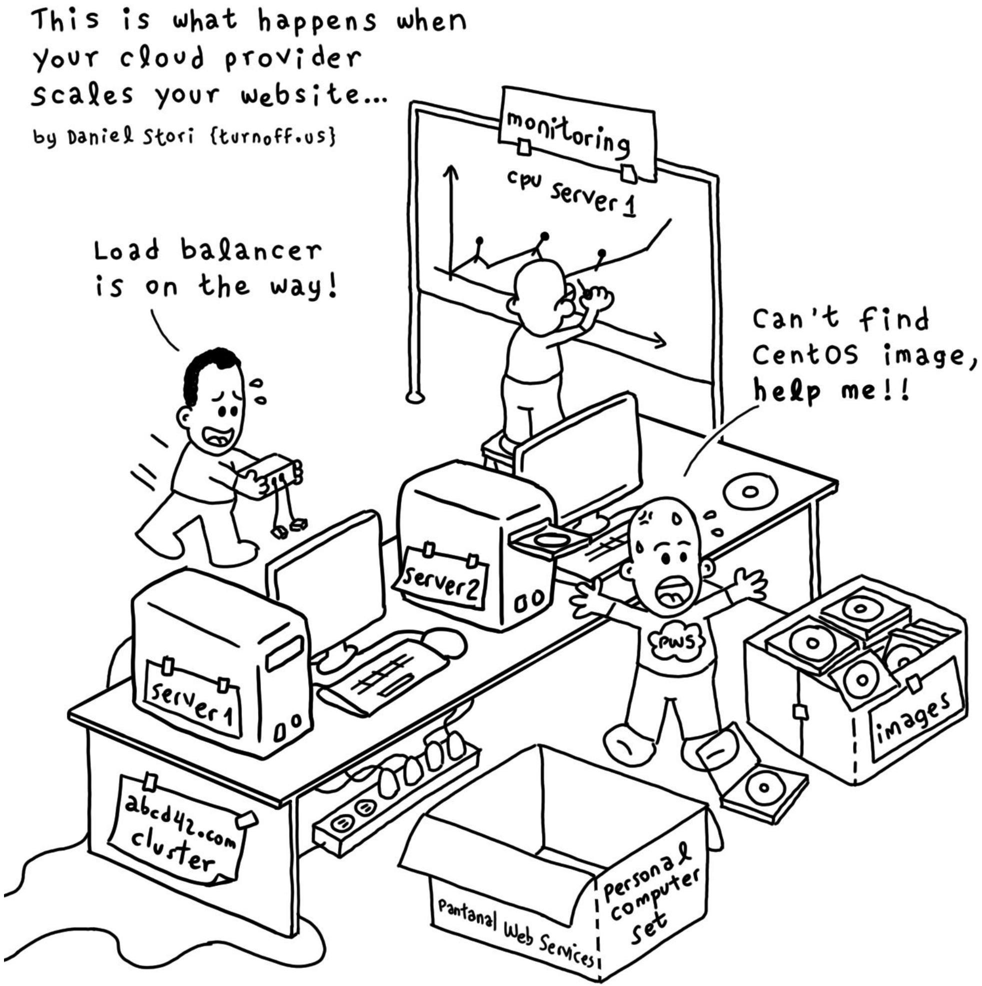
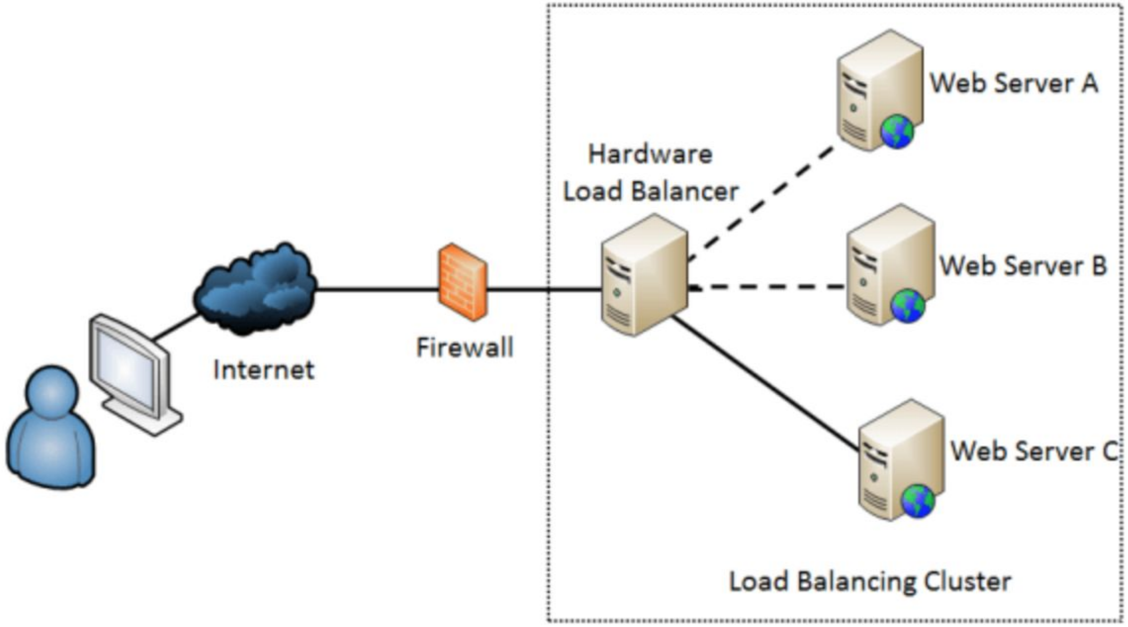
Waiting Room Fun :


This is what happens when your cloud provider scales your website...
by Daniel Stori {turnoff.us}
monitoring
cpu server 1
Load balancer is on the way!
Can't find CentOS image, help me!!
server2
server1
PWS
images
abcd42.com cluster
Pantanal Web Services
Personal computer set

Resource Management

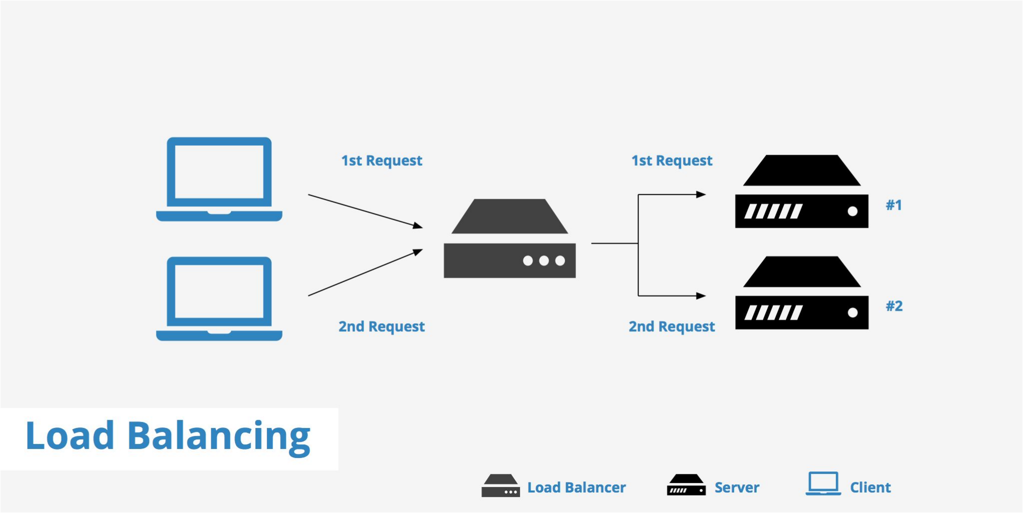
Source : Digital Ocean , avitnetwork

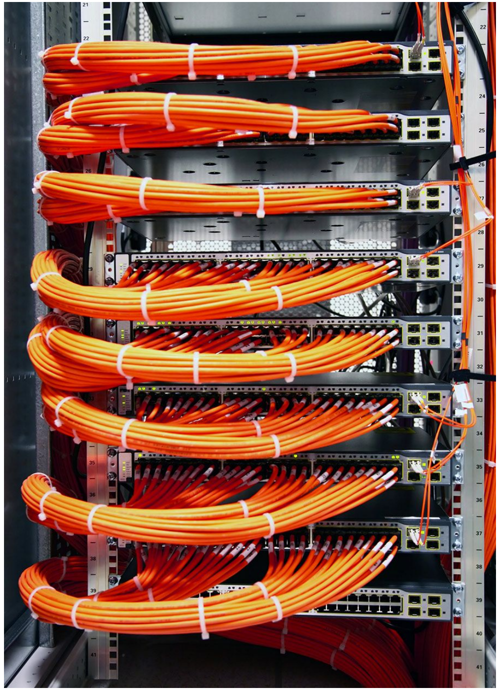
Asst. Prof. Ashwini Mathur



History of Load Balancing

Load balancing got its start in the 1990s as **hardware appliances distributing traffic across a network.** Organizations wanted to improve accessibility of applications running on servers.





What is Load Balancing ?

Load Balancing Definition:

Load balancing is the process of distributing network traffic across multiple servers.

This ensures no single server bears too much demand. By spreading the work evenly, load balancing improves application responsiveness. It also increases availability of applications and websites for users.

Modern applications cannot run without load balancers.

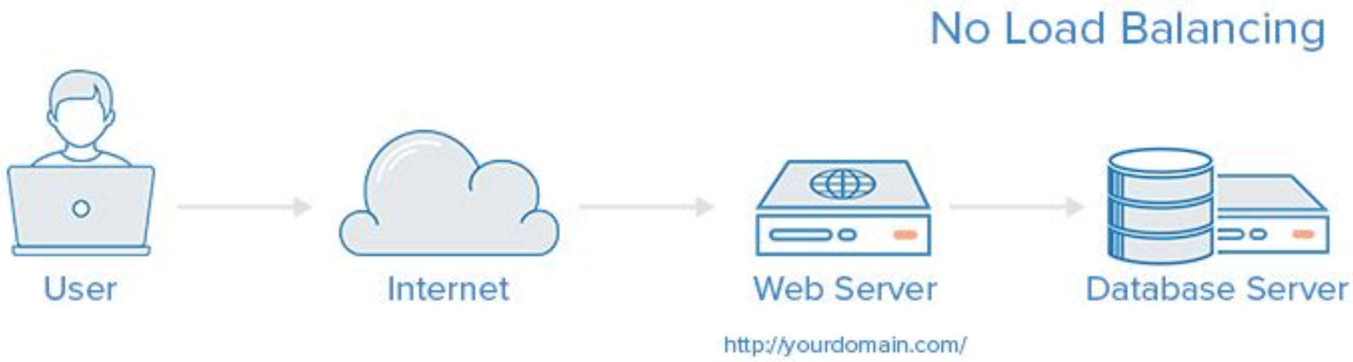
What is a load balancer?

The mechanism or component which performs the load balancing operation is called a load balancer.

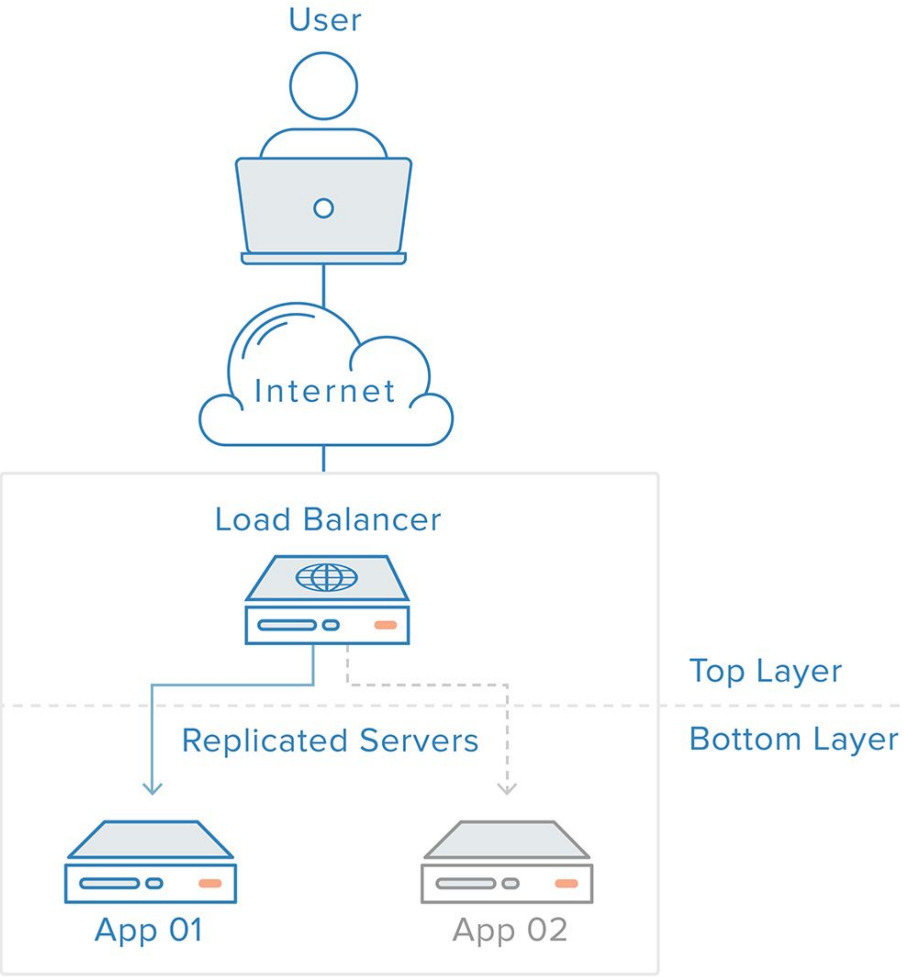
A load balancer may act at:

1. Link level : This is called link load balancing, and it consists in chosing what network link to send a packet to.
2. Network level : This is called network load balancing, and it consists in chosing what route a series of packets will follow.
3. Server level : this is called server load balancing and it consists in deciding what server will process a connection or request.

In this example, the user connects directly to the web server, at *yourdomain.com*. If this single web server goes down, the user will no longer be able to access the website.

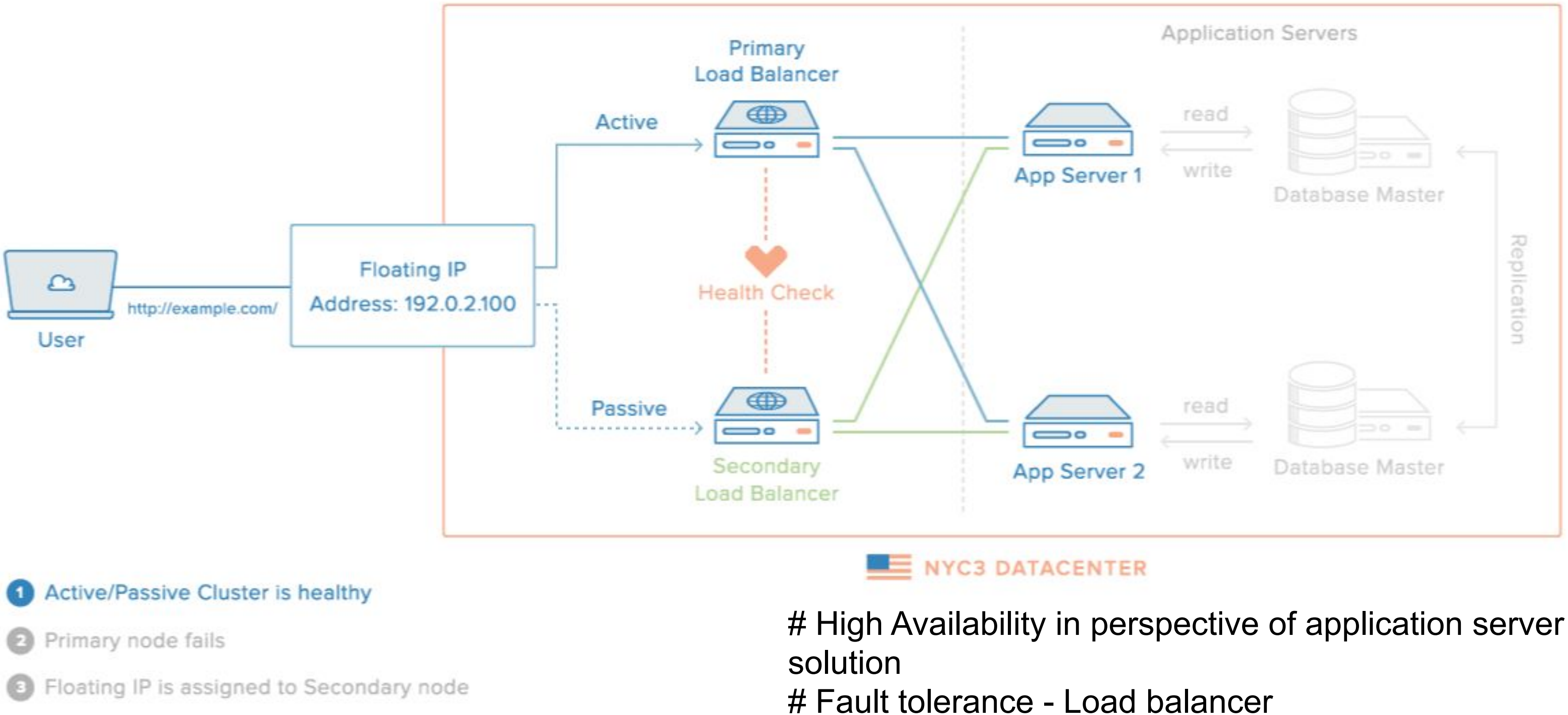


With Load Balancing



In the example illustrated above, the user accesses the load balancer, which forwards the user's request to a backend server, which then responds directly to the user's request.

In this scenario, the single point of failure is now the load balancer itself. This can be mitigated by introducing a second load balancer, but before we discuss that, let's explore how load balancers work.



1 Active/Passive Cluster is healthy

2 Primary node fails

3 Floating IP is assigned to Secondary node

# High Availability in perspective of application server solution

# Fault tolerance - Load balancer

Floating IP concept ...

A floating IP is usually a **public, routable IP address** that is not automatically assigned to an entity. Instead, a project owner assigns them to one or more entities temporarily. The respective entity has an automatically assigned, static IP for communication between instances in a private, non-routable network area, as well as via a manually assigned floating IP. This makes the entity’s services outside a cloud or network **recognizable and therefore achievable**.

In appropriately configured failover scenarios, an IP 'floats' to another active unit in the network so that it can take on the function of a dormant entity **without a time delay**, and can then answer incoming requests.

Load Balancing Algorithms

There is a variety of load balancing methods, which use different algorithms best suited for a particular situation.

- Least Connection Method — directs traffic to the server with the fewest active connections. Most useful when there are a large number of persistent connections in the traffic unevenly distributed between the servers.
- Least Response Time Method — directs traffic to the server with the fewest active connections and the lowest average response time.
- Round Robin Method — rotates servers by directing traffic to the first available server and then moves that server to the bottom of the queue. Most useful when servers are of equal specification and there are not many persistent connections.
- IP Hash — the IP address of the

Vsphere Resource Management Tools

DRS (Distributed Resource Scheduler)

VMware vSphere Distributed Resource Scheduler (DRS) is a feature that enables a virtual environment to automatically balance itself across your ESX hosts in a cluster in an effort to eliminate resource contention. The goals of DRS are:

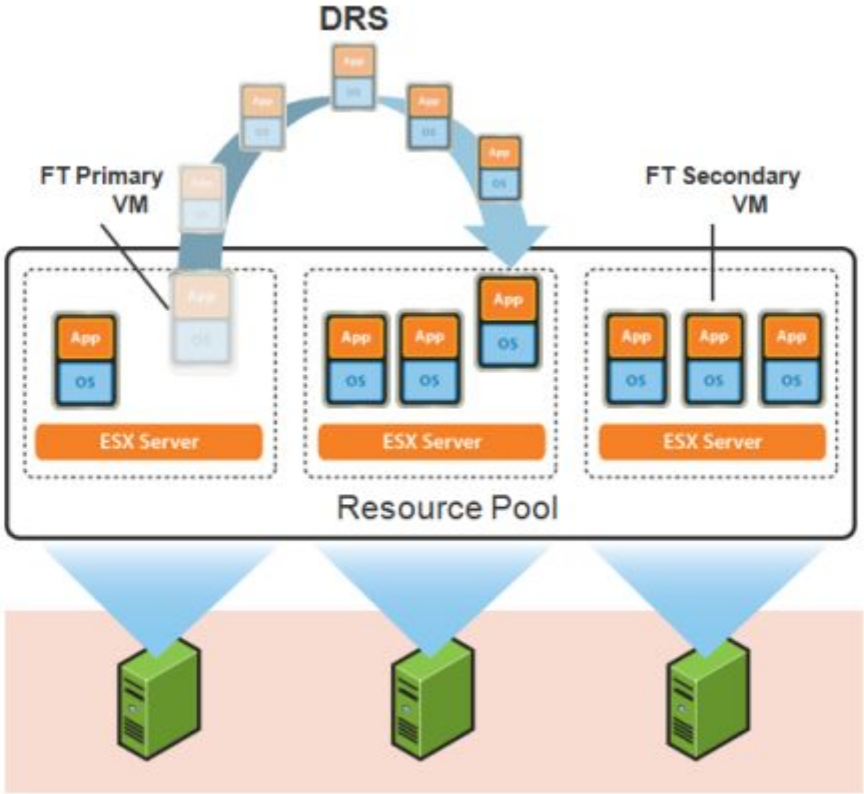
- at startup, DRS attempts to place each VM on the host that is best suited to run that virtual machine.
- while a VM is running, DRS seeks to provide that VM with the required hardware resources while minimizing the amount of contention for those resources in an effort to maintain balanced utilization levels.

If a DRS cluster becomes unbalanced, DRS can migrate VMs from **overutilized ESXi hosts to underutilized hosts**.

DRS performs these migrations of VMs across hosts in the cluster without any downtime by using vMotion. You can determine whether DRS will just display migration recommendations or automatically perform the migration when the cluster becomes unbalanced by defining the automation level.

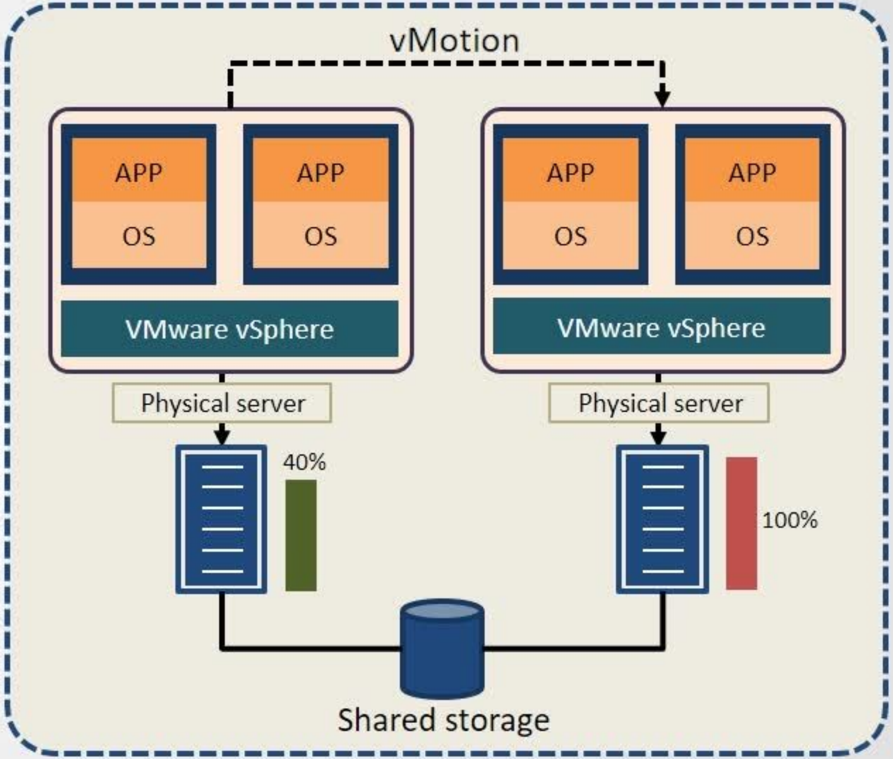
Before using vSphere DRS, the following requirements must be met:

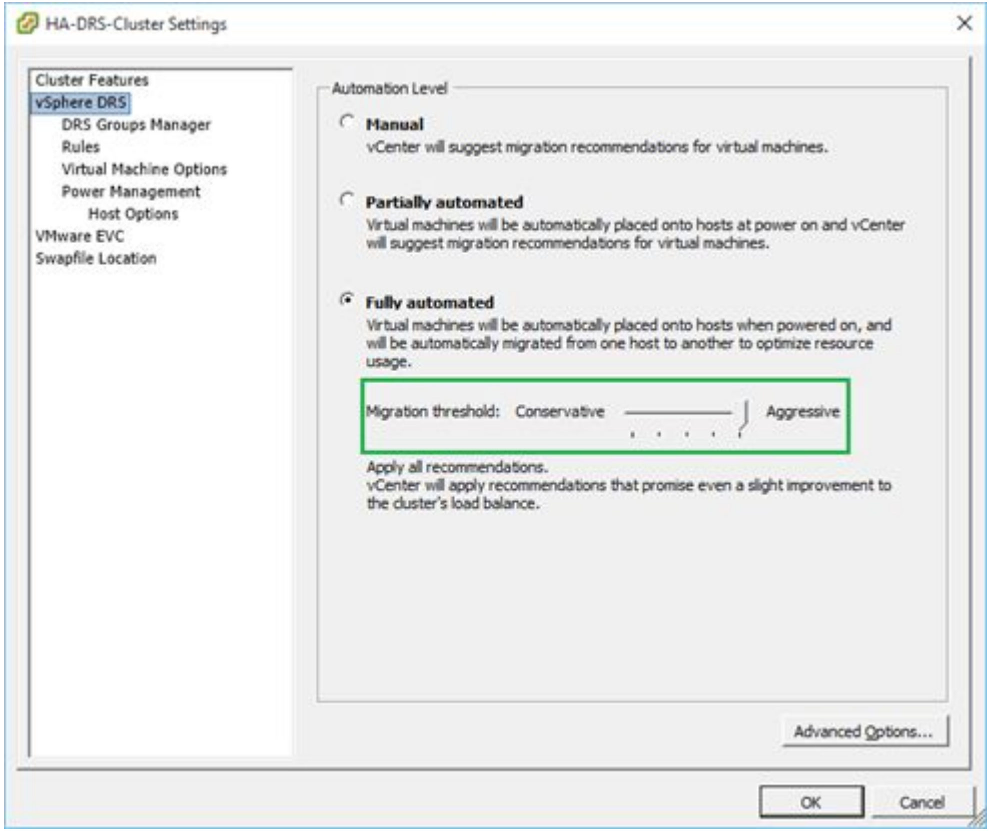
- vCenter Server needs to be installed.
- CPUs in ESXi hosts must be compatible.
- to use DRS for load balancing, hosts in the DRS cluster must be part of a vMotion migration network.
- all hosts should use shared storage, with volumes accessible by all hosts.
- shared storage needs to be large enough to store all virtual disks for the VM.
- DRS works best if the VMs meet vSphere vMotion requirements


DRS
FT Primary VM
FT Secondary VM
App
OS
ESX Server
ESX Server
ESX Server
Resource Pool

Distributed Resource Scheduler (DRS) Basics

- Monitors cluster resources
- Ensures VMs have needed resources
- Monitors ongoing VM performance
- Makes VM migration decisions
- Performs load balancing function





Automation Level

○ **Manual**
vCenter will suggest migration recommendations for virtual machines.

○ **Partially automated**
Virtual machines will be automatically placed onto hosts at power on and vCenter will suggest migration recommendations for virtual machines.

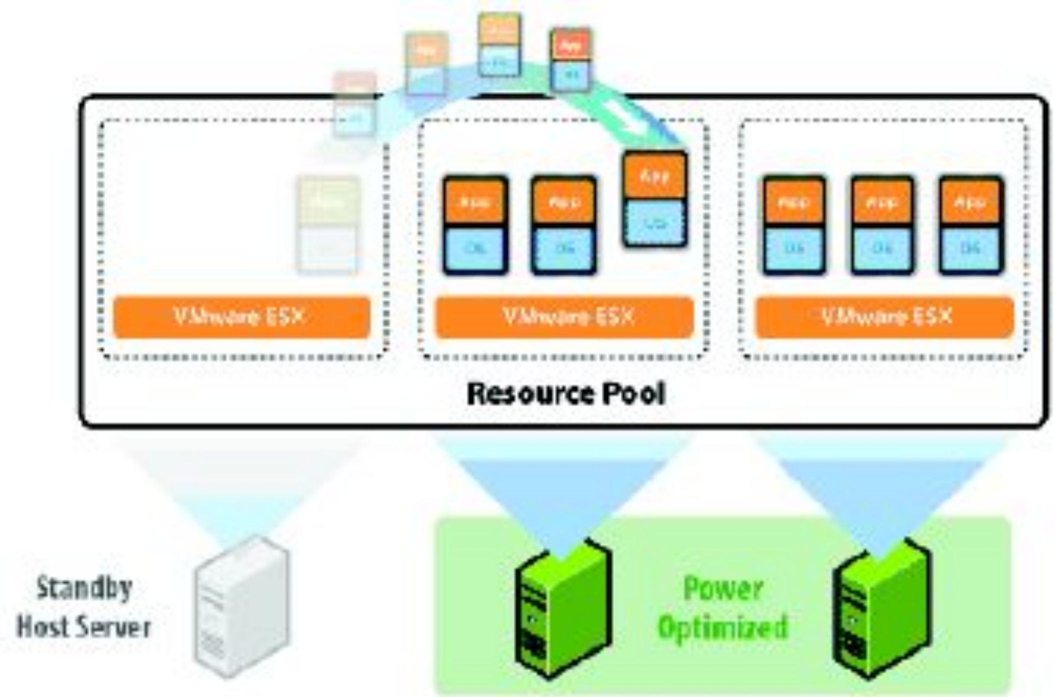
● **Fully automated**
Virtual machines will be automatically placed onto hosts when powered on, and will be automatically migrated from one host to another to optimize resource usage.

Migration threshold: Conservative ———————| Aggressive

Apply all recommendations.
vCenter will apply recommendations that promise even a slight improvement to the cluster's load balance.

https://docs.google.com/document/d/1-d3RdzTnBd7cqDL1RfhATb3Js61Ayq0VssDGvk2vq3w/edit?usp=sharing

DPM - Distributed Power Management


App
OS
VMware ESX
VMware ESX
VMware ESX
Resource Pool
Standby
Host Server
Power
Optimized

Overview

Consolidation of physical servers into virtual machines that share host physical resources can result in significant reductions in the costs associated with hardware maintenance and power consumption.

VMware Distributed Power Management (VMware DPM) provides **additional power savings** beyond this initial benefit by dynamically consolidating workloads even further during periods of low resource utilization. Virtual machines are migrated onto fewer hosts and the **unneeded ESX hosts are powered off**.