负载均衡

为什么需要负载均衡

随着网络业务量的提高，访问量的增加，设备的处理量也相应增大，从而使得设备的负荷加大。从资源上考虑，企业又不能把现有的设备撇弃，因此，负载均衡机制应运而生。

从单机网站到分布式网站，很重要的区别是业务拆分和分布式部署，将应用拆分后，部署到不同的机器上，实现大规模分布式系统。分布式和业务拆分解决了，从集中到分布的问题，但是每个部署的独立业务还存在单点的问题和访问统一入口问题，为解决单点故障，我们可以采取冗余的方式。将相同的应用部署到多台机器上。解决访问统一入口问题，我们可以在集群前面增加负载均衡设备，实现流量分发。

负载均衡（Load Balance），意思是将负载（工作任务，访问请求）进行平衡、分摊到多个操作单元（服务器，组件）上进行执行。是解决高性能，单点故障（高可用），扩展性（水平伸缩）的终极解决方案。

负载均衡的原理

系统的扩展可分为纵向（垂直）扩展和横向（水平）扩展。纵向扩展，是从单机的角度通过增加硬件处理能力，比如CPU处理能力，内存容量，磁盘等方面，实现服务器处理能力的提升，不能满足大型分布式系统（网站），大流量，高并发，海量数据的问题。因此需要采用横向扩展的方式，通过添加机器来满足大型网站服务的处理能力。比如：一台机器不能满足，则增加两台或者多台机器，共同承担访问压力。这就是典型的集群和负载均衡架构：如下图：

应用集群：将同一应用部署到多台机器上，组成处理集群，接收负载均衡设备分发的请求，进行处理，并返回相应数据。 负载均衡设备：将用户访问的请求，根据负载均衡算法，分发到集群中的一台处理服务器。（一种把网络请求分散到一个服务器集群中的可用服务器上去的设备） 负载均衡的作用（解决的问题）：

1.解决并发压力，提高应用处理性能（增加吞吐量，加强网络处理能力）；

2.提供故障转移，实现高可用；

3.通过添加或减少服务器数量，提供网站伸缩性（扩展性）；

4.安全防护；（负载均衡设备上做一些过滤，黑白名单等处理）

负载均衡分类

DNS负载均衡

最早的负载均衡技术，利用域名解析实现负载均衡，在DNS服务器，配置多个A记录，这些A记录对应的服务

器构成集群。大型网站总是部分使用DNS解析，作为第一级负载均衡

优点

使用简单：负载均衡工作，交给DNS服务器处理，省掉了负载均衡服务器维护的麻烦 提高性能：可以支持基于地址的域名解析，解析成距离用户最近的服务器地址，可以加快访问速度，改善性能；

缺点

可用性差：DNS解析是多级解析，新增/修改DNS后，解析时间较长；解析过程中，用户访问网站将失败；扩展性低：DNS负载均衡的控制权在域名商那里，无法对其做更多的改善和扩展； 维护性差：也不能反映服务器的当前运行状态；支持的算法少；不能区分服务器的差异（不能根据系统与服务的状态来判断负载）

IP负载均衡

在网络层通过修改请求目标地址进行负载均衡。

用户请求数据包，到达负载均衡服务器后，负载均衡服务器在操作系统内核进程获取网络数据包，根据负载均衡算法得到一台真实服务器地址，然后将请求目的地址修改为，获得的真实ip地址，不需要经过用户进程处理。 真实服务器处理完成后，响应数据包回到负载均衡服务器，负载均衡服务器，再将数据包源地址修改为自身的ip地址

IP负载均衡，真实物理服务器返回给负载均衡服务器，存在两种方式： 1. 负载均衡服务器在修改目的ip地址的同时修改源地址。将数据包源地址设为自身盘，即源地址转换（snat） 2. 将负载均衡服务器同时作为真实物理服务器集群的网关服务器。

优点:

在内核进程完成数据分发，比在应用层分发性能更好；

缺点:

所有请求响应都需要经过负载均衡服务器，集群最大吞吐量受限于负载均衡服务器网卡带宽；

链路层负载均衡 混合型P负载均衡

负载均衡算法

1. 轮询
2. 随机
3. Hash（源地址散列）
4. 最少链接
5. 加权