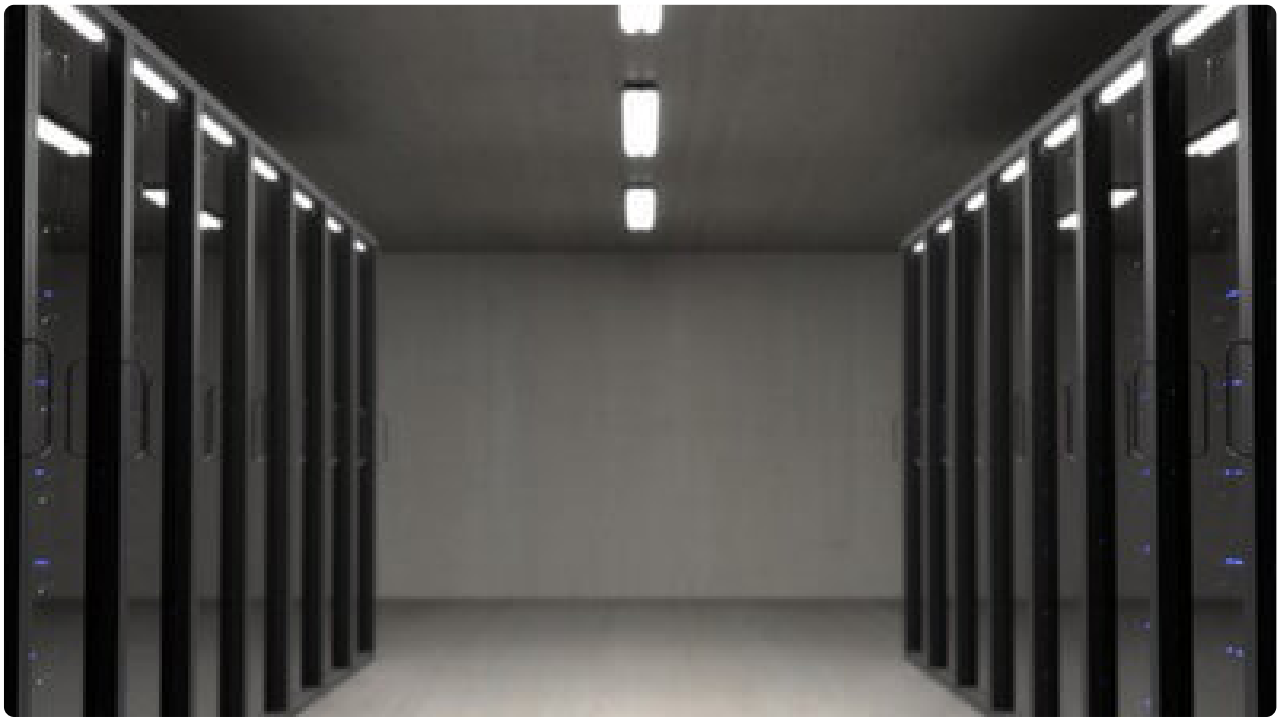
19 何为负载均衡：人多力量大

更新时间：2020-02-19 09:57:26



> 只有在那崎岖的小路上不畏艰险奋勇攀登的人,才有希望达到光辉的顶点。——马克思

前言

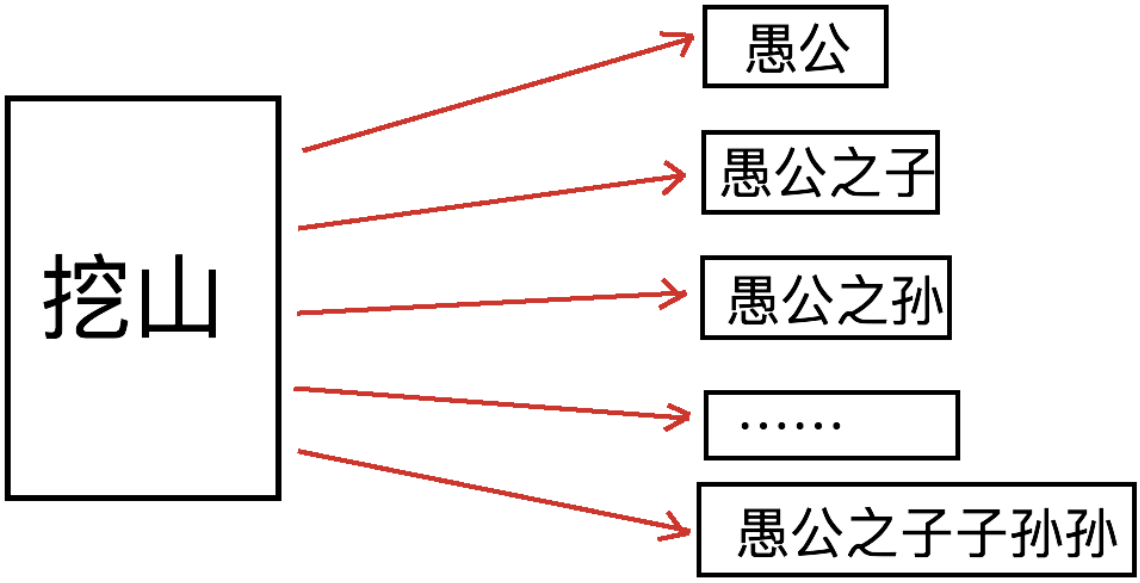
负载均衡的英文叫做 "Load Balance"，它是一种把任务分给多个计算机，由这些计算机共同工作从而完成这个任务，避免某一个机器因为执行的任务过多而造成性能过低，可以提高系统的可用性和稳定性。

我想大家都 "愚公移山" 这个典故应该非常了解，这个故事中有一句非常出名的话：

> 虽我之死，有子存焉。子又生孙，孙又生子；子又有子，子又有孙；子子孙孙无穷匮也，而山不加增，何苦而不平？

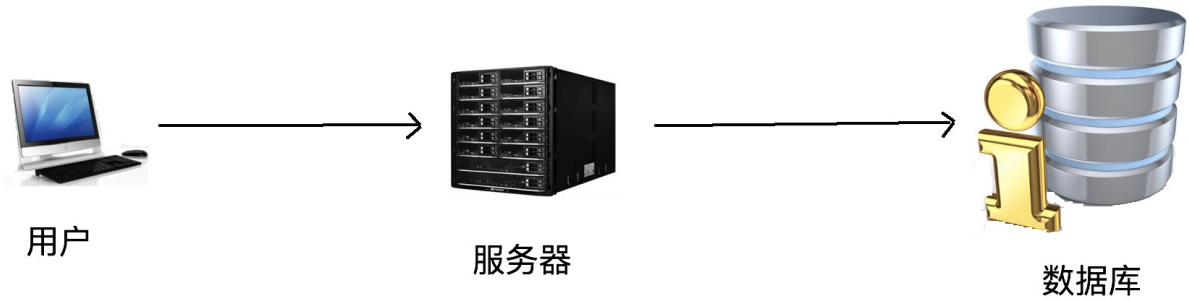
我们从一个计算机系统的角度来理解这句话就可以这么理解：

"挖山"　就是整个任务，这是一个巨大的任务，"愚公"　一个人无法完成这个任务，那么就把这个任务分给　“愚公”，“愚公之子”，“愚公之孙”，“愚公之子子孙孙”，大家一起挖山，最终完成整个任务。其实这就是计算机中的一个典型的负载均衡系统。



为什么要引入负载均衡

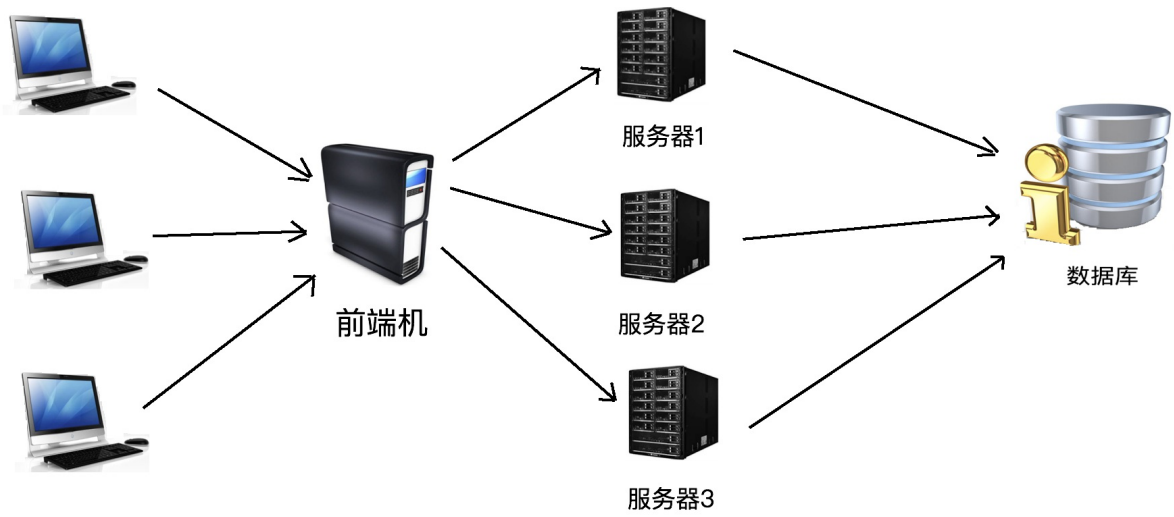
我们先看一个不使用负载均衡时候的系统架构：



上面的这种架构有两个非常大的缺陷：

- 这种架构会引发一个非常著名单点故障问题：如果服务器发生了故障，那么用户就无法获取访问任何服务，这是非常致命的。
- 第二：如果同一时段内有大量的用户访问服务器，超出服务器的处理能力，有可能造成服务瘫痪问题，这也是无法忍受的。

  为了解决上面的两个问题，就出现了负载均衡机制，我们可以在服务器和用户之间增加一个前端机，这个前端机后面可以挂在多个服务器，所有服务器提供相同的服务，前端机将用户的请求按照一定的策略分发给后端服务器即可。



如上图所示，即使服务器1 不能正常工作，还有服务器2 和服务器3 可以提供服务，这样就有效的避免了单点故障 问题。其实这个架构中 前端机 还是有单点故障问题，我们可以通过增加 前端机 个数来解决这个问题。当同时又大量请求到来的时候， 前端机 可以通过一定的策略将请求分配给不同的 服务器，这样就避免了某个服务器因为因为请求过多被拖垮。

负载均衡策略

上面我们提到过， 前端机 就通过一定的策略将用户的请求分发到 服务器 上面。这里的 一定的策略 就是我们常说的负载均衡策略。负载均衡策略可以决定哪台后端服务器会被选中。策略有很多，不同的策略作用适用于不同的环境，我们介绍几个常用的策略帮助大家理解这些概念。

- 轮训：第一个用户请求到来的时候，选择第一个后端服务器，第二个请求选择第二个服务器，以此类推，直到最后一个，然后再次从头开始循环。
- ip 哈希：将用户的源 IP 经过哈希之后得到一个散列值，然后根据这个散列值选择一个后端服务器。这种方式可以保证同一个 IP 的请求总是被同一个后端服务器处理，这样可以保存用户的状态。
- 最小连接：总是选择当前请求数量最少的服务器，因为这个服务器的压力最小，可以提供最优的服务。

总结

这还是一篇偏概念性的介绍，这些都属于基本知识，也是现在服务架构中的最基础原理，深入理解这些概念可以帮助大家设计出更健壮的软件架构。

}

18 反向代理的配置和实例

20 负载均衡模块搭建