负载均衡的原理、分类、实现架构，以及使用场景

题目标签

学习时长：20分钟

题目难度：中等

知识点标签：负载均衡

题目描述

负载过高的时候，通常会使用增加服务器数量来进行横向扩展

1. 面试题分析

根据题目要求我们可以知道：

- 为什么需要负载均衡
- 负载均衡的原理
- 负载均衡的作用
- 负载均衡的分类
- 最常见的四层和七层负载均衡
- 负载均衡应用场景

> 分析需要全面并且有深度

容易被忽略的坑

- 分析片面
- 没有深入

1.为什么需要负载均衡

当系统面临大量用户访问，负载过高的时候，通常会使用增加服务器数量来进行横向扩展，使用集群和负载均衡提高整个系统的处理能力。

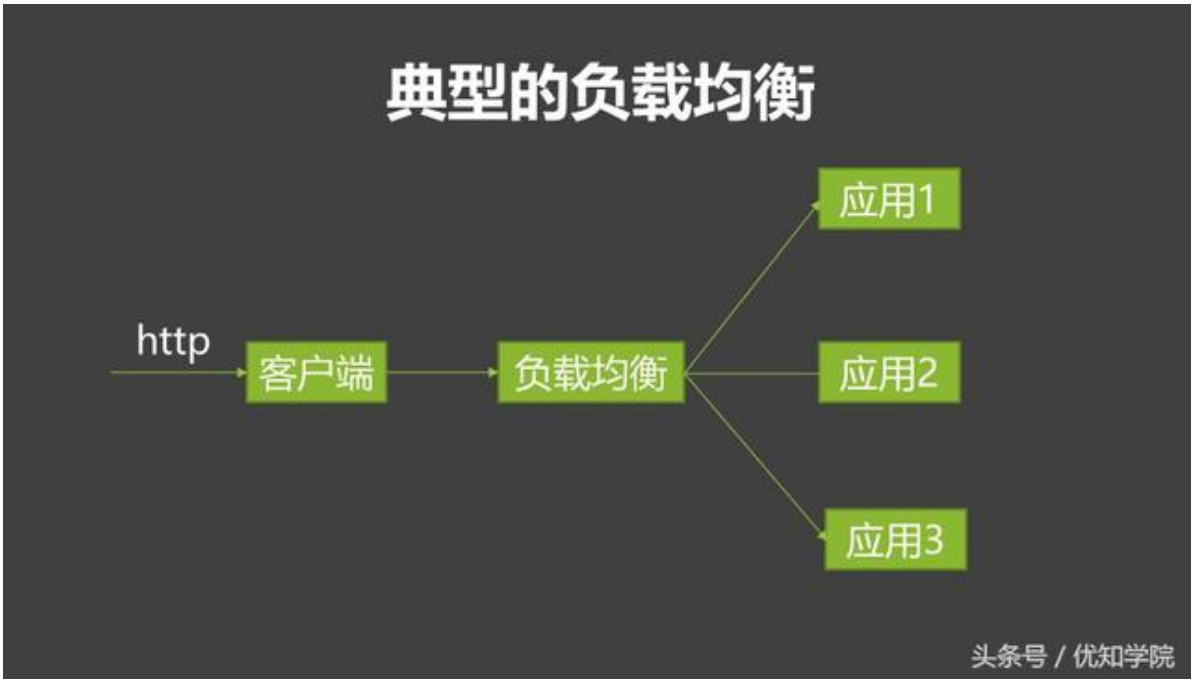
从单机网站到分布式网站，很重要的区别是业务拆分和分布式部署，将应用拆分后，部署到不同的机器上，实现大规模分布式系统。分布式和业务拆分解决了，从集中到分布的问题，但是每个部署的独立业务还存在单点的问题和访问统一入口问题，为解决单点故障，我们可以采取冗余的方式。将相同的应用部署到多台机器上。解决访问统一入口问题，我们可以在集群前面增加负载均衡设备，实现流量分发。

2.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）

3.负载均衡的作用

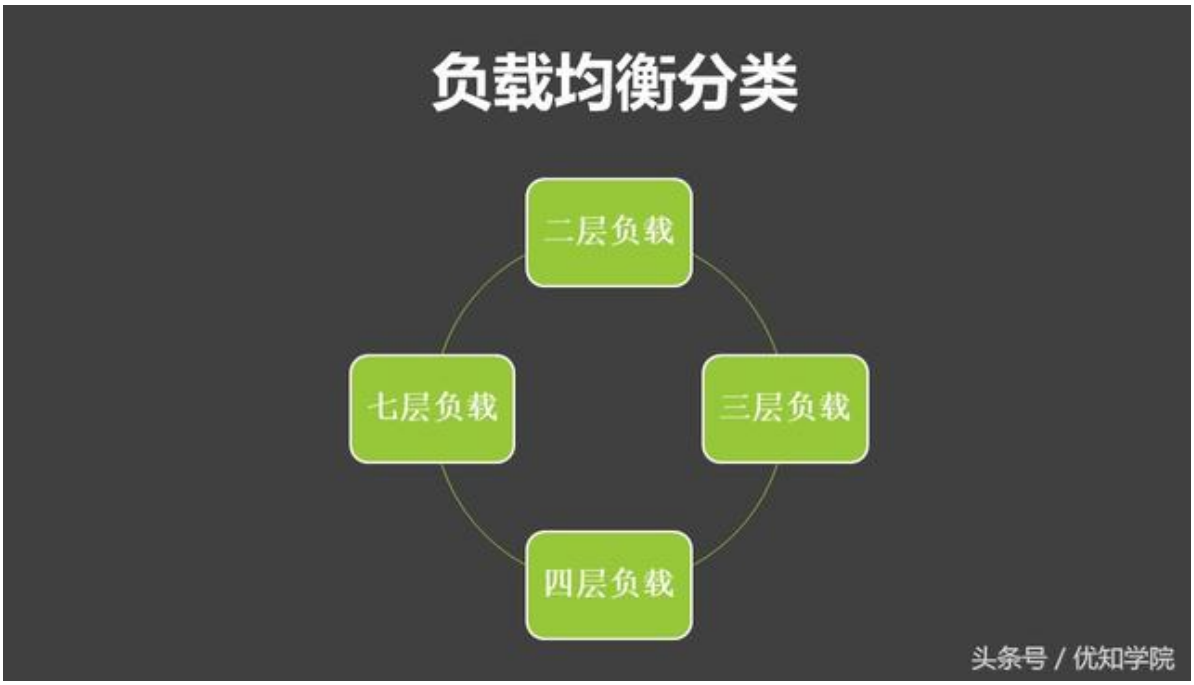
1.解决并发压力，提高应用处理性能（增加吞吐量，加强网络处理能力）；

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）

4.负载均衡的分类



1）二层负载均衡（mac）

根据OSI模型分的二层负载，一般是用虚拟mac地址方式，外部对虚拟MAC地址请求，负载均衡接收后分配后端实际的MAC地址响应）

2）三层负载均衡（ip）

一般采用虚拟IP地址方式，外部对虚拟的ip地址请求，负载均衡接收后分配后端实际的IP地址响应）

3）四层负载均衡（tcp）

在三次负载均衡的基础上，用ip+port接收请求，再转发到对应的机器。

4）七层负载均衡（http）

根据虚拟的url或IP，主机名接收请求，再转向相应的处理服务器。

5.最常见的四层和七层负载均衡

1）四层的负载均衡就是基于IP+端口的负载均衡：在三层负载均衡的基础上，通过发布三层的IP地址（VIP），然后加四层的端口号，来决定哪些流量需要做负载均衡。

对应的负载均衡器称为四层交换机（L4 switch），主要分析IP层及TCP/UDP层，实现四层负载均衡。此种负载均衡器不理解应用协议（如HTTP/FTP/MySQL等等）。

实现四层负载均衡的软件有：

- F5：硬件负载均衡器，功能很好，但是成本很高。
- lvs：重量级的四层负载软件
- nginx：轻量级的四层负载软件，带缓存功能，正则表达式较灵活
- haproxy：模拟四层转发，较灵活

2）七层的负载均衡就是基于虚拟的URL或主机IP的负载均衡

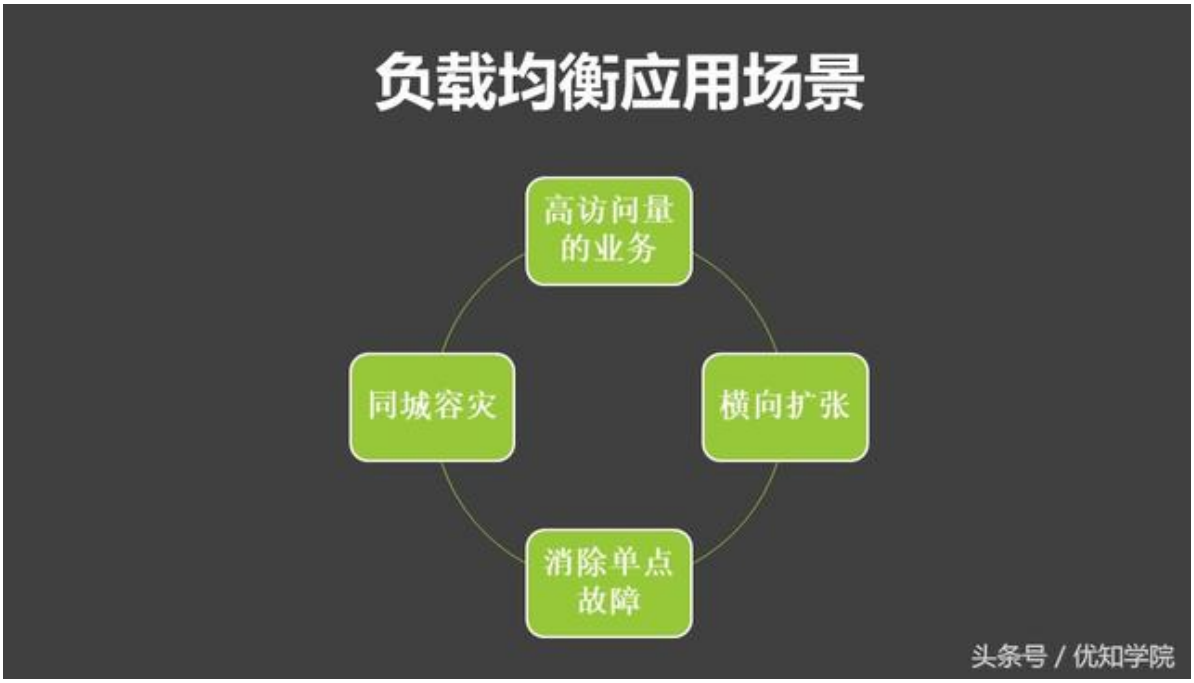
对应的负载均衡器称为七层交换机（L7 switch），除了支持四层负载均衡以外，还有分析应用层的信息，如HTTP协议URI或Cookie信息，实现七层负载均衡。此种负载均衡器能理解应用协议。

实现七层负载均衡的软件有：

- haproxy：天生负载均衡技能，全面支持七层代理，会话保持，标记，路径转移；
- nginx：只在http协议和mail协议上功能比较好，性能与haproxy差不多；
- apache：功能较差
- Mysql proxy：功能尚可。

总的来说，一般是lvs做4层负载；nginx做7层负载；haproxy比较灵活，4层和7层负载均衡都能做。

6.负载均衡应用场景



场景一：应用于高访问量的业务

如果应用访问量很高，可以通过配置监听规则将流量分发到不同的服务器上。

场景二：横向扩张系统

可以根据业务发展的需要，通过随时添加和移除服务器，来扩展应用系统的服务能力，适用于各种Web服务器和App服务器。

场景三：消除单点故障

当其中一部分服务器发生故障后，负载均衡会自动屏蔽故障的服务器，将请求分发给正常运行的服务器，保证应用系统仍能正常工作。

场景四：同城容灾 （多可用区容灾）

为了提供更加稳定可靠的负载均衡服务，当主可用区出现机房故障或不可用时，负载均衡仍然有能力在非常短的时间内切换到另外一个备可用区恢复服务能力；当主可用区恢复时，负载均衡同样会自动切换到主可用区提供服务，保证服务依然正常运行。

2. 扩展内容

- 分布式、集群、负载均衡、分布式数据一致性的区别与关联
- 分布式Session共享的4类技术方案，与优劣势比较
- 双11秒杀系统如何设计?