一、什么是高可用

高可用HA（High Availability）是分布式系统架构设计中必须考虑的因素之一，它通常是指，通过设计减少系统不能提供服务的时间。

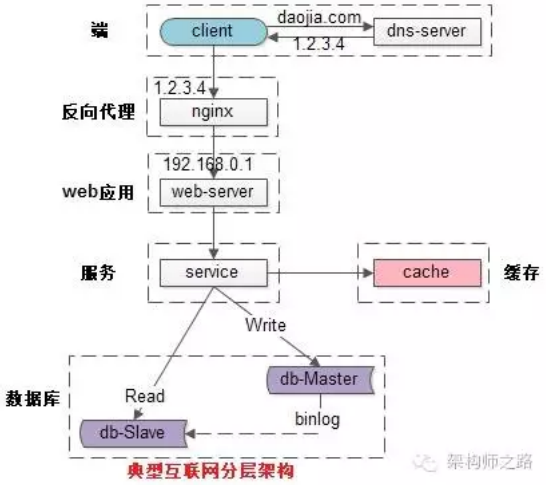
二、如何保障系统的高可用

我们都知道，单点是系统高可用的大敌，单点往往是系统高可用最大的风险和敌人，应该尽量在系统设计的过程中避免单点。方法论上，高可用保证的原则是“集群化”，或者叫“冗余”：只有一个单点，挂了服务会受影响；如果有冗余备份，挂了还有其他backup能够顶上。

保证系统高可用，架构设计的核心准则是：冗余。

有了冗余之后，还不够，每次出现故障需要人工介入恢复势必会增加系统的不可服务实践。所以，又往往是通过“自动故障转移”来实现系统的高可用。

三、常见的互联网分层架构



常见互联网分布式架构如上，分为：

（1）客户端层：典型调用方是浏览器browser或者手机应用APP

（2）反向代理层：系统入口，反向代理

（3）站点应用层：实现核心应用逻辑，返回html或者json

（4）服务层：如果实现了服务化，就有这一层

（5）数据-缓存层：缓存加速访问存储

（6）数据-数据库层：数据库固化数据存储

整个系统的高可用，又是通过每一层的冗余+自动故障转移来综合实现的。

和实现高并发一样,在每一层都做处理

四,总结

高可用HA（High Availability）是分布式系统架构设计中必须考虑的因素之一，它通常是指，通过设计减少系统不能提供服务的时间。

方法论上，高可用是通过冗余+自动故障转移来实现的。

整个互联网分层系统架构的高可用，又是通过每一层的冗余+自动故障转移来综合实现的，具体的：

（1）【客户端层】到【反向代理层】的高可用，是通过反向代理层的冗余实现的，常见实践是keepalived + virtual IP自动故障转移

（2）【反向代理层】到【站点层】的高可用，是通过站点层的冗余实现的，常见实践是nginx与web-server之间的存活性探测与自动故障转移

（3）【站点层】到【服务层】的高可用，是通过服务层的冗余实现的，常见实践是通过service-connection-pool来保证自动故障转移

（4）【服务层】到【缓存层】的高可用，是通过缓存数据的冗余实现的，常见实践是缓存客户端双读双写，或者利用缓存集群的主从数据同步与sentinel保活与自动故障转移；更多的业务场景，对缓存没有高可用要求，可以使用缓存服务化来对调用方屏蔽底层复杂性

（5）【服务层】到【数据库“读”】的高可用，是通过读库的冗余实现的，常见实践是通过db-connection-pool来保证自动故障转移

（6）【服务层】到【数据库“写”】的高可用，是通过写库的冗余实现的，常见实践是keepalived + virtual IP自动故障转移

> https://mp.weixin.qq.com/s?__biz=MjM5ODYxMDA5OQ==&mid=2651959728&idx=1&sn=933227840ec8cdc35d3a33ae3fe97ec5&chksm=bd2d046c8a5a