目录

[服务器基础入门 1](#_Toc477185558)

[基本名词 1](#_Toc477185559)

# 服务器基础入门

## 基本名词

首先对吞吐量（）、QPS、并发数、响应时间（RT）几个概念一直比较模糊，也不知道哪些指标可以较好的衡量系统的性能。今天特意查了些资料做一些记录：首先看一些概念（来自百度百科）   
1. 响应时间(RT)   
　　响应时间是指系统对请求作出响应的时间。直观上看，这个指标与人对软件性能的主观感受是非常一致的，因为它完整地记录了整个计算机系统处理请求的时间。由于一个系统通常会提供许多功能，而不同功能的处理逻辑也千差万别，因而不同功能的响应时间也不尽相同，甚至同一功能在不同输入数据的情况下响应时间也不相同。所以，在讨论一个系统的响应时间时，人们通常是指该系统所有功能的平均时间或者所有功能的最大响应时间。当然，往往也需要对每个或每组功能讨论其平均响应时间和最大响应时间。   
　　对于单机的没有并发操作的应用系统而言，人们普遍认为响应时间是一个合理且准确的性能指标。需要指出的是，响应时间的绝对值并不能直接反映软件的性能的高低，软件性能的高低实际上取决于用户对该响应时间的接受程度。对于一个游戏软件来说，响应时间小于100毫秒应该是不错的，响应时间在1秒左右可能属于勉强可以接受，如果响应时间达到3秒就完全难以接受了。而对于编译系统来说，完整编译一个较大规模软件的源代码可能需要几十分钟甚至更长时间，但这些响应时间对于用户来说都是可以接受的。   
  
2. 吞吐量(Throughput)   
     吞吐量是指系统在单位时间内处理请求的数量。对于无并发的应用系统而言，吞吐量与响应时间成严格的反比关系，实际上此时吞吐量就是响应时间的倒数。前面已经说过，对于单用户的系统，响应时间（或者系统响应时间和应用延迟时间）可以很好地度量系统的性能，但对于并发系统，通常需要用吞吐量作为性能指标。   
　　对于一个多用户的系统，如果只有一个用户使用时系统的平均响应时间是t，当有你n个用户使用时，每个用户看到的响应时间通常并不是n×t，而往往比n×t小很多（当然，在某些特殊情况下也可能比n×t大，甚至大很多）。这是因为处理每个请求需要用到很多资源，由于每个请求的处理过程中有许多不走难以并发执行，这导致在具体的一个时间点，所占资源往往并不多。也就是说在处理单个请求时，在每个时间点都可能有许多资源被闲置，当处理多个请求时，如果资源配置合理，每个用户看到的平均响应时间并不随用户数的增加而线性增加。实际上，不同系统的平均响应时间随用户数增加而增长的速度也不大相同，这也是采用吞吐量来度量并发系统的性能的主要原因。一般而言，吞吐量是一个比较通用的指标，两个具有不同用户数和用户使用模式的系统，如果其最大吞吐量基本一致，则可以判断两个系统的处理能力基本一致。   
  
3. 并发用户数   
　　并发用户数是指系统可以同时承载的正常使用系统功能的用户的数量。与吞吐量相比，并发用户数是一个更直观但也更笼统的性能指标。实际上，并发用户数是一个非常不准确的指标，因为用户不同的使用模式会导致不同用户在单位时间发出不同数量的请求。一网站系统为例，假设用户只有注册后才能使用，但注册用户并不是每时每刻都在使用该网站，因此具体一个时刻只有部分注册用户同时在线，在线用户就在浏览网站时会花很多时间阅读网站上的信息，因而具体一个时刻只有部分在线用户同时向系统发出请求。这样，对于网站系统我们会有三个关于用户数的统计数字：注册用户数、在线用户数和同时发请求用户数。由于注册用户可能长时间不登陆网站，使用注册用户数作为性能指标会造成很大的误差。而在线用户数和同事发请求用户数都可以作为性能指标。相比而言，以在线用户作为性能指标更直观些，而以同时发请求用户数作为性能指标更准确些。   
4. QPS每秒查询率(Query Per Second)   
　　每秒查询率QPS是对一个特定的查询服务器在规定时间内所处理流量多少的衡量标准，在因特网上，作为域名系统服务器的机器的性能经常用每秒查询率来衡量。对应fetches/sec，即每秒的响应请求数，也即是最大吞吐能力。   
  
从以上概念来看吞吐量和响应时间是衡量系统性能的重要指标，QPS虽然和吞吐量的计量单位不同，但应该是成正比的，任何一个指标都可以含量服务器的并行处理能力。当然Throughput更关心数据量，QPS更关心处理笔数。   
  
QPS提升带来什么？QPS提升说明单台服务器处理能力提升，如果QPS提升1倍，服务器资源减少1半，或者说服务器不变可以支撑2倍的请求量。   
如何提升QPS？   
1）减少CPU的使用时间（哪些代码会消耗CPU：循环、字符串拼接\查找\替换、编码\解码、序列化\反序列化、压缩）   
2）增加CPU的数量   
3）减少同步锁   
（如果CPU不能被压到85%以上，并且此时的QPS已经达到了峰值，则说明另有瓶颈，接下去关注内存）   
RT提升带来什么？   
响应速度提升说明单词请求的处理速度提升，用户感觉任务处理速度更快，系统反应速度更快。当然在处理能力不变的情况下，RT的提升必然会提升QPS。   
如何提升RT？   
1）减少I/O的响应时间   
2）减少I/O的调用次数   
3）减少CPU使用时间（当然在I/O占大头的应用里，这方面优化效果肯定不明显）