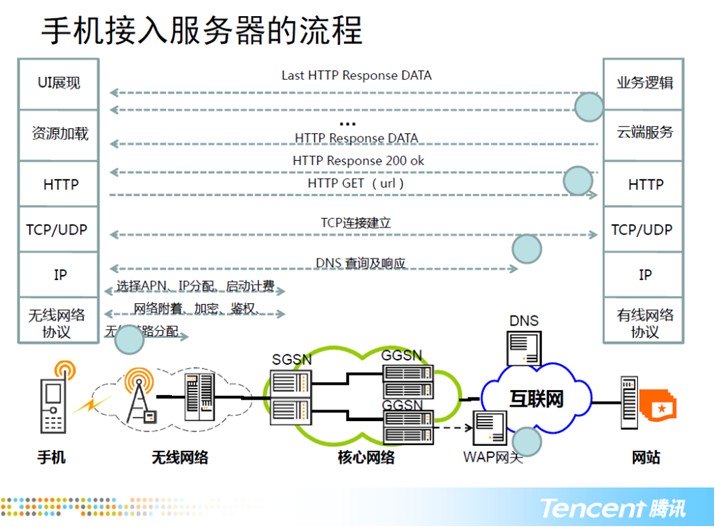
**一秒钟法则：来自腾讯无线研发的经验分享**

InfoQ  发表于 2014.4.14     2727浏览    0讨论   [查看原文](http://www.infoq.com/cn/articles/1sec-rule-from-tencent)

在2014年4月11日的腾讯分享日活动上， 来自腾讯MIG的移动互联网事业群运营总监/T4专家，负责运营QQ手机浏览器、腾讯PC浏览器、腾讯手机安全管家、腾讯电脑管家产品的刘昕介绍了移动无线产品研发中的“一秒钟法则”。本文根据该演讲内容整理形成。

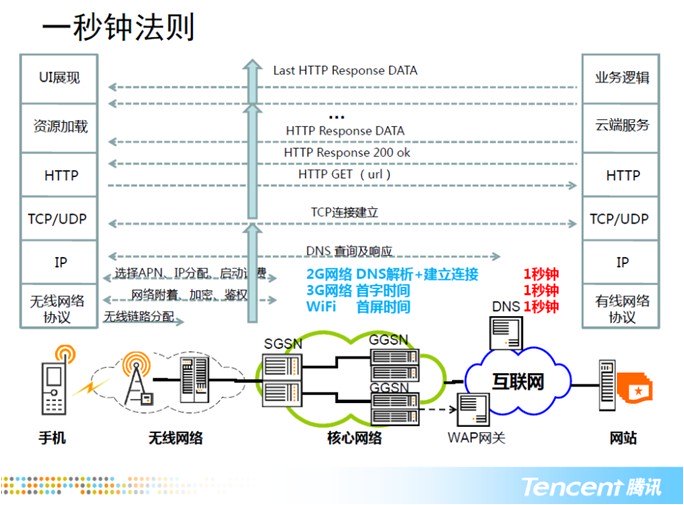


移动互联网的一个很大问题在于无线网络跟以前的有线网络不一样，无论是网络的组织形态、架构、通讯机制，跟有线网络都有很大差异，这带来很多挑战。今天介绍的“一秒钟法则”就是根据我们在移动互联网研发、运营过程中总结出来的一条解决的原则。  
   
手机接入服务器的流程

  
   
首先，手机要通过无线网络协议，从基站获得无线链路分配，才能跟网络进行通讯。

无线网络基站、基站控制器这方面，会给手机进行信号的分配，已完成手机连接和交互。  
   
获得无线链路后，会进行网络附着、加密、鉴权，核心网络会检查你是不是可以连接在这个网络上，是否开通套餐，是不是漫游等。核心网络有SGSN和GGSN，在这一步完成无线网络协议和有线以太网的协议转换。  
   
再下一步，核心网络会给你进行APN选择、IP分配、启动计费。

再往下面，才是传统网络的步骤：DNS查询、响应，建立TCP链接，HTTP GET，RTTP RESPONSE 200 OK，HTTP RESPONSE DATA，LAST HTTP RESPONSE DATA，开始UI展现。  
   
这是手机通过无线网络接入服务器的全过程。整个过程当中有几个困扰开发者的问题：  
 1.无线网络是怎么给手机分配到无线链路的？  
2.核心网络有接入点（APN），这里的CMNET和CMWAP有什么区别，仅仅是协议不同吗吗？数据转发又有什么区别？一个数据包在不同网络上传输有不同吗？  
3.用户怎么最快的找到正确的服务器？内容怎么快速有效的加载，在第一时间显示出来？  
   
这几个问题的重点在于其中的几个连接点：  
 1.无线链路分配。这是一个物理实连接。  
2.IP层链接。这是一个逻辑虚连接。  
3.TCP层链接。这是一个逻辑虚连接。  
4.HTTP层链接。这是一个逻辑虚连接。  
5.用户在线。这是一个逻辑虚连接。  
   
即使TCP连接建立，看到用户在线，也必须在手机获得无线链路分配的情况下，一个完整的通信才能真正完成，上行下行数据才能发送。这是移动互联网非常重要的特性。在现实中，手机已经分配ip也可能是没有无线链路，为什么？无线网络的资源是有限的，必须有效利用，这里由无线网络的信令机制完成无线网络资源的分配与释放。  
   
以用手机打电话的场景示例：用户在手机上拨号出去后，手机会跟网络申请无线链路，呼叫申请会发给电路域的核心网，通过电话交换机找寻被叫电话，被叫方接通电话，无线链路建立；完成通话，挂断的时候，手机给网络发送指令，表示服务使用结束，把已经分配的无线链路释放。  
   
上网的情况就比较复杂一些了。什么时候决定无线链路的分配？什么时候决定通讯完成？对于这两个时间点，不同的网络制式、不同的运营商都是不同的，不过大致上有几个区间值：  
   
在2G Edge网络下，差不多是1秒钟不传数据，就释放物理连接，回收给其他人备用。3G网络会延长几秒钟。  
   
这样的设定是有原因的。比如现在我们这个会场里有200人，那么我们200人同时上网的前提是共享同一个基站的资源，共享资源必须要有规则，比如要有排序，根据资源情况、用户链接活跃决定分配还是回收，这都是通过无线网络信令控制的。  
   
给一个手机分配无线信道的信令又有好几个情况，比如基站跟手机，基站跟基站控制器、核心网。举个例子，服务器从后台发送push消息，移动网络可能不知道这个手机是否活跃，不知道在哪个小区，移动网络就会发一个寻呼，在各个小区找这个手机，当然这个不能基于IP，而是其他的网络标识。找到了之后，这个手机再去申请信道资源，然后才能接受push。所以，这种场景下信令的消耗可能会在很多小区产生。  
   
根据以上情况，就形成无线网络的一大特点：秒级状态管理，秒级状态转换。这两个操作都在几百ms到几秒之间进行，对于维持连接来说时间太短，对于从无连接到有连接的转换来说时间又太长。  
   
相比之下，有线网络的状态管理如ip分配、tcp连接释放，都是分钟级，而状态转换则是毫秒级。  
   
这些通讯机制，同时加上无线网络的高延迟、高丢包。如何保证移动互联网的产品提供稳定的、可预期的服务质量，成为非常大的挑战：  
 •2G网络上无线部分数据传输的延迟有几百ms，4G网络上无线部分传输延迟减少到几十ms，核心网状态转换、协议转换30~100ms，IP骨干网上的延迟又跟物理距离以及运营商互联互通质量有关，跨运营商50-400ms，同运营商5-80ms，这个还要取决于网络拥塞的情况。  
•无线网络误码率比有线高两个数量级，在不同时间段的波动也非常巨大。  
   
怎么基于移动网络的特性去优化服务？这就是我们总结的一秒钟法则：在一秒内要完成的规定动作。  
 •2g网络：1秒内完成dns查询、和后台服务器建立连接  
•3g网络：1秒内完成首字显示（首字时间）  
•wifi网络：1秒内完成首屏显示（首屏时间）  
   
这些指标需要在终端度量，必须跟用户体验相关：首字时间、首屏时间都必须是用户可以直观感受到的。  
   
优化思路

  
   
接入调度优化  
   
接入调度优化首先要考虑的是减少DNS的影响。移动网络的DNS有如下特点：  
 1.骨干网无法识别移动用户在哪个城市，东西南北各个地方的调度没有充分调用。目前有一部分全国范围的DNS承载了超过40%的全网用户  
2.很多山寨机的终端local dns设置是错误的  
   
另外还有一些有线网络也一样会遇到的问题，如终端DNS解析滥用、域名劫持、DNS污染、老化、脆弱等。不过对于这些问题，桌面的自愈性会比较好，而在手机上则比较难以解决。  
   
对于DNS的问题，有两条主要的解决思路：  
 1.减少DNS的请求、查询、更新，也就是做DNS缓存  
2.在终端配置server list，直接访问IP，不用DNS  
   
但仅仅这么做还不够，因为用户可能来自国内外不同的运营商，还需要进一步优化调度策略：  
 1.DNS缓存需要多建立接入点，用不同域名区分  
2.IP列表需要更新以适应不同网络情况，要做到主动调度。好比最早我们只服务好移动用户就行，保证移动用户的接入质量优先，因为绝大多数用户集中在移动；现在国内有三个运营商，用户分布的比例在慢慢接近，要区分清楚；智能手机会用wifi，接入的是电信、联通还是哪个运营商，不知道，所以你不可能预先设置场景再if then，必须通过后台调度能力来解决。  
   
再进一步优化，就产生一种融合的方式：  
 1.先做域名解析，客户端直接连接解析的IP，可以用http协议，也可以用tcp socket  
2.多端口、多协议组合：不同协议有不同的限制，有些只能http，有些只能tcp socket，各种环境都要适应，客户端不能只支持一种协议  
3.终端测速：接入点越来越多，接入哪个合适，要选择，可以通过终端测速来选择最快的。你当然可以每一次新建连接都做测速，但是这样建立连接时间可能会很长；我们可以给用户先建立连接后，在后台根据长期速度监控、当前测速的结果，来做动态调度。也就是说，第一次连接可能不是最优，连接建立后动态测速，再转移到最快接入点。更进一步就是建立网络profile，终端学习的思路。  
   
测速采样的粒度我也说一下，移动互联网取IP段是没用的，比较好的粒度是到网元级别，比如广东有20多个wap网关，每一个网关的情况都不一样，这就是一个比较合适的粒度。  
   
另外我们后面还有一个SET模型，可以就近提供服务。  
   
最后想强调一个所有的接入调度原则：不要把调度逻辑写死在客户端，一定要由后台完成。  
   
协议优化  
   
协议参数优化这块就简单列一下，是我们长期运营过程中总结的一些经验，在启动移动互联网服务时作为运营的规范，可以少走很多弯路：  
•关闭TCP快速回收  
•Init RTO不低于3秒  
•初始拥塞控制窗口不小于10。因为大部分页面在10kB以下，很多请求在慢启动阶段已经结束，改为10可以降低小页面资源传输时延。内容越大，这个选项的效果就比较不明显。  
•Socket buffer > 64k  
•TCP滑动窗口可变  
•控制发包大小在1400字节以下，避免分片  
   
协议优化的原则总结下来是这么几条：  
 •连接重用  
•并发连接控制  
•超时控制  
•包头精简  
•内容压缩  
•选择更高效率的协议。无论是TCP、HTTP、UDP、长连接、GZIP、SPDY、WUP还是WebP，每一种协议、方案都有其道理，没有最优，只有是否适合你的产品和服务特点，需要大家在运营过程验证和取舍。  
   
WAP接入点优化  
   
关于WAP接入点优化，可能有些人会说，我们的App是高端大气上档次的应用，是不是就不用做WAP优化？实际上我们的统计显示，目前有5%-20%的用户选择的接入点是＊WAP（CMWAP、3GWAP、CTWAP），这甚至包括一些iPhone终端。实际上，WAP网关本质是个代理，不完全是落后的东西，随着技术的进步也在演进，以后在组网架构中可能有综合网关、内容计费网关来取代目前的WAP网关，所以建议也要一并考虑。以下是做WAP优化需要注意的一些问题：  
 •资费提醒页面  
•302跳转处理  
•X-Online-Host使用与处理  
•包大小限制  
•劫持与缓存  
•正确获取资源包大小  
   
业务逻辑优化  
   
简化逻辑：交互繁琐的内容尽量用标识更新。举一个例子，我们在老版的手机QQ上做过一个测试：假如我有100个好友，用手机QQ完成登陆，完成好友列表更新一遍，需要3.5分钟。这肯定是不合理的。建议用信令状态来通知是否需要更新，同时合理利用缓存。在比如玩游戏，好友给你送了很多星星，是让用户一次一次点还是批量点？从优化的角度肯定是批量点，从用户体验的角度这也更加舒服。  
   
另一方面，延长域名图标的缓存时间也可以有效地优化访问次数。我们把手机腾讯网图标的缓存时长从120分钟延长到2天后，访问次数优化了差不多35%。  
   
柔性可用：这个意思就是在网络质量好的时候给高清大图，不好的时候先给用户看小图，点一下再拉取原图。举一个极端的例子，比如万一地震了，基站毁掉20%，用户要给家人报平安，这时候产品上就必须优化，比如只发送文字，合理降低网络消耗。另外在响应很慢的时候，需要给用户一些合理的页面提示，比如提示用户再过5秒会发送，所以你不要一直刷屏，这也可以减少访问对后台服务、对网络的冲击。  
   
其实腾讯公司的很多产品在业务逻辑优化，更好的适应移动互联网场景上，有很多非常好的思路，今天由于时间关系就不再展开来。  
   
最后谈谈对优化方法的实践和结果的评估。QQ手机浏览器从4.5版本、5.0版本到5.1版本，我们对2G网络下的连接时间、3G网络下的首字耗时、wifi网络下的首屏耗时进行持续监控，耗时降到一秒钟以下还在不断的改进，每个新的版本平均值均有所压缩。这个结果是从每天用户实际使用的运营数据中得到的，覆盖到绝大多数的手机终端和网络环境。不过平均值只是一方面，我们另一方面还要看“有多少比例的数据满足了一秒钟法则”这个维度，因为无线网络的长尾数据波动很大，这一个维度也非常重要。目前现状是我们2G网络做到79%，3G网络做到73%，wifi网络做到69%。目前我们的目标是达到80%，实现之后，再进一步挑战90%的比例，不断追求极致。