

网络架构对 OTT 业务体验及空口速率需求影响分析 2014H1



目 录

1 简介.....	1
2 研究背景.....	2
3 用户体验与业务 TCP 吞吐率关系分析	3
3.1 主流 OTT 业务传输机制分析	3
3.2 用户体验与业务 TCP 吞吐率关系分析	4
3.2.1 TCP 吞吐率与 RTT 之间的关系	4
3.2.2 基于 TCP 业务的用户体验公式	5
3.2.3 Web 业务 TCP 传输特征分析	6
3.2.4 Video 业务 TCP 传输特征分析	8
4 典型网络场景用户体验对比	11
4.1 典型网络场景 E2E RTT 分析.....	11
4.2 UMTS 与 LTE 两种制式体验比较	13
4.2.1 UMTS 与 LTE Web 体验比较.....	13
4.2.2 UMTS 与 LTE Video 体验比较	13
4.3 Gi 口 Cache 体验提升比较	14
4.3.1 LTE Gi Cache 下 Web 业务体验.....	14
4.3.2 UMTS Gi Cache 下 Web 业务体验	15
4.3.3 LTE Gi Cache 下 Video 业务体验	15
4.3.4 UMTS Gi Cache 下 Video 业务体验	16
4.4 国内国外体验比较	16
4.4.1 国内国外 Top30 Web 体验比较	16
4.4.2 国内国际 Video 体验比较	17
4.5 中国东西部城市体验比较	17
4.5.1 中国东西部城市 Web 体验比较.....	17
4.5.2 中国东西部城市 Video 体验比较	18
4.6 典型网络 E2E RTT 和体验总结	19
5 总结.....	20
5.1 业务来水量决定空口需求	20
5.2 空口速率与 RTT 共同决定无线网络 OTT 业务速率	21

5.3 E2E RTT 影响因素.....	21
5.4 空口优化与架构优化协同实现最佳用户体验	22
5.5 mLAB 体验与网络需求研究中心的例行工作介绍	22
5.5.1 工作目标	22
5.5.2 工作方式	22
5.5.3 联系方式	23
5.5.4 免责声明	23
A 专业术语	24
B 参考文献.....	27

插图目录

图 2-1 空口不受限场景下 LTE 与 UMTS 比较.....	2
图 3-1 主流 OTT 业务传输机制.....	3
图 3-2 TCP 慢启动 TCP 吞吐率与 E2E RTT 关系.....	5
图 3-3 TCP 业务体验公式	6
图 3-4 Web 页面潜入对象个数与大小统计分析.....	6
图 3-5 Web 业务 TCP 传输特征统计分析	7
图 3-6 页面显示时延与页面嵌入对象数量之间关系(LTE)	8
图 3-7 Video 业务 TCP 传输特征统计分析	9
图 3-8 Video 业务 TCP 传输特征统计分析	10
图 4-1 典型网络 E2E RTT 分析	11
图 4-2 UMTS 与 LTE Web 业务体验比较	13
图 4-3 UMTS 与 LTE Video 业务体验比较	13
图 4-4 LTE Gi Cache 与无 Cache 场景 Web 体验比较.....	14
图 4-5 UMTS Gi Cache 与无 Cache 场景 Web 体验比较.....	15
图 4-6 LTE Gi Cache 与无 Cache 场景 Video 体验比较	15
图 4-7 UMTS Gi Cache 与无 Cache 场景 Video 体验比较	16
图 4-8 国内国际 Web 业务体验比较	16
图 4-9 国内国际 Video 业务体验比较	17
图 4-10 中国东西部城市 Web 体验比较	17
图 4-11 中国东西部城市 Video 体验比较	18
图 5-1 空口速率需求与 RTT 及 TCP 并发连接数之间的关系	20
图 5-2 空口速率需求与 RTT 之间的关系	21
图 5-3 E2E RTT 影响因素	21
图 5-4 空口和架构优化协同提升用户体验.....	22

表格目录

表 3-1 TCP 慢启动阶段 TCP 吞吐率与 RTT 关系	4
表 3-2 Video 初始缓冲和分片统计	9
表 4-1 典型网络 E2E RTT 组成	12
表 4-2 典型网络 E2E RTT 和体验	19

1 简介

mLAB 在 OTT 业务体验及网络需求研究中发现，在空口都不受限的情况下，LTE 仍然获得了比 UMTS 更好的用户体验，针对上述问题，同时为了研究无线网络能力所能达到的最佳体验极限，本文在主流 OTT 业务传输机制研究基础上探索了网络架构（本文着重用 E2E RTT 来表示网络架构的差异）对用户体验的影响。

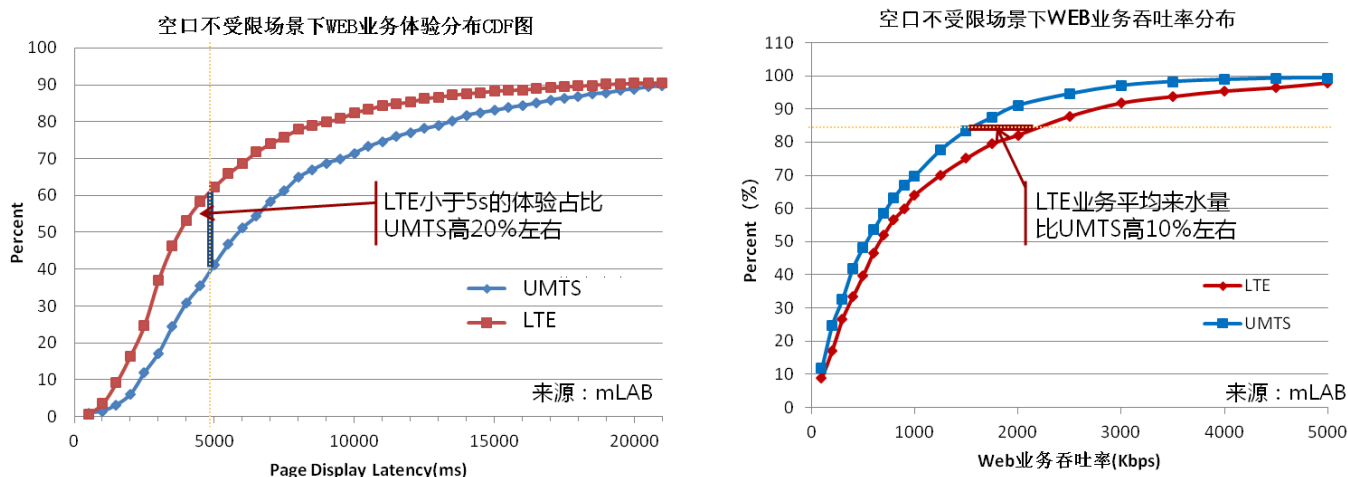
研究的关键发现有：

- ✓ 在空口不受限时：
 - ✓ 用户可获得体验和业务速率，主要取决于业务内容大小和业务 TCP 吞吐率
 - ✓ RTT 越小，TCP 吞吐率会在越短时间内达到或接近空口速率
 - ✓ RTT 越小，业务 TCP 吞吐率越高，对空口带宽的需求也越高
- ✓ 影响 E2E RTT 最为关键的要素是网络架构，在相同的条件下，不同的网络架构导致了不同的 RTT 和 TCP 层吞吐量；为达到最佳的用户体验，不仅要进行基于空口覆盖的建网，而且需要对网络架构进行持续改善

2 研究背景

通过对接入同一核心网且 Gi 出口相同的 UMTS 和 LTE 网络进行体验对比测试，在 UMTS 和 LTE 空口都不受限情况下访问相同的 TOP Web 页面，体验结果如图 2-1 所示。

图2-1 空口不受限场景下 LTE 与 UMTS 比较



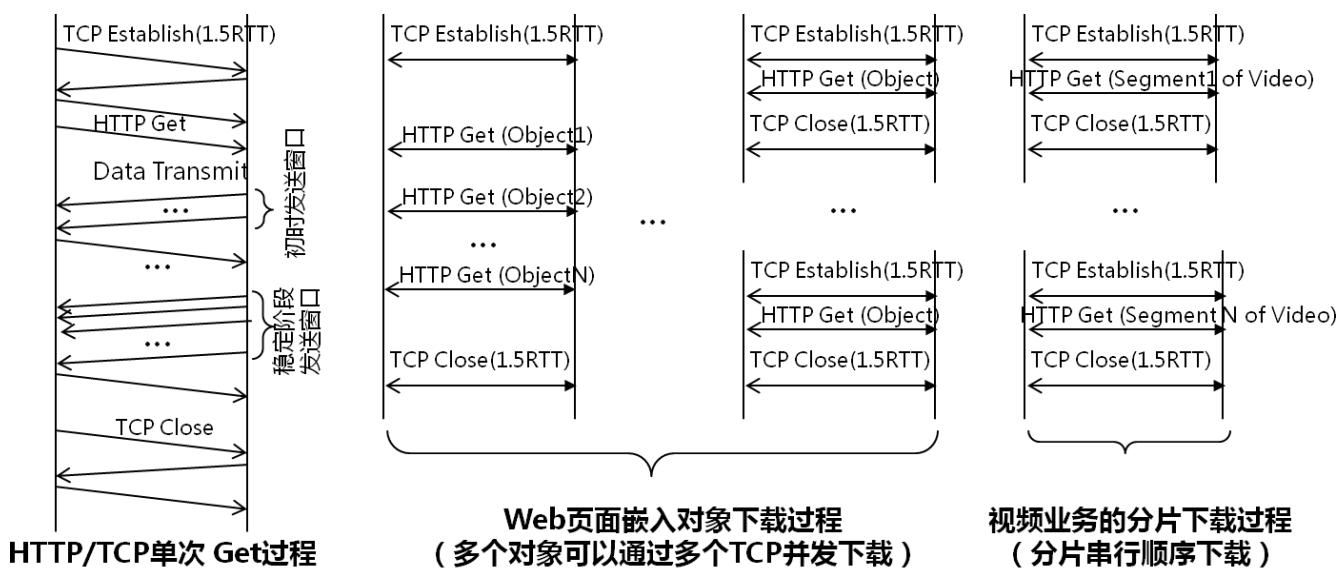
参考图 2-1，可以看出 LTE 小于 5s 的体验占比 UMTS 高 20%左右，同时 LTE 网络下 Web 业务平均吞吐率比 UMTS 高 10%，即 LTE 比 UMTS 获得了更高的业务来水量。

由于所测试 UMTS 和 LTE 两个网络空口都没有其他负载，业务速率基本不受空口带宽的限制，因此两者的主要差异体现在 UMTS RAN 和 LTE RAN 所带来 RTT 的不同。前期 mLAB 已经研究并发布了空口带宽对用户体验的影响，详细可参考 mLAB 发布的《OTT 业务体验及网络能力需求参考 V2.5 2013H2》（本报告年度刷新一次）。为了研究无线网络能力所能达到的最佳体验极限，本文重点分析也可能成为体验瓶颈的 TCP 吞吐率及网络架构（本文着重用 E2E RTT 来分析架构的差异）对 TCP 吞吐率的影响。

3 用户体验与业务 TCP 吞吐率关系分析

3.1 主流 OTT 业务传输机制分析

图3-1 主流 OTT 业务传输机制



- 业务吞吐率取决于 TCP 并发连接数以及单 TCP 的吞吐率，Web 业务单个主机默认情况下可以并发建立 10 个 TCP（单 Web 页面可能存在多个主机），而视频业务当前多数 OTT 每个分片建立一个 TCP，分片串行下载（由于视频有缓存机制，因此分片下载并非连续）
- 单 TCP 吞吐率取决于 RTT、发送窗口（初时发送窗口和稳定阶段最大发送窗口）
- 在链路带宽（如空口带宽）不受限的情况下，TCP 总吞吐率取决于单 TCP 吞吐率和 TCP 并发数

3.2 用户体验与业务 TCP 吞吐率关系分析

3.2.1 TCP 吞吐率与 RTT 之间的关系

1. 慢启动阶段 TCP 吞吐率

- 慢启动阶段前 N 个 RTT 内的 TCP 平均吞吐率（不考虑丢包）：

$$\text{TCP 吞吐率} = (2^N - 1) * \text{MSS} / ((N + 1.5) * \text{RTT})$$

- 慢启动阶段第 N 个 RTT 时刻的 TCP 平均吞吐率（不考虑丢包）：

$$\text{TCP 吞吐率} = 2^{N-1} * \text{MSS} / \text{RTT}$$

- 第 N 个 RTT 的发送窗口大小为 $2^{N-1} * \text{MSS}$
- 如果用 PMMS 表示单个 TCP 报文所承载的数据净荷大小，则前 N 个 RTT 传输数据块净荷大小为： $\text{TCP-PayLoad} = (2^N - 1) * \text{PMSS}$

表3-1 TCP 慢启动阶段 TCP 吞吐率与 RTT 关系

N	WIN	前N时刻平均吞吐率 (kbps)			第N时刻吞吐率 (kbps)			时延(ms)			前N个RTT传输块大小 (KB)	第N时刻发送的数据块 (KB)
		RTT=50	RTT=100	RTT=200	RTT=50	RTT=100	RTT=200	RTT=50	RTT=100	RTT=200		
0	1	0	0	0	240	120	60	75	150	300	0	1
1	2	80	40	20	480	240	120	125	250	500	1	3
2	4	180	90	45	960	480	240	175	350	700	4	6
3	8	336	168	84	1920	960	480	225	450	900	10	11
4	16	600	300	150	3840	1920	960	275	550	1100	21	23
5	32	1063	531	266	7680	3840	1920	325	650	1300	44	46
6	64	1890	945	473	15360	7680	3840	375	750	1500	90	91
7	128	3387	1693	847	30720	15360	7680	425	850	1700	181	183
8	256	6120	3060	1530	61440	30720	15360	475	950	1900	364	365
9	512	11149	5575	2787	122880	61440	30720	525	1050	2100	729	730
10	1024	20460	10230	5115	245760	122880	61440	575	1150	2300	1459	1460
11	2048	37791	18895	9448	491520	245760	122880	625	1250	2500	2919	2920
12	4096	70200	35100	17550	983040	491520	245760	675	1350	2700	5839	5840
13	8192	131056	65528	32764	1966080	983040	491520	725	1450	2900	11679	11680
14	16384	245745	122873	61436	3932160	1966080	983040	775	1550	3100	23359	23360
15	32768	462593	231296	115648	7864320	3932160	1966080	825	1650	3300	46719	46720
16	65535	873800	436900	218450	15728400	7864200	3932100	875	1750	3500	93439	93439

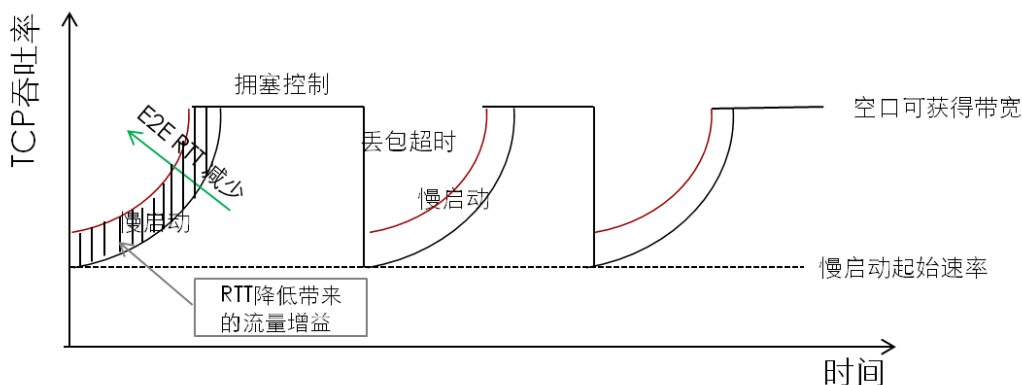
注：为了方便描述，上表窗口大小采用报文个数来描述，对应 TCP 协议中窗口大小需要乘以 MSS，本表 MSS=1500Bytes

在 TCP 传输机制（初始发送窗口，慢启动机制，服务器负载控制）一定的情况下，RTT 越小，业务对空口速率需求越高。

- 1、从表 3-1 可以看出，小于 90KB 的对象最多在 5.5RTT（不包括 TCP 建立的 1.5RTT）内就可以完成发送

- 2、 当发送窗口为 64(MSS=1500Bytes),RTT 为 50ms 时,此时 TCP 吞吐率为 15.36Mbps; 而当 RTT 增加为 100ms, TCP 吞吐率为 7.68Mbps, 此时对空口的需求也下降一半
- 3、 由此可以看出 RTT 越小, TCP 吞吐率越快速地接近最低链路带宽(譬如空口速率)的稳态速率

图3-2 TCP 慢启动 TCP 吞吐率与 E2E RTT 关系



2. 拥塞避免阶段（稳定阶段）内的 TCP 吞吐量

$$\text{TCP 吞吐率} = \min(W_m / \text{RTT}, \text{Link-Bandwidth}), \quad W_m = \min(CWIN, RWIN)$$

- W_m : 发送窗口大小
- $CWIN$: 拥塞窗口大小, 取决于链路的丢包率和时延抖动, 以及服务器发送性能和单用户吞吐率限制, TCP 慢启动之后, 对于持续数传, 在丢包率为 0 以及没有超时的情况下最终会达到 $RWIN$;
- $RWIN$: 接收窗口, 目前基本都比较大, 通过扩展因子最大可以达到 1GB, 因此接收窗口一般不会成为瓶颈;
- Link-Bandwidth 为终端与服务器之间节点或链路带宽的最小值, 空口带宽可以认为是当前 MBB 网络的主要瓶颈

3.2.2 基于 TCP 业务的用户体验公式

在不考虑丢包的情况, TCP 吞吐率主要取决于 TCP 发送窗口和 E2E RTT(Round-Trip Time); 在慢启动阶段 RTT 越小 TCP 发送窗口增长越快, 可以传输的数据越多; 而在拥塞控制阶段 TCP 发送窗口主要取决于 E2E 链路带宽, 链路带宽越大 TCP 发送窗口也会越大。单用户所访问业务的所有 TCP 吞吐量之和最终会达到或接近 E2E 链路中最可能成为瓶颈的空口可获得速率。

图3-3 TCP 业务体验公式

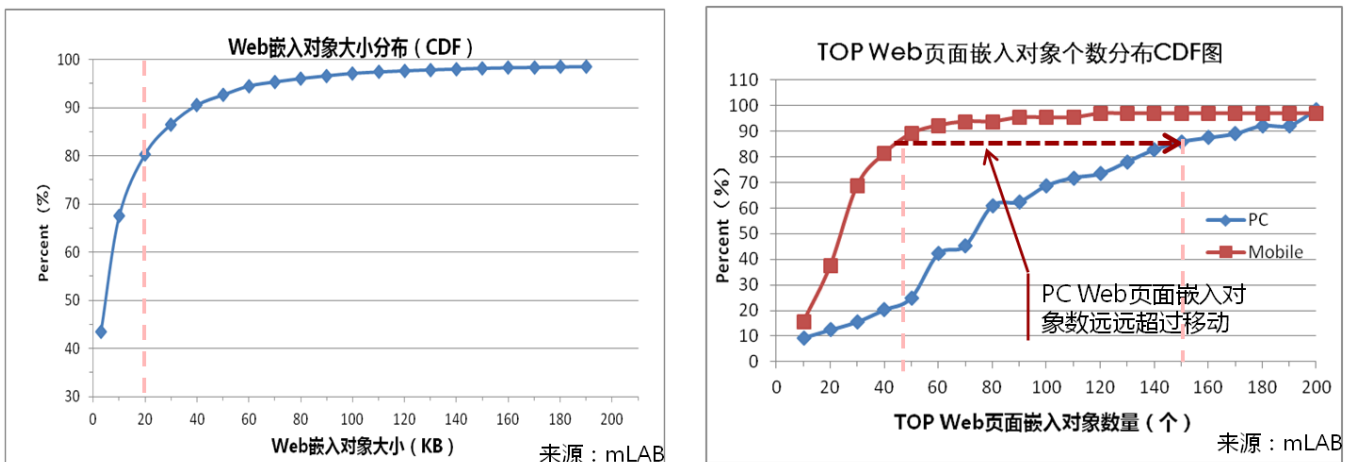
$$\text{MBB用户体验} = \frac{\text{业务内容大小}}{\text{MIN} \left\{ \sum_{\text{TCP并发数}} (\text{TCP发送窗口} / \text{E2E RTT}), \text{空口可获得速率} \right\}}$$

在空口速率的不受限时，业务 TCP 总吞吐率主要与 TCP 发送窗口、RTT 以及 TCP 并发数相关，降低业务 E2E RTT 是促进空口速率提升的关键措施。

- 在空口可获得速率不是瓶颈时，业务 TCP 吞吐率取决于 RTT，RTT 越小用户体验越好；
- 当 RTT 降低到一定程度后，业务吞吐率就会超越空口可获得速率，此时空口可获得速率将成为用户体验瓶颈

3.2.3 Web 业务 TCP 传输特征分析

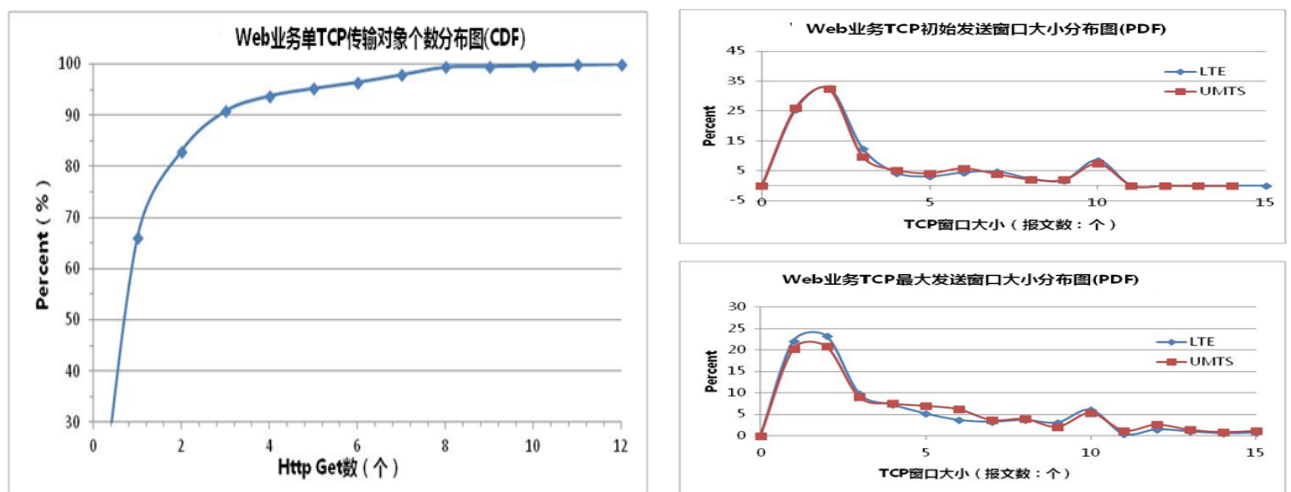
图3-4 Web 页面嵌入对象个数与大小统计分析



- 从图 3-4 可以看出目前 Web 页面嵌入对象 80%小于 20KB，一般在 3.5 RTT 内即可下载完（不包括 TCP 建立的 1.5RTT）
- 根据 mLAB 对 TOP 页面的统计，移动网页嵌入对象 90%在 50 个以下，而 PC 一般在 150 个以下
- Web 页面的加载时延包括主页面加载时延、主页面解析时延(主要取决于客户端性能，本文后续计算忽略此部分)、页面嵌入对象加载时延、以及页面渲染时延（本文后续计算未考虑此时延）几部分，其中主页面加载时延和页面嵌入对象加载时延可以按照如下方式进行估算：

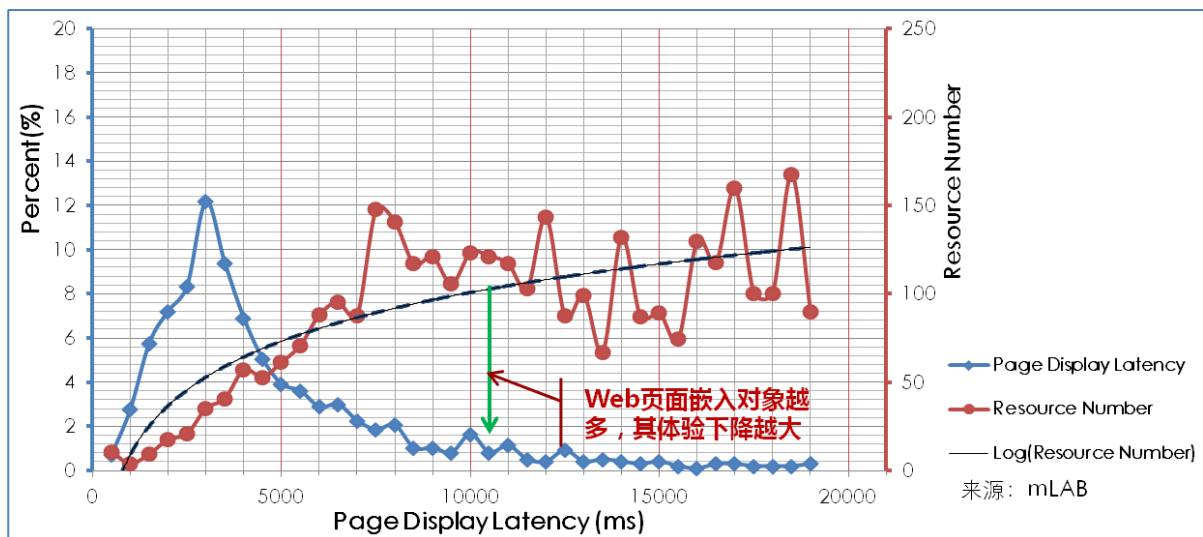
1. 主页面加载时延包括：主页面 DNS 请求时延（可以按照 1RTT 计算）、主页面 TCP 建立时延（1.5RTT）、主页面加载时延
 2. 对象加载时延可以参考如下计算：（对象 DNS 时延（如果与主文件属同一主机则忽略）+ TCP 建立时延+对象加载累计时延）×（对象总个数÷单 TCP 加载对象个数÷TCP 并发数）
- 对于嵌入对象个数 50 个，主文件以及嵌入对象大小均为 20KB 的网页为例（总大小约为 1MB），页面加载时延计算如下：
 1. 如果慢启动初始窗口为 1，根据根据表 3-1，主文件（20KB）和单个嵌入对象（20KB）均需要 3.5 RTT 完成加载，按照 TCP 并发率为 10 且每个 TCP 加载 2 个对象来计算，两个对象连续的加载时间为 3.5RTT（首对象加载时延）+1RTT(Get 请求)+1RTT(第二个对象加载时延)=5.5RTT，则一个页面加载时间大约为主文件加载时延(1RTT(DNS 请求)+1.5RTT（TCP 建立）+3.5RTT)+嵌入对象加载时延（1.5RTT（TCP 建立）+5.5RTT+1RTT（两条 TCP 之间的间隔））×(50 ÷ 2 ÷ 10)≈30RTT 左右；
 2. 如果慢启动初始窗口为 2，则主页面和 TCP 首个对象加载时延均减少 1RTT，则页面加载时延为主文件加载时延(1RTT+1.5RTT+2.5RTT)+ 嵌入对象加载时延（1.5RTT+4.5RTT+1RTT）×(50 ÷ 2 ÷ 10)≈26RTT。
 3. 因此对于本例所述网页，其加载时延可以认为在 26RTT~30RTT 之间，用户体验到的页面显示时延还需要考虑页面解析和渲染时延。

图3-5 Web 业务 TCP 传输特征统计分析



- 参考图 3-5 中 Web 业务单 TCP 传输对象个数分布图, 66%的 TCP 连接仅传送 1 个对象, 28%的 TCP 连接传输 2~4 对象, 传输 5 个以上对象的 TCP 连接约占 6%左右, 平均为 1.78 个对象
- 参考图 3-5 中 Web 业务 TCP 初始发送窗口大小分布图, Web 业务 TCP 初始发送窗口 LTE 和 UMTS 基本一致, 一般在 10 (初始连续下行报文数量) 以下, 25%左右初始发送窗口为 1, 75%初始发送窗口在 2~10 之间, 初始发送窗口为 10 的占比为 8.6%, 平均为 3.4
- 参考图 3-5 中 Web 业务 TCP 最大发送窗口大小分布图, Web 业务 TCP 最大发送窗口 LTE 和 UMTS 也基本一致, 87%的 TCP 最大发送窗口在 10 以下 (主要是因为 80%对象小于 20KB, 在窗口达到 8 时已经传完)
- 根据统计分析, Web 业务单主机 TCP 并发率一般是 10, 如果 Web 页面嵌入对象分布在多个主机上, TCP 并发率会超过 10 个

图3-6 页面显示时延与页面嵌入对象数量之间关系(LTE)



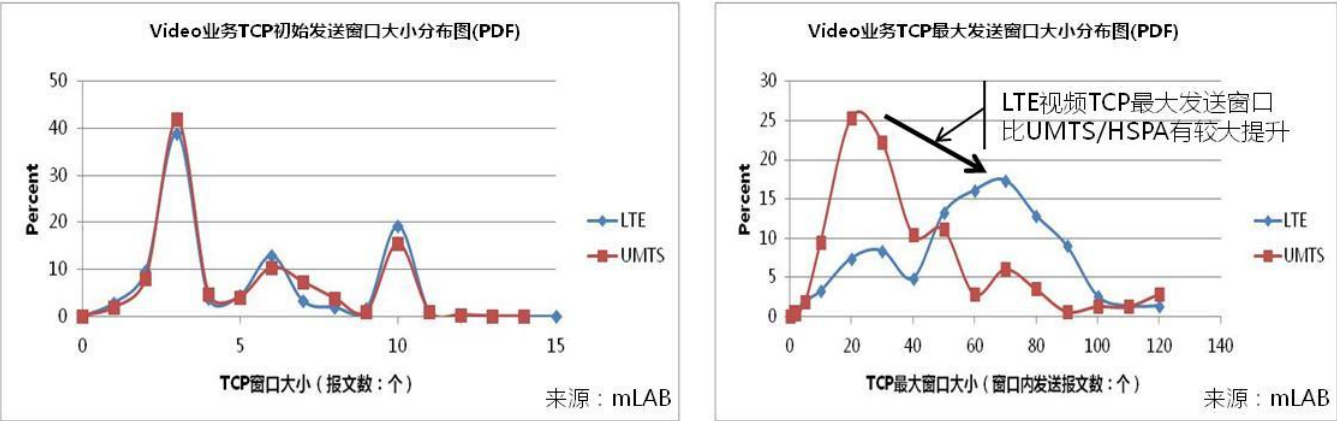
注：以5000ms点为例，蓝线上的点表示Page Display Latency在【4500，5000】区间占比为3.9%，红线上的点表示此区间所有网页平均嵌入对象为60

从图 3-6 可以获知页面嵌入对象的数量也是影响体验的关键因素，按照前述分析，嵌入对象个数为 50 个时，其页面加载时延一般为 26~30RTT，按 $RTT=120ms$ 计算，页面加载时延一般 3~3.6s，页面显示时延需要在此基础上加上各对象 URL 的 DNS 请求以及页面处理和渲染等时延；从上图可以看出嵌入对象超过 50 时，其页面显示时延基本都在 4s 以上。

3.2.4 Video 业务 TCP 传输特征分析

1. Video 业务 TCP 传输特征

图3-7 Video 业务 TCP 传输特征统计分析



注：上述数据来源于 mLAB LTE 网络下的测试数据，上述数据仅包括视频分片下载的 TCP 连接，非分片下载 TCP 连接没有统计在内

- 视频应用在下载分片时目前基本都是一个 TCP 仅传输一个视频分片
- 参考图 3-7 Video 业务 TCP 初始发送窗口大小分布图，LTE 和 UMTS 在下载视频分片时初始发送窗口大小基本一致，视频应用在下载分片时的 TCP 初始发送窗口 50%左右为 2~3，初始发送窗口为 10 的 TCP 也比较多，在 15%~20%之间
- 参考图 3-7 Video 业务 TCP 最大发送窗口大小分布图，视频应用在下载分片时 TCP 最大发送窗口 LTE 普遍比 UMTS 要大，LTE 最大发送窗口大于 30 的占比达 80%左右，而 UMTS 最大发送窗口大于 30 的占比只有 40%左右，LTE TCP 最大发送窗口大于 60 的 TCP 占比在 45%左右，而 UMTS 只有 15%左右
- 初始发送窗口大小决定分片下载初始速率，增加初始窗口可以降低慢启动的影响，从而提升用户体验；最大发送窗口决定 TCP 最高可达速率，最大窗口越大 TCP 峰值速率越高

2. 视频初时缓冲大小以及播放分片大小统计

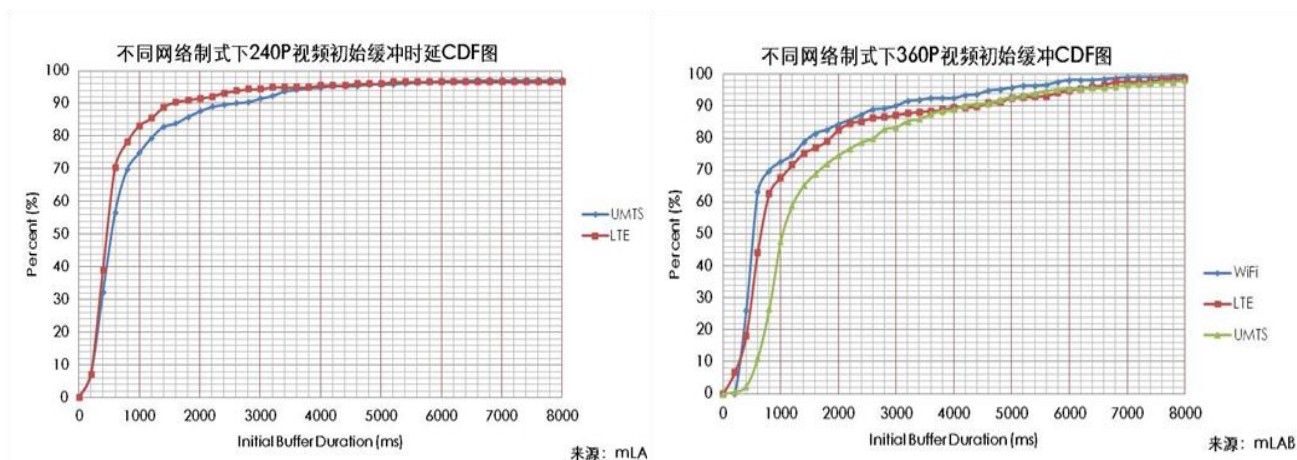
表3-2 Video 初始缓冲和分片统计

OTT	初时缓冲大小(KB)			播放过程分片大小 (KB)		
	240P	360P	720P	240P	360P	720P
YouTube	250	600	2200	3000	7000	15000
优酷	300	600	1200	14000	22000	30000
搜狐	250	600	1200	1000	3000	6000

- 视频客户端目前一般是采用单 TCP 对视频分片进行下载，参考“3.2.1 TCP 吞吐率与 RTT 之间的关系”，初时窗口按照 2 进行计算：
 - ✓ 对于 250~300KB 的初时缓存量，一般在 8RTT ($1.5RTT+6.5RTT$) 内完成加载，
 - ✓ 600KB 的初时缓冲量，一般在 10RTT ($1.5RTT+8.5RTT$) 内完成加载
 - ✓ 1200KB 的初时缓冲量，一般在 11RTT ($1.5RTT+9.5RTT$) 内完成加载
- 由上可以看出，在空口速率和视频服务器性能不是瓶颈时（对于 100ms RTT 而言，在第 7, 8 和 9 个 RTT 时刻的速率将分别达到 15.3Mbps, 30.7Mbps, 61.4Mbps），视频的体验主要与 RTT 相关
- 由于现网要满足每个用户随时随地 15Mbps 以上的带宽是很难的，在 TCP 吞吐率达到空口瓶颈之后 TCP 窗口趋于稳定（参考图 3-7 统计，一般会稳定在 $64 \times 1500\text{Bytes}$ 左右），因此现网的体验时延往往会大于上述理论分析值。

3. 视频体验与 RTT 关系分析

图3-8 Video 业务 TCP 传输特征统计分析

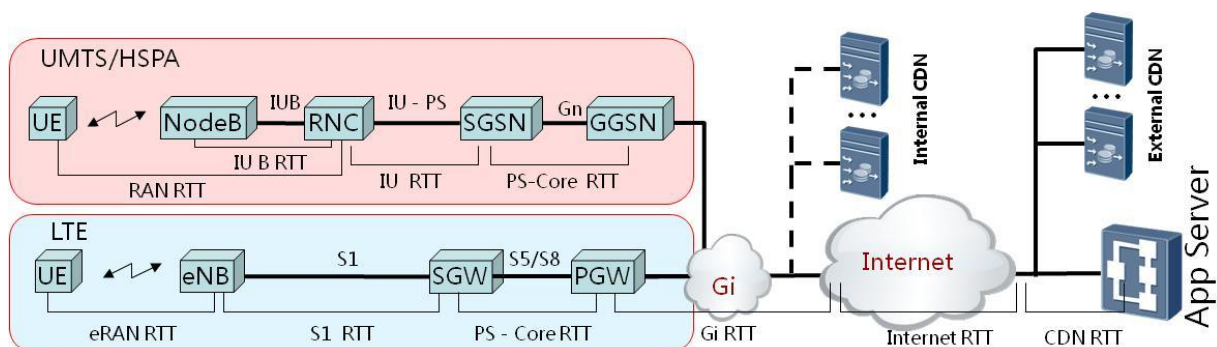


从图 3-8 中统计数据可以看出，240P 视频的初始缓冲时间，LTE 略高于 UMTS 约 0.2s 左右，360P 视频 LTE 比 UMTS 提升 0.5s 左右，而所测试网络 UMTS 比 LTE 的 RTT 高 25~40ms 左右。

4 典型网络场景用户体验对比

4.1 典型网络场景 E2E RTT 分析

图4-1 典型网络 E2E RTT 分析



由图4-1可以看出,E2E RTT 由(e)RAN RTT 、S1 RTT 或 IU RTT、PS Core RTT、Gi RTT、Internet RTT、以及 CDN RTT 各部分组成。

本文仅分析 UMTS/HSPA 和 LTE 两个制式，重点关注以下几个场景 RTT 对用户体验的影响：

- UMTS 和 LTE 两种制式网络场景下 RTT 对用户体验的影响，重点关注 UMTS 和 LTE 两种制式导致的 RTT 差异及对体验的影响；
- Gi Cache 和无 Cache 网络场景下 RTT 对用户体验的影响，重点关注 Internet 导致的 RTT 及对体验的影响；
- 国内国际 Internet 网络场景下 RTT 对用户体验的影响，重点关注 Internet 距离的不同导致的 RTT 以及对体验的影响；

- 国内东西部网络场景下 RTT 对用户体验的影响，重点关注因 CDN 部署位置的差异导致的 RTT 以及对体验的影响。

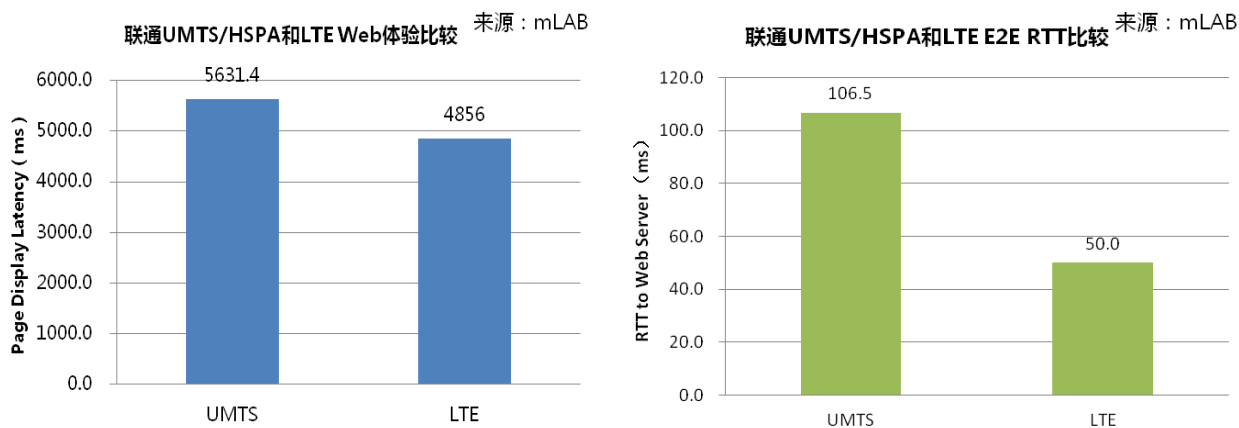
表4-1 典型网络 E2E RTT 组成

关键因素		分析
UMTS	LTE	
RAN RTT (HSPA 一般小于 50ms)	eRAN RTT (NGMN 要求小于 10ms，现网一般小于 20ms)	LTE 与 UMTS 空口调度机制以及架构都有较大差异，eRAN 时延已经提升到 10ms 以下，同时 LTE 比 UMTS 减少了 RNC 节点，同时也少了 Iub-RTT 部分时延；另外 RAN RTT 还取决于 RNC 的负载和性能，而 eRAN 仅受 eNB 的负载和性能影响
IU RTT	S1 RTT	IU 接口一般通过运营商骨干网互联，时延相对较低，S1 RTT 与 IUB RTT 相近。NGMN 要求 eRAN 包括 S1 接口的环回时延小于 10ms，最好 5ms
PS Core RTT		主要取决于网络负载状况，NGMN 基本要求为 10ms，最好小于 5ms
Gi RTT		主要取决于运营商防火墙以及相关业务监测网元（如 DPI 等）的部署和性能
Internet RTT		取决于 OTT 内容分发网络与用户接入网络之间是否存在直接互联，以及两者之间的传送性能；当 OTT 在 Gi 或 RAN 有 Cache 或 OTT CDN 与 Gi 口有直连链路时，这部分时延可以忽略
CDN RTT		主要取决于 OTT 服务器内部架构，包括 CDN 以及 OTT 服务器互联的网络架构
Client 收发处理时延		目前多数终端在单任务场景下均可满足当前 APP 的处理性能
Server 收发处理时延		在忙时，OTT 服务器处理能力可能成为 RTT 的瓶颈

4.2 UMTS 与 LTE 两种制式体验比较

4.2.1 UMTS 与 LTE Web 体验比较

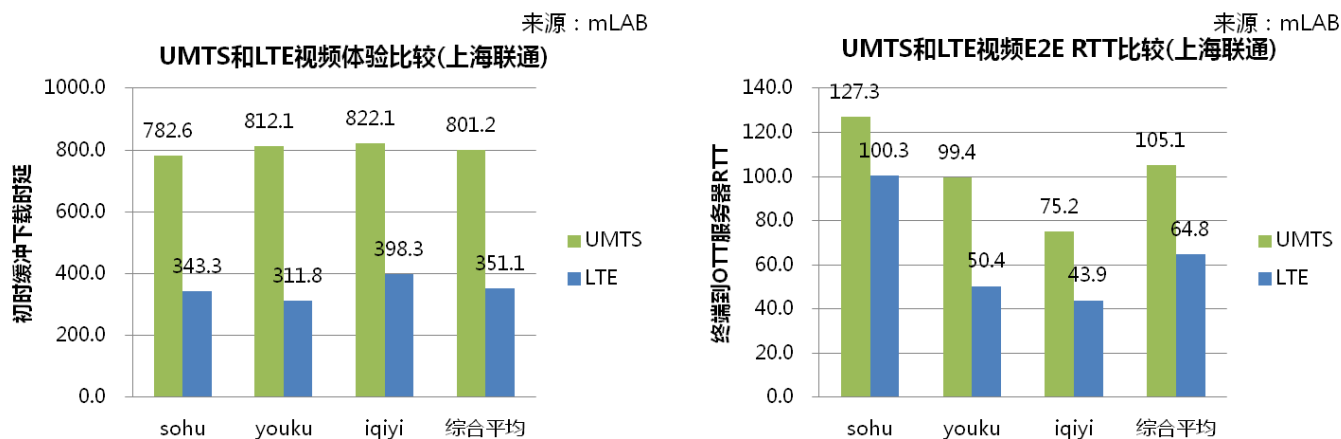
图4-2 UMTS 与 LTE Web 业务体验比较



- 注1：所测试Web页面是中国国内TOP 30 Web URL，UMTS和LTE访问相同的网页，每个网页测试为10次以上
- 注1：测试地点为上海，联通UMTS/HSPA通过Speedtest进行Ping测试其平均RTT为93ms，联通LTE通过Speedtest进行Ping测试其平均RTT为70ms
- 联通LTE网络访问中国TOP 30 Web URL，其平均RTT(客户端到服务器)为50ms，页面显示时延平均为4.856s；
- 联通UMTS网络访问中国TOP 30 Web URL，其平均RTT(客户端到服务器)为106.5ms，页面显示时延平均为5.63s。

4.2.2 UMTS 与 LTE Video 体验比较

图4-3 UMTS 与 LTE Video 业务体验比较

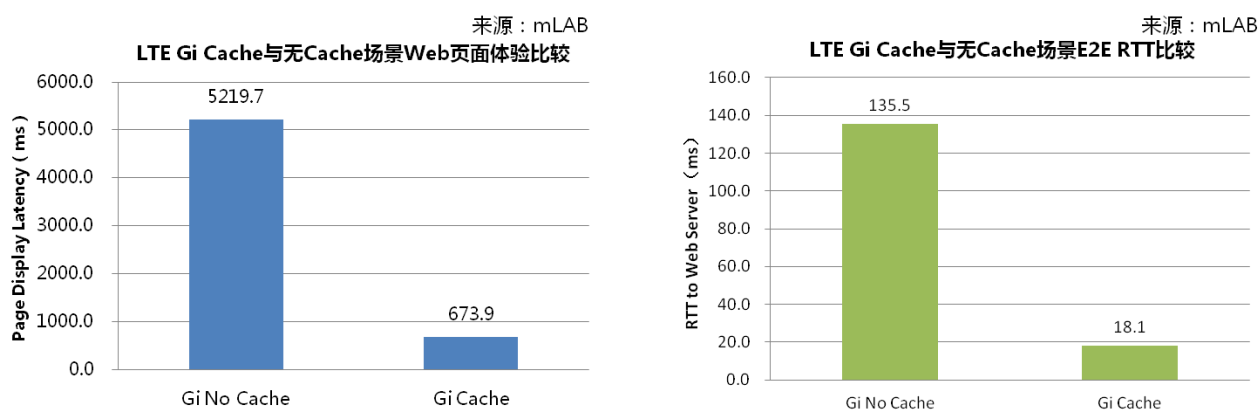


- 注 1: 测试对象为 sohu, youku 和 iqiyi, 每 OTT 取三个高清视频, 每个视频测试为 10 次以上
- 注 2: 所用网络为中国联通 UMTS/HSPA, 联通 UMTS/HSPA 通过 Speedtest 进行 Ping 测试其平均 RTT 为 93ms, 联通 LTE 通过 Speedtest 进行 Ping 测试其平均 RTT 为 70ms,
- 注 3: 在 LTE 测试搜狐视频时, 所测试片源的服务器恰好位于广东电信的网络, 所以 RTT 较大
- 上海联通 LTE 访问中国大陆 Top 视频站点, 其平均 RTT(客户端到服务器)为 64.8ms, 视频初始缓冲下载时延平均为 0.351s, 相对 UMTS 提升 50%以上;
- 上海联通 UMTS/HSPA 访问中国大陆 Top 视频站点, 其平均 RTT(客户端到服务器)为 105.1ms, 视频初始缓冲下载时延平均为 0.801s。

4.3 Gi 口 Cache 体验提升比较

4.3.1 LTE Gi Cache 下 Web 业务体验

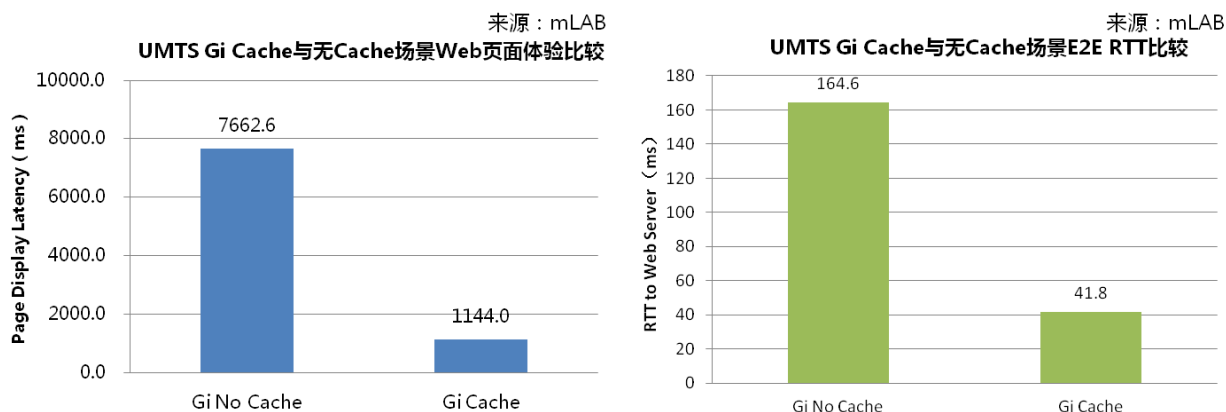
图4-4 LTE Gi Cache 与无 Cache 场景 Web 体验比较



- mLAB LTE 网络(Gi 出口在深圳)访问中国 TOP 30 Web URL, 其平均 RTT(客户端到服务器)为 135ms, 页面显示时延平均为 5.22s;
- 通过将上述 Top Web 页面 Cache 到部署在 Gi 口的服务器进行访问, 其平均 RTT(客户端到服务器)为 18ms, 页面显示时延平均为 0.67s, 而且各 Web 站点体验差异较小, 用户体验非常稳定。

4.3.2 UMTS Gi Cache 下 Web 业务体验

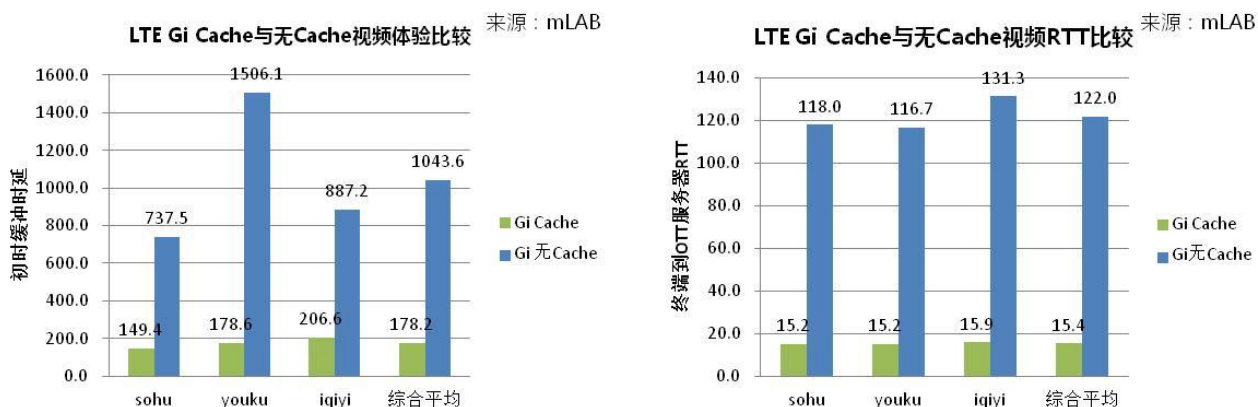
图4-5 UMTS Gi Cache 与无 Cache 场景 Web 体验比较



- mLAB UMTS 网络(Gi 出口在深圳)访问中国 TOP 30 Web URL,其平均 RTT(客户端到服务器)为 165ms, 页面显示时延平均为 7.663s
- 通过将上述 Top Web 页面 Cache 到部署在 Gi 口的服务器进行访问,其平均 RTT(客户端到服务器)为 42ms, 页面显示时延平均为 1.144s, 而且各 Web 站点体验差异较小, 用户体验非常稳定。

4.3.3 LTE Gi Cache 下 Video 业务体验

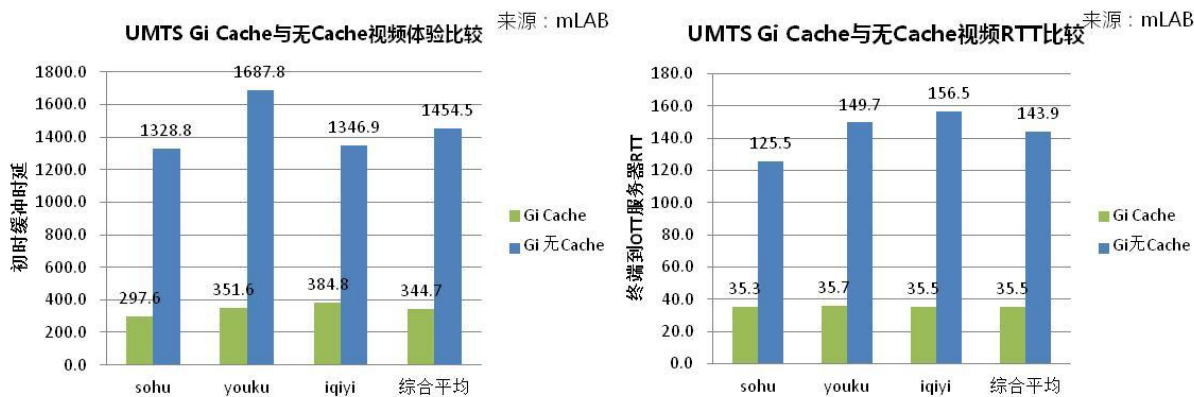
图4-6 LTE Gi Cache 与无 Cache 场景 Video 体验比较



- 通过 mLAB LTE (Gi 口在深圳) 访问中国大陆 Top 视频站点,其平均 RTT(客户端到服务器)为 122ms, 视频初始缓冲下载时延平均为 1.04s;
- 将视频 Cache 到部署在 Gi 口的服务器进行访问,其平均 RTT 为 15.4ms, 视频初始缓冲下载时延为 178.2ms, 达到零等待要求, 而且不同 OTT 视频的体验非常较小, 用户体验非常稳定。

4.3.4 UMTS Gi Cache 下 Video 业务体验

图4-7 UMTS Gi Cache 与无 Cache 场景 Video 体验比较

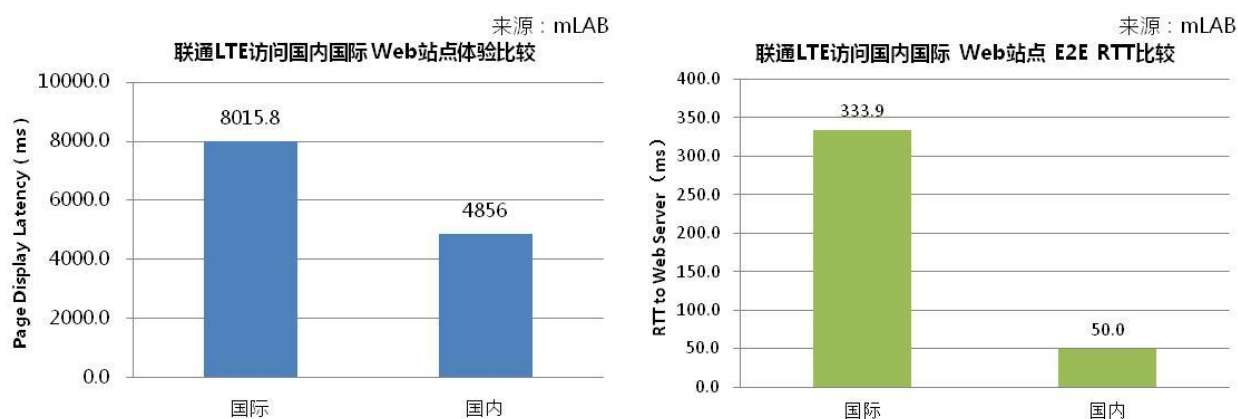


- 通过 mLAB UMTS (Gi 口在深圳) 访问中国大陆 Top 视频站点，其平均 RTT(客户端到服务器)为 143.9ms，页视频初始缓冲下载时延平均为 1.454s；
- 将视频 Cache 到部署在 Gi 口的服务器进行访问，其平均 RTT 为 35.5ms，视频初始缓冲下载时延为 0.345s，而且不同 OTT 视频的体验差异较小，用户体验比较稳定。

4.4 国内国外体验比较

4.4.1 国内国外 Top30 Web 体验比较

图4-8 国内国际 Web 业务体验比较

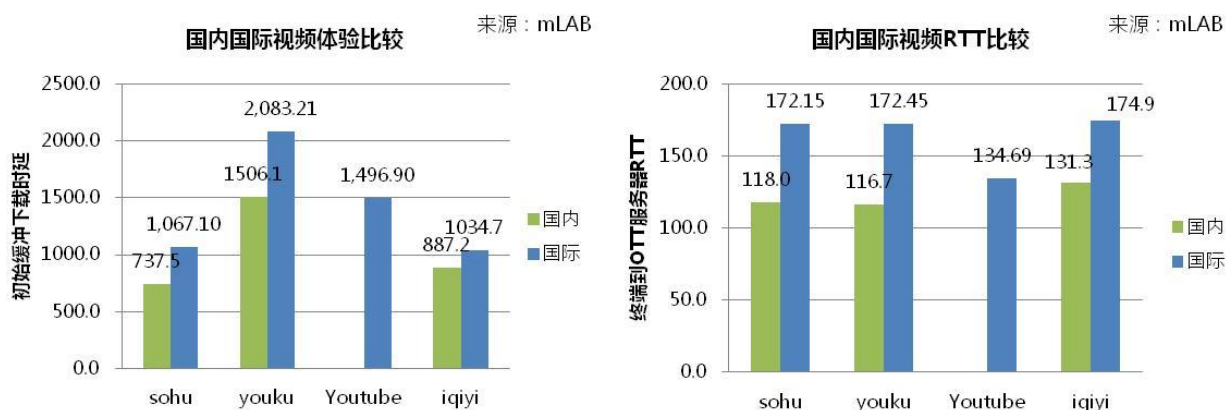


- 通过联通 LTE 网络访问中国 TOP 30 Web URL，其平均 RTT(客户端到服务器)为 50ms，页面显示时延平均为 4.856s

- 通过联通 LTE 网络访问国外同类 TOP 30 Web URL，其平均 RTT(客户端到服务器)为 333.9ms，页面显示时延平均为 8.016s。

4.4.2 国内国际 Video 体验比较

图4-9 国内国际 Video 业务体验比较

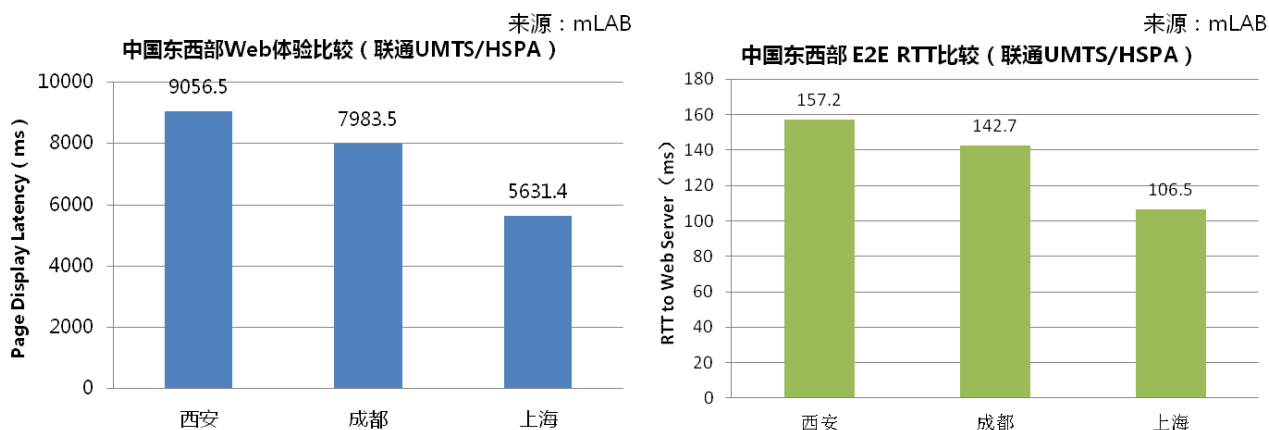


- 从 mLAB LTE 网络 (Gi 出口在深圳) 访问中国大陆 Top 视频站点，其平均 RTT(客户端到服务器)为 121ms，视频初始缓冲下载时延下载平均为 1.04s；
- 从 mLAB LTE 网络 (Gi 出口在香港) 访问中国大陆 Top 视频站点，其平均 RTT(客户端到服务器)为 164ms，页视频初始缓冲下载时延平均为 1.42s。

4.5 中国东西部城市体验比较

4.5.1 中国东西部城市 Web 体验比较

图4-10 中国东西部城市 Web 体验比较



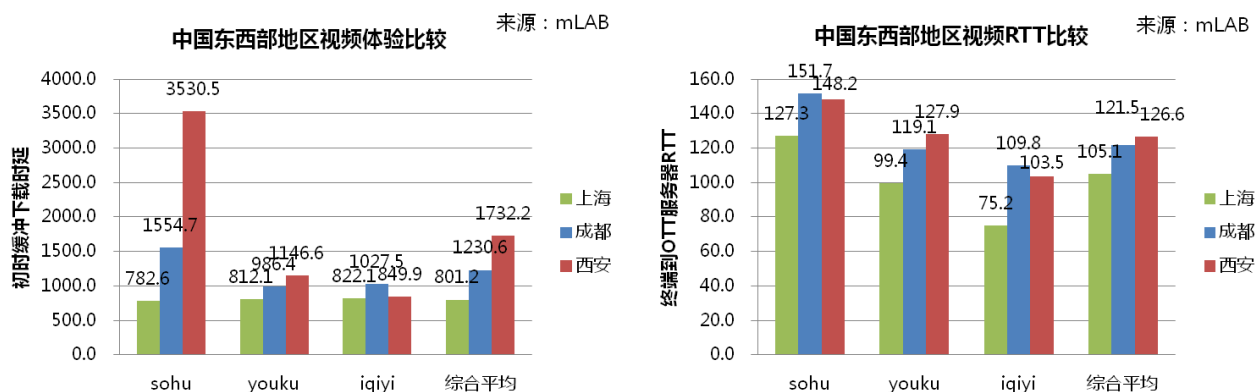
- 注 1: 所测试 Web 页面是中国国内 TOP 30 Web URL, 三地基本同一时段测试, 每个 URL 测试为 10 次以上
- 注 2: 所用网络为中国联通 UMTS/HSPA, 在上海通过 Speedtest 进行 Ping 测试其平均 RTT 为 93ms, 在成都通过 Speedtest 进行 Ping 测试其平均 RTT 为 108ms, 而在西安通过 Speedtest 进行 Ping 测试其平均 RTT 为 151ms

上述结果为中国 Top 30 Web 页面的测试平均值, 我们可以看出

- 中国东部城市上海 Web 体验 (页面显示时延) 好于成都和西安, 平均页面显示时延上海比西安小 3.4s, 比成都小 2.3s
- 中国东部城市上海 E2E RTT 也比成都和西安都小, 比西安小 50ms, 比成都小 36ms

4.5.2 中国东西部城市 Video 体验比较

图4-11 中国东西部城市 Video 体验比较



- 注 1: 测试对象为 sohu, youku 和 iqiyi, 每家 OTT 取三个高清视频, 三地基本同一时段测试, 每个视频测试为 10 次以上
- 注 2: 所用网络为中国联通 UMTS/HSPA, 在上海通过 Speedtest 进行 Ping 测试其平均 RTT 为 93ms, 在成都通过 Speedtest 进行 Ping 测试其平均 RTT 为 108ms, 而在西安通过 Speedtest 进行 Ping 测试其平均 RTT 为 151ms
- 中国东部城市上海访问中国大陆 Top 视频站点, 其平均 RTT (客户端到服务器) 为 105.1ms, 视频初始缓冲下载时延平均为 0.801s;
- 中国西部城市成都访问中国大陆 Top 视频站点, 其平均 RTT (客户端到服务器) 为 121.5ms, 页视频初始缓冲下载时延平均为 1.230s; 而西安访问中国大陆 Top 视频站点, 其平均 RTT (客户端到服务器) 为 126.6ms, 页视频初始缓冲下载时延平均为 1.732s;

4.6 典型网络 E2E RTT 和体验总结

表4-2 典型网络 E2E RTT 和体验

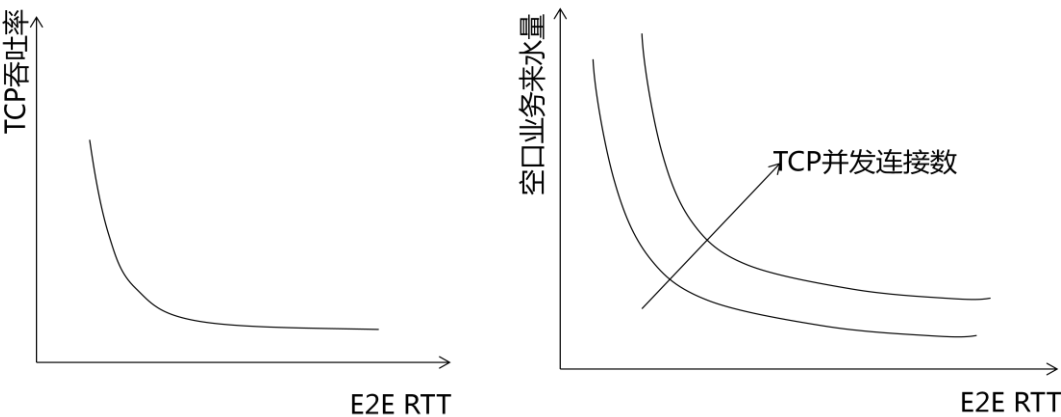
典型网络	Speed Test RTT (ms)	中国Top30 Web		中国Top3 视频	
		RTT (ms)	页面显示时延 (ms)	RTT (ms)	初始缓冲下载时延 (ms)
UMST/HSPA(GiCache) (mLAB)	-	42	1144	36	345
LTE(Gi Cache) (mLAB)	-	18	674	15	178
联通UMTS/HSPA (上海)	93	107	5631	105	801
联通UMTS/HSPA (成都)	108	143	7983	122	1230
联通UMTS/HSPA (西安)	151	157	9056	127	1732
联通LTE (上海)	70	50	4856	65	351

- 。 注：通过联通 LTE 访问国外 TOP30 站点，E2E RTT 平均为 334ms，页面显示时延平均值为 8.16s
- Gi Cache 带来的体验增益极大，而且体验比较稳定，LTE 基本能够实现“零等待”
 - 联通 LTE 视频体验相对 UMTS 有大幅提升，超清视频初始缓冲下载时延降低到 0.5s 以下；
 - 现网由于 OTT 大多将服务器部署在东部沿海等发达城市，因此东部城市终端与服务器之间的 E2E RTT 较西部城市小，相差约 20~50ms；
 - RTT 较大时，终端与服务器之间路径较长，中间网络节点较多，因此导致 RTT 波动的因素较多，RTT 波动 (RTT Jitter) 就比 RTT 小时较大，因此对应的用户体验也更差，后续将进一步研究 RTT 波动对用户体验的影响。

5 总结

5.1 业务来水量决定空口需求

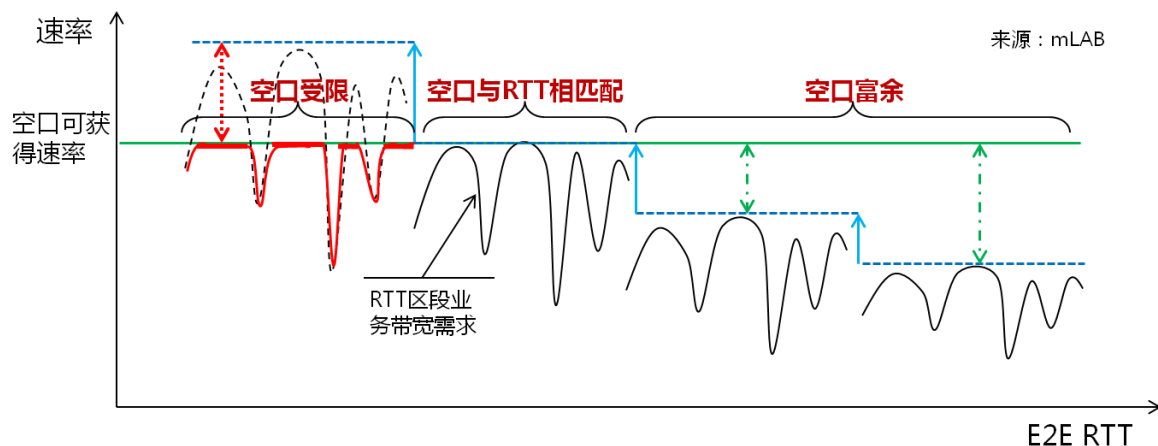
图5-1 空口速率需求与 RTT 及 TCP 并发连接数之间的关系



根据当前 OTT 的业务设计模式，RTT 和 TCP 并发连接数一起决定了业务的 TCP 吞吐率，同时也决定了对空口速率的需求（即业务来水量决定空口速率需求），当 E2E 带宽不是瓶颈时，RTT 的能力提升（降低时延）与用户体验的提升基本成正比例关系，持续优化 RTT 是提升用户体验的关键。

5.2 空口速率与 RTT 共同决定无线网络 OTT 业务速率

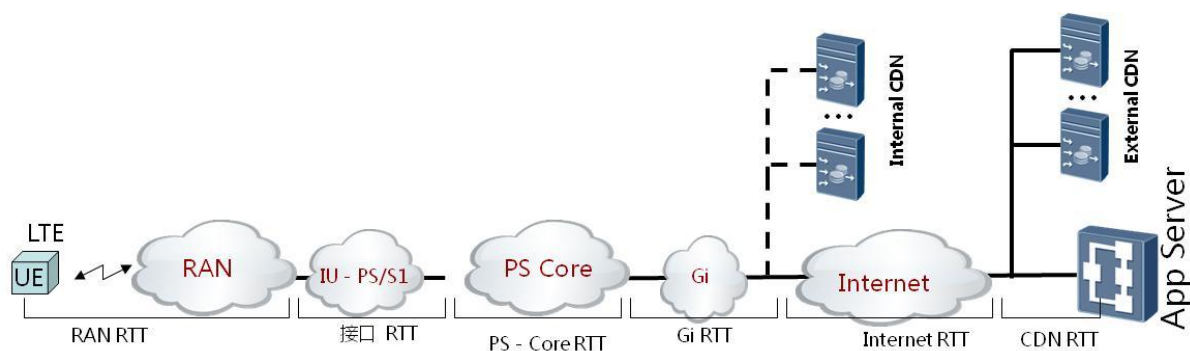
图5-2 空口速率需求与 RTT 之间的关系



当 E2E RTT 较小时，业务来水量（业务实际速率）较高，对空口带宽的需求也较高，此时空口带宽就相对容易出现受限；当 E2E RTT 较大时，业务来水量相对就会变低，对空口带宽的需求也较低，此时空口带宽相对就容易出现富余。

5.3 E2E RTT 影响因素

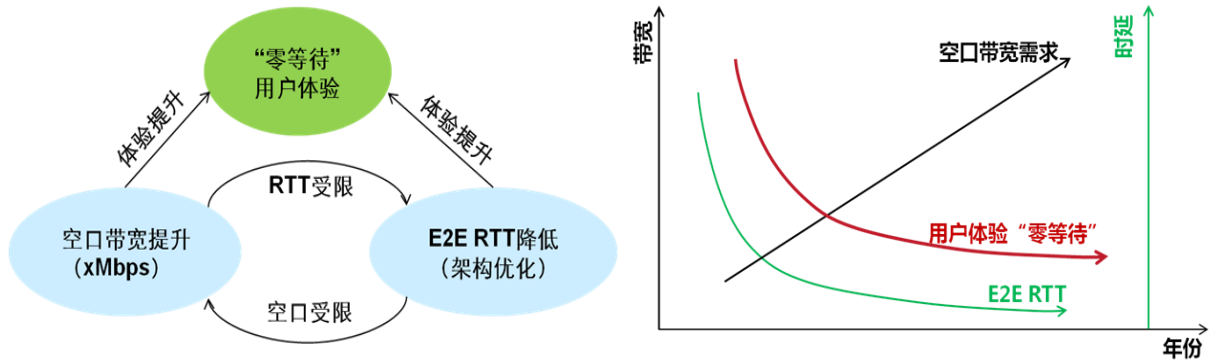
图5-3 E2E RTT 影响因素



终端与 OTT 服务器之间的 E2E RTT 影响因素较多，包括无线网络性能，OTT CDN 架构（CDN RTT）和部署位置（Internet RTT），以及应用服务器和终端的性能等。而无线网络部分的影响包括无线网络部署架构（如核心网的融合部署策略，网元之间的接口网络，Gi 口向 OTT CDN 开放等），网络容量，空口覆盖，网络制式，空口调度算法和相关配置参数，以及网络所部署的 QoS 策略等。针对这些领域，华为提供了系列的产品特性和解决方案（如 xMbps）来帮助客户构建最佳用户体验网络，同时提供了相应的网络架构优化服务来持续降低 E2E RTT。

5.4 空口优化与架构优化协同实现最佳用户体验

图5-4 空口和架构优化协同提升用户体验



- 用户体验不仅取决于空口覆盖（xMbps），而且取决于网络架构所导致的 RTT；
- 空口不受限时，用户体验主要与 RTT 相关，网络应重点关注网络架构的优化以降低 RTT，包括无线网路自身 RTT 的优化，也包括 OTT 网络架构优化（如 CDN 部署等）；
- 在链路带宽不是瓶颈时，RTT 越低业务 TCP 吞吐率越高，业务对空口带宽的需求也越大；
- 空口带宽与 RTT 优化需相互配合同步进行才能以最低的成本实现最大体验提升。

5.5 mLAB 体验与网络需求研究中心的例行工作介绍

5.5.1 工作目标

- ✓ 洞察 OTT 体验现状和变化趋势；
- ✓ 洞察 OTT 网络需求(Bandwidth and RTT)及发展变化。

5.5.2 工作方式

- ✓ 在 OTT 体验和网络需求研究领域，mLAB 目前不仅具有专业高效的工具 MBB Explorer,而且拥有多位有丰富经验的领域专家，同时 mLAB 也与多个国内外知名大学或专业研究机构建立了长期的合作；
- ✓ mLAB 体验和网络需求研究中心，每半年度都有例行输出，相关研究成果已经在华为建设或运维的网络得到广泛的应用，是华为 xMbps 解决方案的基础；
- ✓ mLAB 同时对主流运营商网络进行 QoE 和 RTT 例行监测，为运营商网络优化提供参考。



5.5.3 联系方式

本文作者：贺文胜/00150170

mLAB 体验需求研究中心负责人： 贺文胜/00150170 hewensheng@huawei.com

mLAB(MBB lab)联系方式：MBBlab@huawei.com

关注方法：

1, 扫描右方二维码进行关注
2, 在通讯录点击“添加朋友”，搜索微信号：**mbblab**
3, 点击右上角，选择 查看并关注官方账号



5.5.4 免责声明

本分析报告由华为 mLAB 体验及网络需求研究中心撰写，报告中提供的信息仅供参考。报告的数据源自 mLAB MBB Explorer 工具，由于样本数量受限，以及所测试网络 and OTT 业务的快速发展，华为 mLAB 保留后续更改本文中相关描述的权利；且不对更改造成的一切影响负责。

本报告不能作为投资研究决策的依据，不能作为道义的、责任的和法律的依据或者凭证，无论是否已经明示或者暗示。mLAB 将随时补充、更正和修订有关信息，但不保证及时发布。对于本报告所提供信息所导致的任何直接的或者间接的投资盈亏后果不承担任何责任。

本报告版权为华为 mLAB 所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布。如引用发布，需注明出处为华为 • mLAB，且不得对本报告进行有悖原意的引用、删节和修改。

A 专业术语

名词	解释
APP	应用
CDN	Content Distribution Network，即内容分发网络
DNS	Domain Name System，域名系统
GGSN	Gateway GPRS Support Node，网关 GPRS 支持节点
Gn	同一 PLMN 中 SGSN 与 GGSN 间的接口为 Gn 接口
HTTP	Hypertext transfer protocol，超文本传输协议
IUB	RNC 和 Node B 之间的逻辑接口
IU-PS	分组域 RNC 和 SGSN 接口
LTE	Long Term Evolution
MBB Explorer	mLAB 开发的 OTT 业务体验评估系统，包括 APP 客户端和 QoE 数据中心，能够完成 Web 和视频的体验测试，同时能够进行 DNS 和 Ping RTT 测试，并集成了 Speedtest 测速
MSS	TCP Maximum Segment Size 最大分段
OTT	Over the Top，本文泛指运行于无线网络之上的各类 MBB 业务或业务运营商
PGW	Packet Gateway，EPC 核心网网关
RAN	Radio Access Network，无线接入网
RTT	Round-Trip Time，往返时延

S1	LTE 基站（eNB）与 EPC SGW 之间的接口
SGSN	Serving GPRS Support Node，服务 GPRS 支持节点
SGW	Service Gateway, 服务网关
TCP	Transmission Control Protocol 传输控制协议
TCP 并发连接数	同时存在的 TCP 连接数量
TCP 初始窗口	TCP 初始发送窗口
TCP 发送窗口	TCP 发送方可以连续发送的最大字节数
TCP 接收窗口	TCP 接收方允许发送方可以连续发送的最大字节数
TCP 慢启动	慢启动，是传输控制协议使用的一种阻塞控制机制。慢启动也叫做指数增长期。慢启动是指每次 TCP 接收窗口收到确认时都会增长。增加的大小就是已确认段的数目。这种情况一直保持到要么没有收到一些段，要么窗口大小到达预先定义的阈值。如果发生丢失事件，TCP 就认为这是网络阻塞，就会采取措施减轻网络拥挤。一旦发生丢失事件或者到达阈值，TCP 就会进入线性增长阶段。这时，每经过一个 RTT 窗口增长一个段。
TCP 吞吐量	TCP 单位时间传输的数据块大小，也成为 TCP 速率
TCP 拥塞控制窗口	拥塞窗口(congestion window)，记为 cwnd。TCP 连接建立时拥塞窗口被初始化为 1 个报文段（即另一端通告的报文段大小）。每收到一个报文的 ACK，拥塞窗口就增加一个报文段（cwnd 以字节为单位，但是慢启动以报文段大小为单位进行增加）。发送方取拥塞窗口与通告窗口（TCP 接收窗口）中的最小值作为发送上限。拥塞窗口是发送方使用的流量控制，而通告窗口则是接收方使用的流量控制。
TCP 最大发送窗口	一个 TCP 连接的客户端或服务器端连续发送的最大字节数
UMTS	Universal Mobile Telecommunications System, 即通用移动通信系统
Web 对象	<p>Web Object: HTTP object, which is used for processing and rendering the webpage and is referenced by the page mark-up or script. However, these objects are not necessarily visible on the fully rendered page (cf. elements).</p> <p>Web Element: Visual content of a webpage, which is displayed to the user on the rendered webpage. E.g. Text, Pictures, Widgets, Videos, etc.</p>
Web 页面显示时延	从用户点击 Web 链接或从用户输入 URL 到整个 Web 显示出来之间的时延。
空口受限	指业务会话的实际吞吐量受到无线空口可以为用户提供的传输能力限制，或者说当空口带宽不是瓶颈时，业务实际吞吐率会比当前更高
空口业务来水量	基站每秒收到的下行数据量，或终端应用层发送到终端无线模块的上行数据量
视频初始缓冲时延	从点击视频开始播放，到出现视频画面的等待时间

视频初始缓冲下载时延	从点击视频到出现视频画面之间，视频客户端与视频服务器之间有数据交互的那部分时延
------------	---

B 参考文献

1. Next Generation Mobile Networks Beyond HSPA & EVDO A White Paper By Board Of NGMN Limited
2. Next Generation Mobile Networks Radio Access Performance Evaluation Methodology By the NGMN Alliance
3. Understanding Website Complexity: Measurements, Metrics, and Implications Michael Butkiewicz UC Riverside butkiewm@cs.ucr.edu, Harsha V. Madhyastha UC Riverside harsha@cs.ucr.edu, Vyas Sekar Intel Labs vyas.sekar@intel.com
4. TCP / IP 详解, 卷 1: 协议
5. 李文斌, 耿博, “TCP 吞吐量理论分析 0.2”, 2009.6
6. M. Mathis, J. Semke, J. Mahdavi and T. Ott, “The macroscopic behavior of the TCP congestion Avoidance Algorithm,” Computer Communications Review, vol. 27, no. 3, pp 67-82, July 1997.
7. J. Padhye, V. Firoiu, D. Towsley and J. Kurose, “Modeling TCP Reno performance: A simple model and its empirical validation,” IEEE/ACM Trans. on Networking, vol. 8, no. 2, pp. 133-145, April 2000.
8. 韩涛, 朱耀庭, “考虑慢启动影响的 TCP 吞吐量模型”, 电子学报, vol. 30, no10, pp 1482-1483, 2002.10
9. 潘亚军, “TCP 吞吐量测量算法研究硕士论文”, 北京师范大学, 2005.6
10. RFC 2988, “Computing TCP's Retransmission Timer”, November 2000
11. Stas Khirman and Peter Henriksen. Relationship between Quality-of-Service and Quality-of-Experience for Public Internet Service. 2000.3 3950 Fabian Way, Palo Alto, CA 94303
12. Alessandro Anzaloni and Alexander Bento Melo. TCP PERFORMANCE IN WIRELESS CHANNELS. 2002.5. Instituto Tecnológico de Aeronáutica- BRAZIL