How YouTube Performance is Improved in the T-Mobile Network

Jie Hui★, Kevin Lau★, Ankur Jain†, Andreas Terzis†, Jeff Smith★
T Mobile★, Google†
Speakers

Jie Hui

Kevin Lau

Ankur Jain

Andreas Terzis

Jeff Smith
Executive Summary

**Background:** Since 2012, Google and T-Mobile have been working together to optimize the delivery of YouTube traffic in the T-Mobile network.

**Scope of this talk:** By bypassing TCP proxy for YouTube, T-Mobile users enjoyed a higher quality video experience while saving battery life at the same time.

Other optimization experiments will be discussed in next steps and future talks.

**Note:** YouTube traffic was not prioritized or treated preferentially in any other way.
Make the customer experience more enjoyable!

QoE is measurement of the enjoyment of the customer experience.
Mobile QoE 2.0

1.0  
Network Experience

1.5  
Real User Speed Experience

2.0  
Real User Application Experience

Network-instrumented  
Crowd-sourced from devices

Figure originally from: I. Grigorik, "High Performance Browser Networking: What every web developer should know about networking and web performance". O'Reilley Media, 2013
Unique Mobile Challenges
Customer Evolution

From

- High speed WiFi
- Wireless but mostly local/stationary
- Shared with few users
- Machines with access to power

To

- High speed cellular
- Full mobility and everywhere!!!
- Shared with high number of users
- Power limited devices
Mobile Data Usage is Growing Exponentially

Data Demand Growth

HTTP Traffic
Wireless 101

- At the RAN
  - Devices communicate with the eNodeB and the RAN is a complex system managing scheduling, wireless channel prediction/transmission, and mobility handovers.
  - The air medium is shared and inherently lossy; signals levels are time varying and different per user. User experience and device power use varies.

- At the Core
  - There are a number of gateways and proxies to “normalize” RAN & external network characteristics (packet loss, delay)
  - Concealment of packet drops allow the use of protocols like TCP over wireless
  - As RAN moves to LTE, Core network also need to evolve to improve data performance, especially for streaming video

Figure originally from: I. Grigorik, "High Performance Browser Networking: What every web developer should know about networking and web performance". O'Reilley Media, 2013
YouTube Video Delivery in Mobile Network
Life of a Mobile Video Playback

Radio Access Network

Core Network (EPC)

External Network

Figure originally from: I. Grigorik, “High Performance Browser Networking: What every web developer should know about networking and web performance”. O'Reilley Media, 2013
Mobile Proxies Hide Cellular Variability - Important for 2G & 3G

Split TCP

Low latency and variability
easy for TCP

Medium latency and variability
difficult for TCP

High latency and variability
even more difficult for TCP
Proxies Apply Various Optimizations to Traffic - Just in Time delivery is one example.

Direct Connection

May have more wastage depending on when/if the video is abandoned.

Proxied

Attempts to limit wastage by reducing the delivery rate.

Volume Consumed Over Time

Video Delivery Rate
Experiment

- *What happens if we bypass the proxy?*
  - Quality of User Experience
  - Network usage

- To answer this question we bypassed TCP proxies for YouTube traffic and measured difference
Experiment and Results
Being a good network citizen

- Without the proxies, the servers need to protect and fully utilize the network
  - Chunked delivery
  - FQ/Pacing: limit Tx rate
    - Especially important after being idle
    - Tx rate ≤ 2*cwnd/RTT
  - cwnd clamp: limit the amount of Data In transit
    - cwnd ≤ Bandwidth Delay Product
  - TCP stack that tolerates RTT variance and packet reordering
Evaluation Metrics I: Network level

1. Distributions of
   a. Minimum round trip times (RTT)
   b. Retransmission rates
   c. Connection throughputs
   d. median(RTT)/min(RTT)

2. Network Usage
Comparing distributions: Quantile-Quantile plots

- We have distributions of proxied and direct connections
- How do we compare the two distributions?
- One answer: Quantile-Quantile plot
Direct connections have fewer retransmitted bytes.

Direct is better than Proxied.
Direct connections have higher throughput

Throughput Distribution

- Direct
- Proxied

Lower throughput due to clamp

Lower throughput peak
Proxied connections are more bloated
Removing proxy did not significantly change overall network traffic

Slight increase in busy hour and daily volume
No significant change in other metrics
1. Join Latency: $T_1 - T_0$
2. Playback time: $T_P = (T_2 - T_1) + (T_4 - T_3) + (T_7 - T_6) + (T_9 - T_8)$
3. Total Rebuffer time: $T_3 - T_2$
4. Battery Lifetime (Power consumed during $[T_0, T_9]$)
Worse Performance

Direct is better

Direct connections rebuffer for less time
Direct connections have lower join latency

Join Latency (sec)
~20% Reduction in Battery Consumption

Direct Connection

Measured Power Data

Direct Connection Battery Savings

<table>
<thead>
<tr>
<th>Access</th>
<th>2.3 Min</th>
<th>5+ Min</th>
<th>15+ Min</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTE</td>
<td>20.11%</td>
<td>19.28%</td>
<td>21.62%</td>
<td>21.84%</td>
</tr>
<tr>
<td>HSPA</td>
<td>16.49%</td>
<td>15.10%</td>
<td>19.64%</td>
<td>19.46%</td>
</tr>
</tbody>
</table>

Proxied Connection

Measured Power Data

Proxied

Direct

PCH
Low power, listening only

FACH
Medium Power, Shared Channel, low speed

DCH
High Power, dedicated channel, High speed
Summary Findings from Proxy Bypass

Bypassing proxies

- Lowers retransmission rates, increases throughput, and decreases bufferbloat
- Does not negatively impact network traffic
- Improves quality of experience for video
- Increases battery lifetime
What else?
Other components impacting QoE
(completed, in progress)

- Burst optimization, Congestion Notification (In/Out band)
- YouTube workload analysis shows caching at base station not effective in reducing congestion on RAN backhaul.
- GGC supports IPv6. Also bypasses CGN device that could impact performance
- Deploy GGC to serve video from location closer to user.

![Diagram of network components: eNodeB, Core Network (CN), Radio Access Network (RAN), Internet]
Examples of RAN optimizations

- **Burst Optimization**
  - Send bigger chunks at longer intervals to amortize cost of radio wakeup/timeout

- **Pacing traffic at the server for congested cells**
  - in-band (e.g., ECN)
  - out-of-band

Such optimizations require applications to have knowledge of partial network configuration/state
Open exchange of information

- T-Mobile and Google experimenting with APIs for programmatic exchange of information
  - T-Mobile → Google
    - Congestion level/radio state transition timers of cells, throughput guidance, etc
    - Use this information for server/client optimizations
  - Google → T-Mobile
    - Aggregated, anonymized QoE metrics
    - Diagnostics to identify problems
- Exploring out-of-band, in-band and device APIs
- Future work to standardize these APIs so that other carriers & content providers also benefit
Take Away

• Mobile QoE 2.0 requires a new collaborative approach

• Future directions to standardize
  – mobile application RUM
  – application awareness of network

• Connect with us
  - T-Mobile: DeviceQoELab@T-Mobile.com
  - Google: mobiledelivery@google.com