# Accelerating YouTube & Google Search

Andreas Terzis

# YouTube Statistics

- YouTube is a large fraction of Internet traffic *globally*[1]
  - 17% NA, 25% Europe, 33% LATAM, 23% APAC of fixed-line traffic
- Mobile makes ~40% of YouTube's global watch time
- Over 6B hours of video watched each month on YouTube[2]
- 100 hours of video are uploaded to YouTube every minute
- ~8M users concurrently saw Felix Baumgartner jump from space

[1] https://www.sandvine.com/downloads/general/global-internet-phenomena/2013/2h-2013-global-internet-phenomena-report.pdf
[2] http://www.youtube.com/yt/press/statistics.html

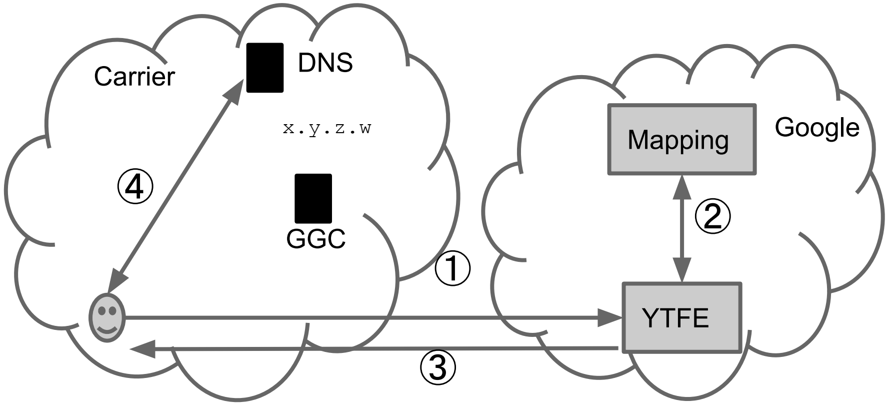
# How YouTube works: Mapping

① Client issues HTTP(S) request for manifest from YT Front End: `GET /watch?v=n_6p-1J551Y`

② Mapping infrastructure determines cache that the user should contact

③ YTFE returns Manifest with videoplayback URLs for different encoding schemes/rates/video sizes

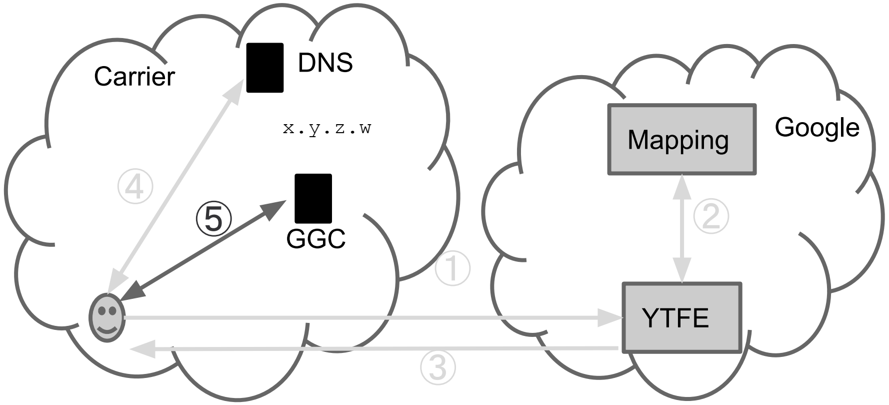
④ Client resolves `xxx.googlevideo.com -> x.y.z.w` (inside carrier's addr space)



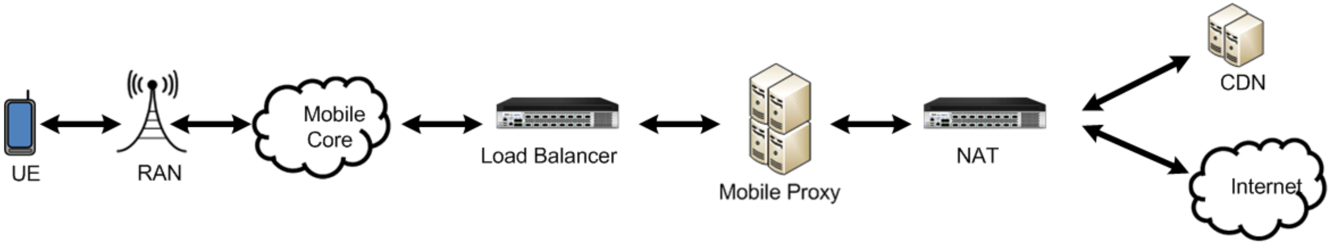
# How YouTube works today: Video Playback

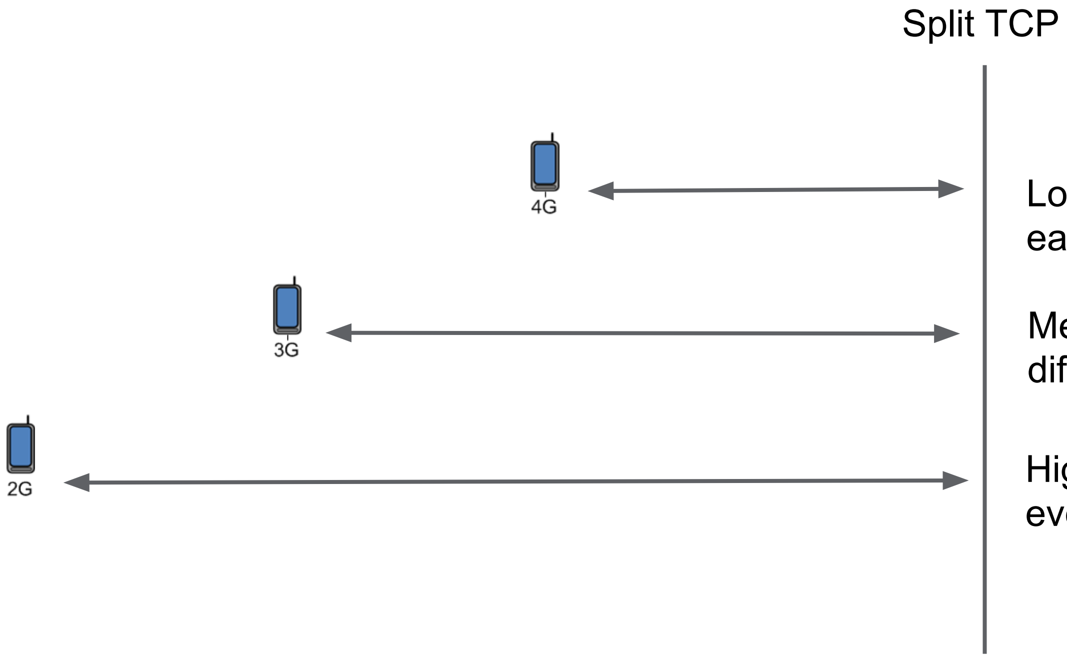
⑤ Client issue HTTP(S) videoplayback requests from nearest Google Global Cache (GGC)

- HTTP range requests for *video chunks* (100's KB - MB)
- ABR algorithm at the client determines requested format for the next video chunk
  - ABR selection depends on multiple factors: network rate, screen size, client resources, etc.



# Delivering Video to Mobile Networks: TCP Proxies





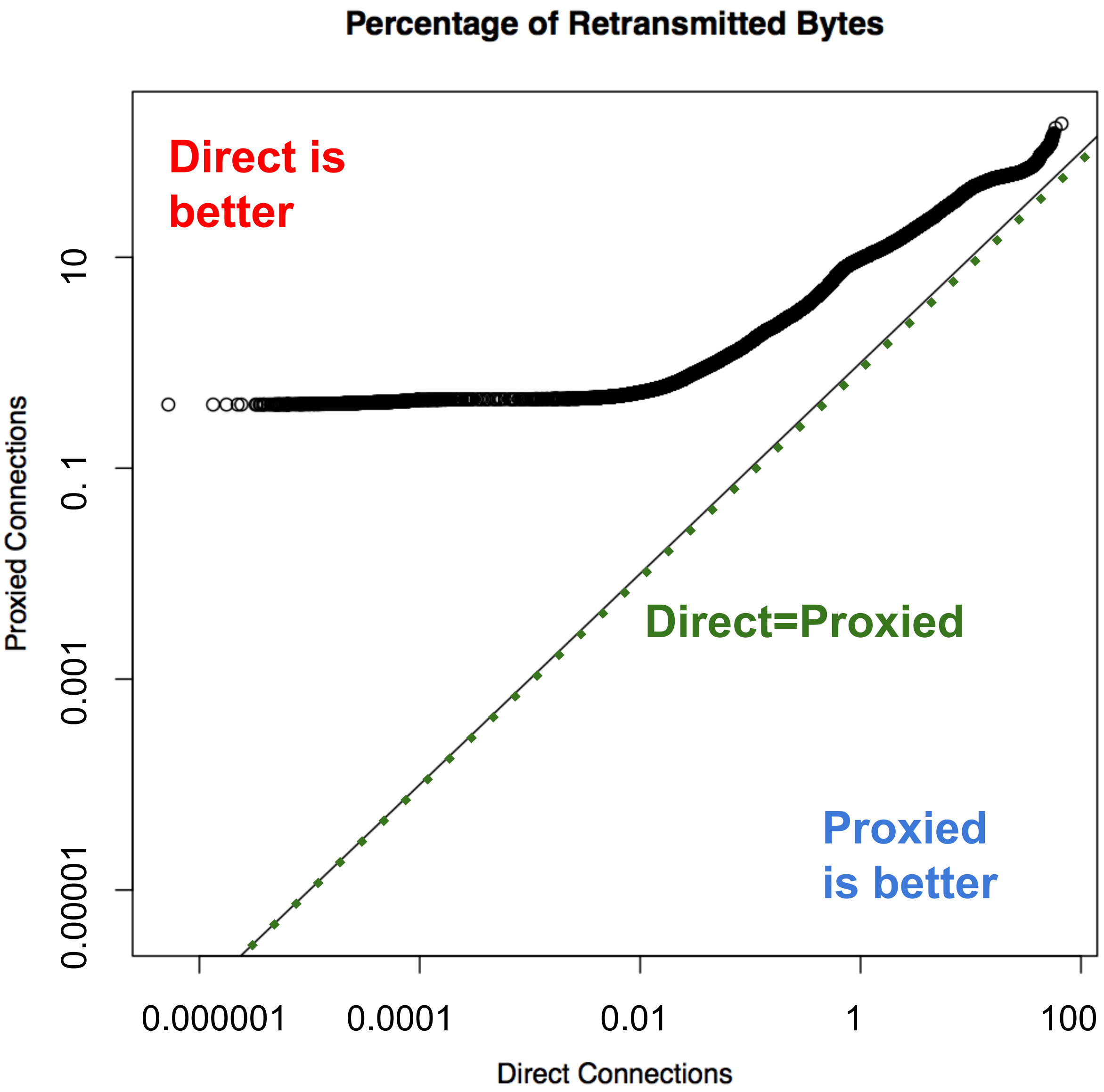
Low latency and variability easy for TCP

Medium latency and variability difficult for TCP

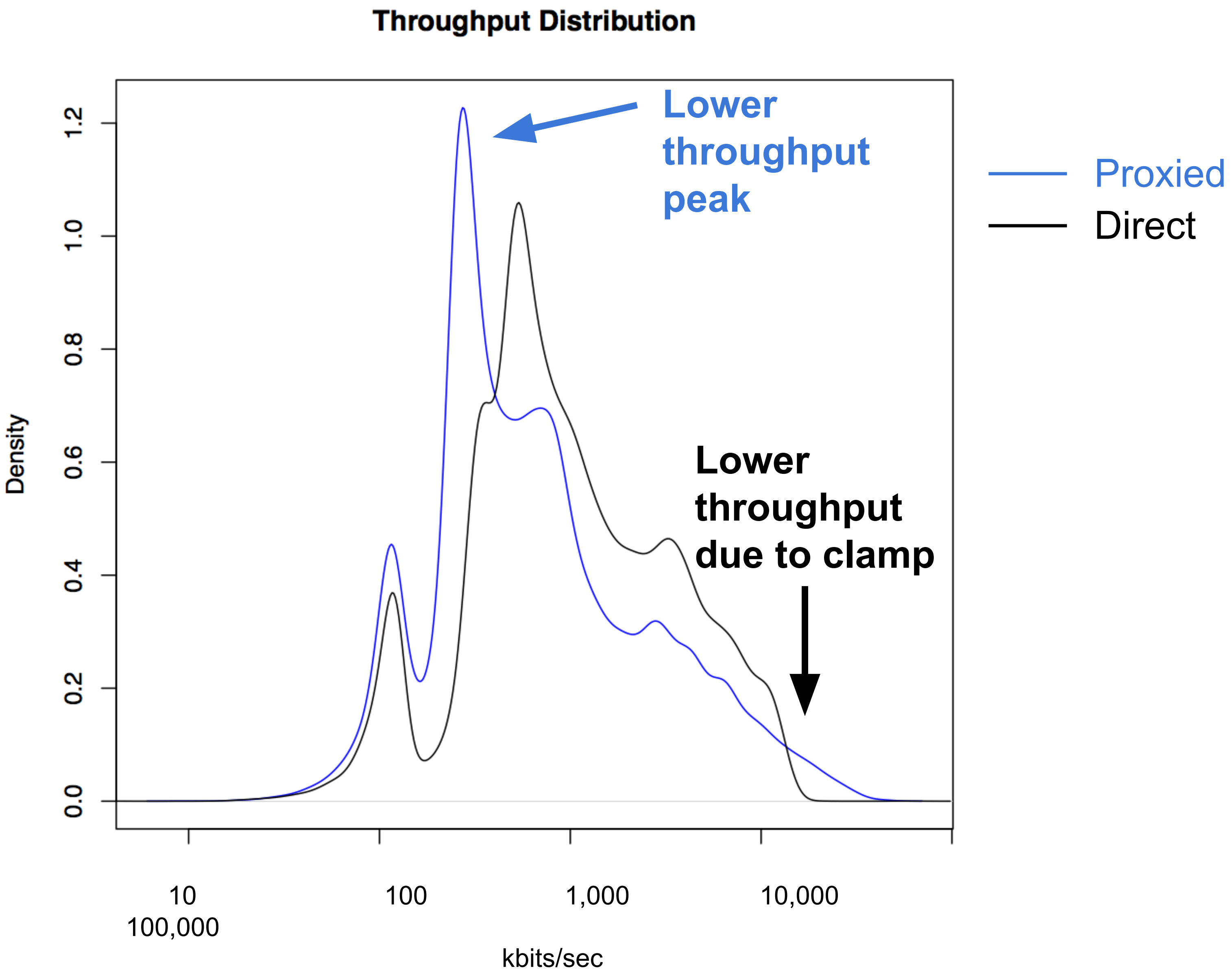
High latency and variability even more difficult for TCP

# Challenging the assumption

- Conventional wisdom suggests that TCP proxies improve performance in cellular networks
- *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



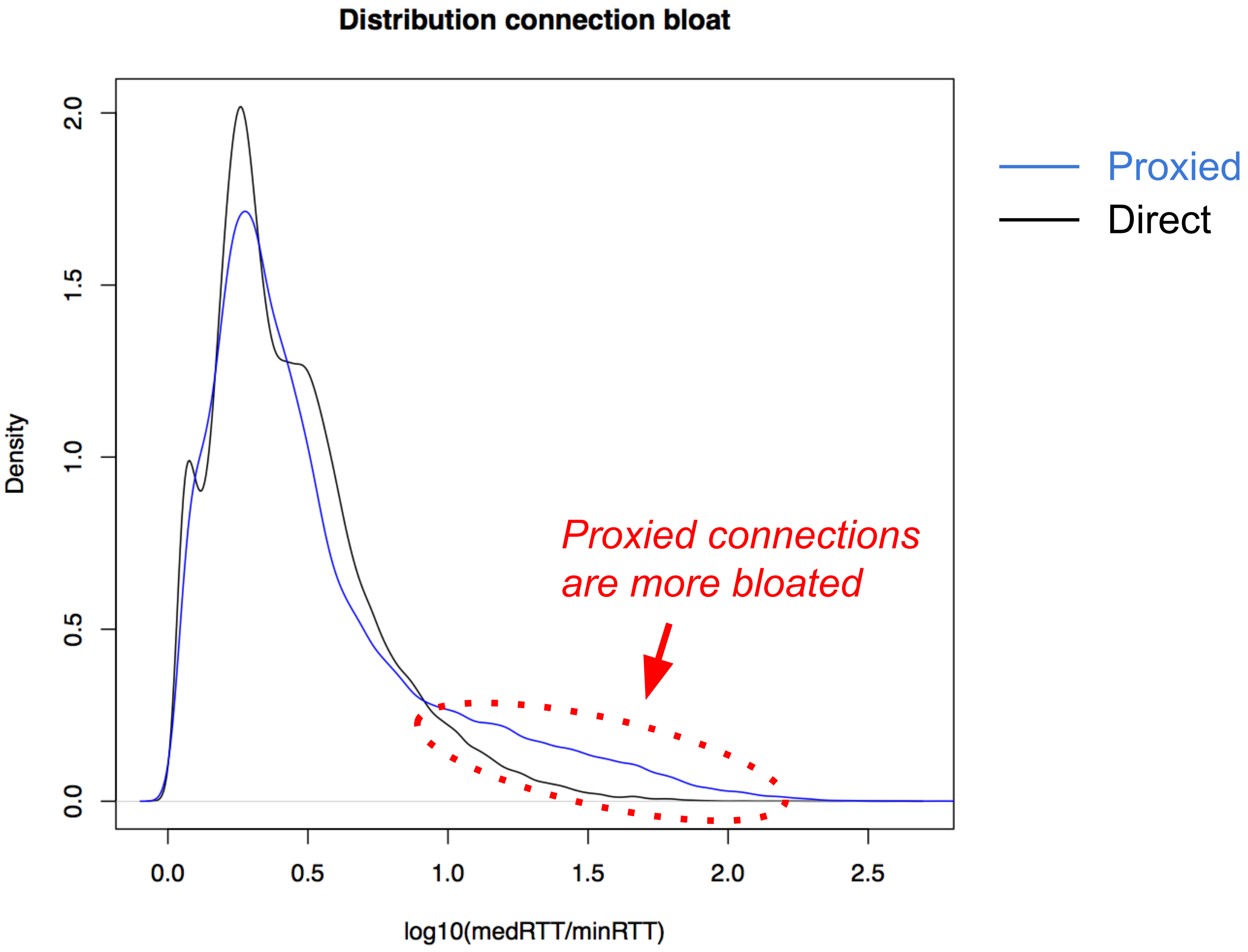
*Direct connections have fewer retransmitted bytes*



**Lower throughput peak**

**Lower throughput due to clamp**

*Direct connections have higher throughput*



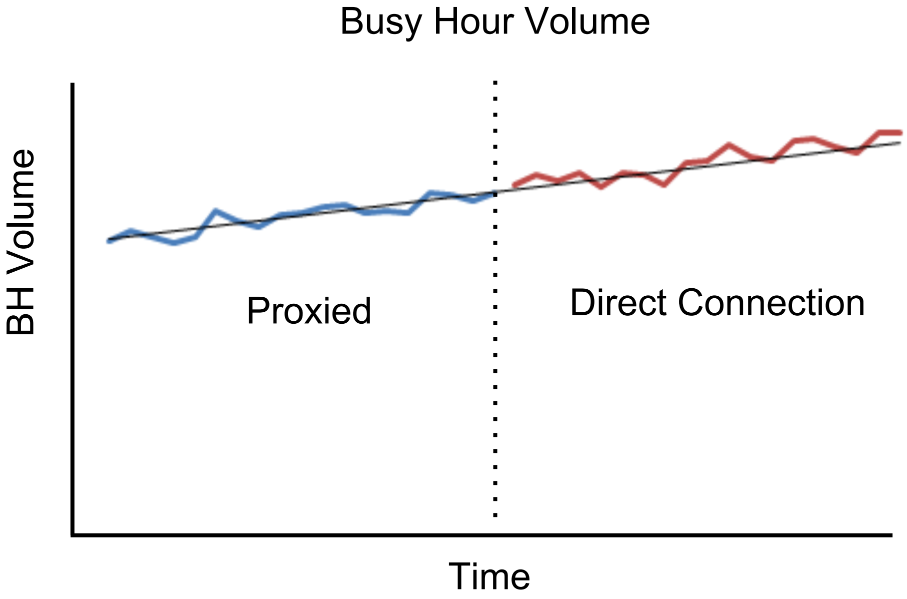
Distribution connection bloat

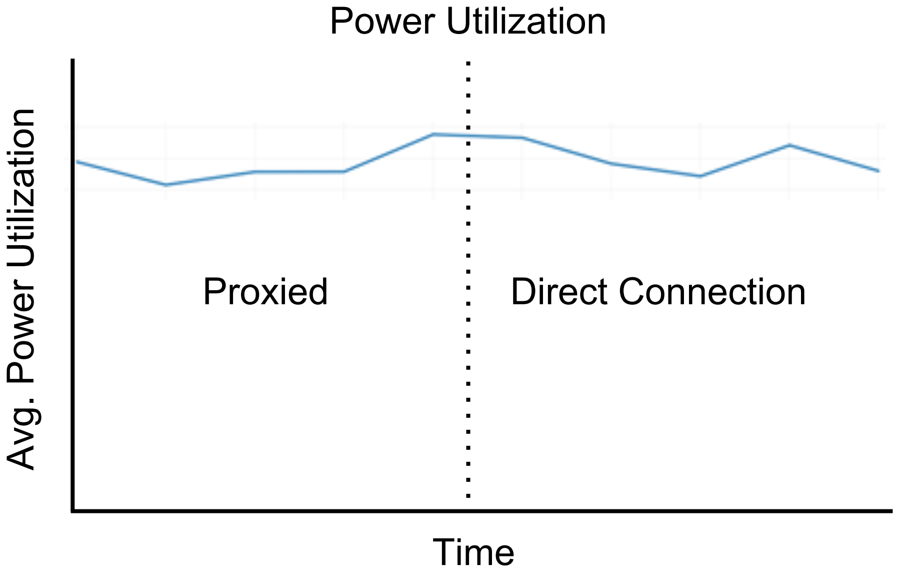
*Proxied connections are more bloated*

# Removing proxy did not significantly change overall network traffic

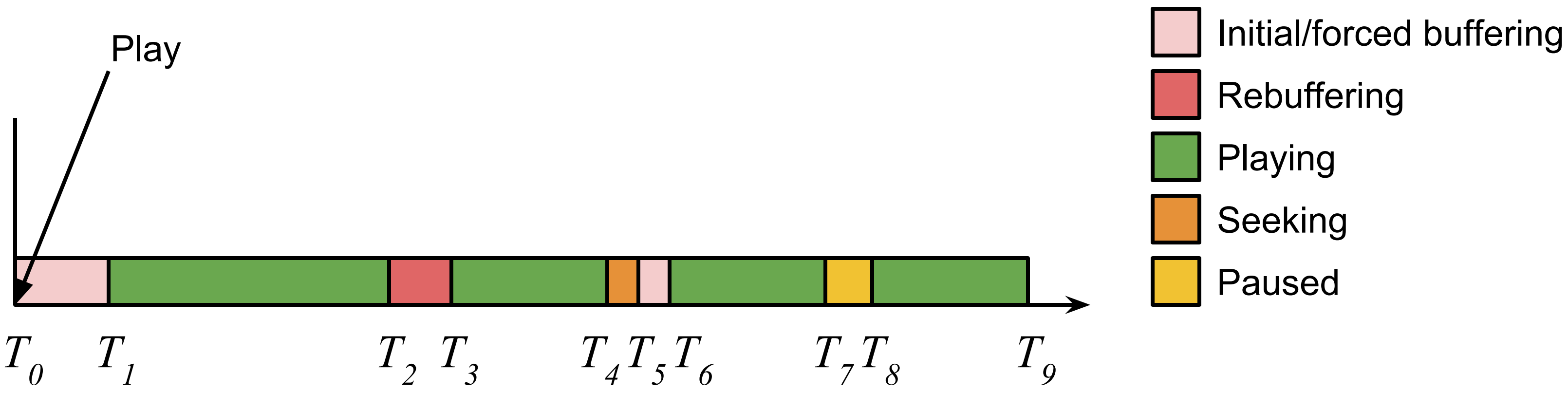
Slight increase in busy hour and daily volume

No significant change in other metrics



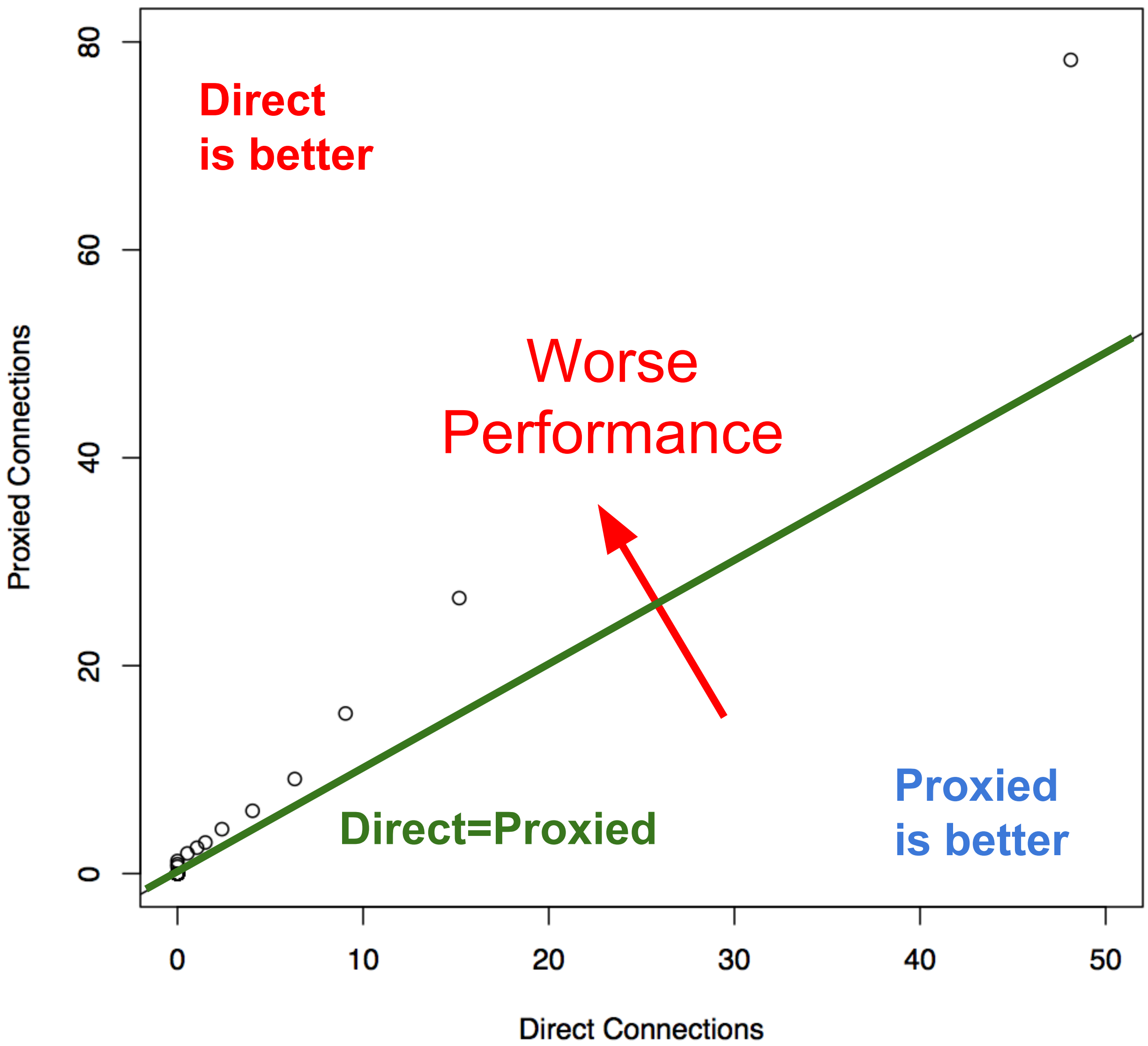


# Evaluation Metrics II: Quality of Experience

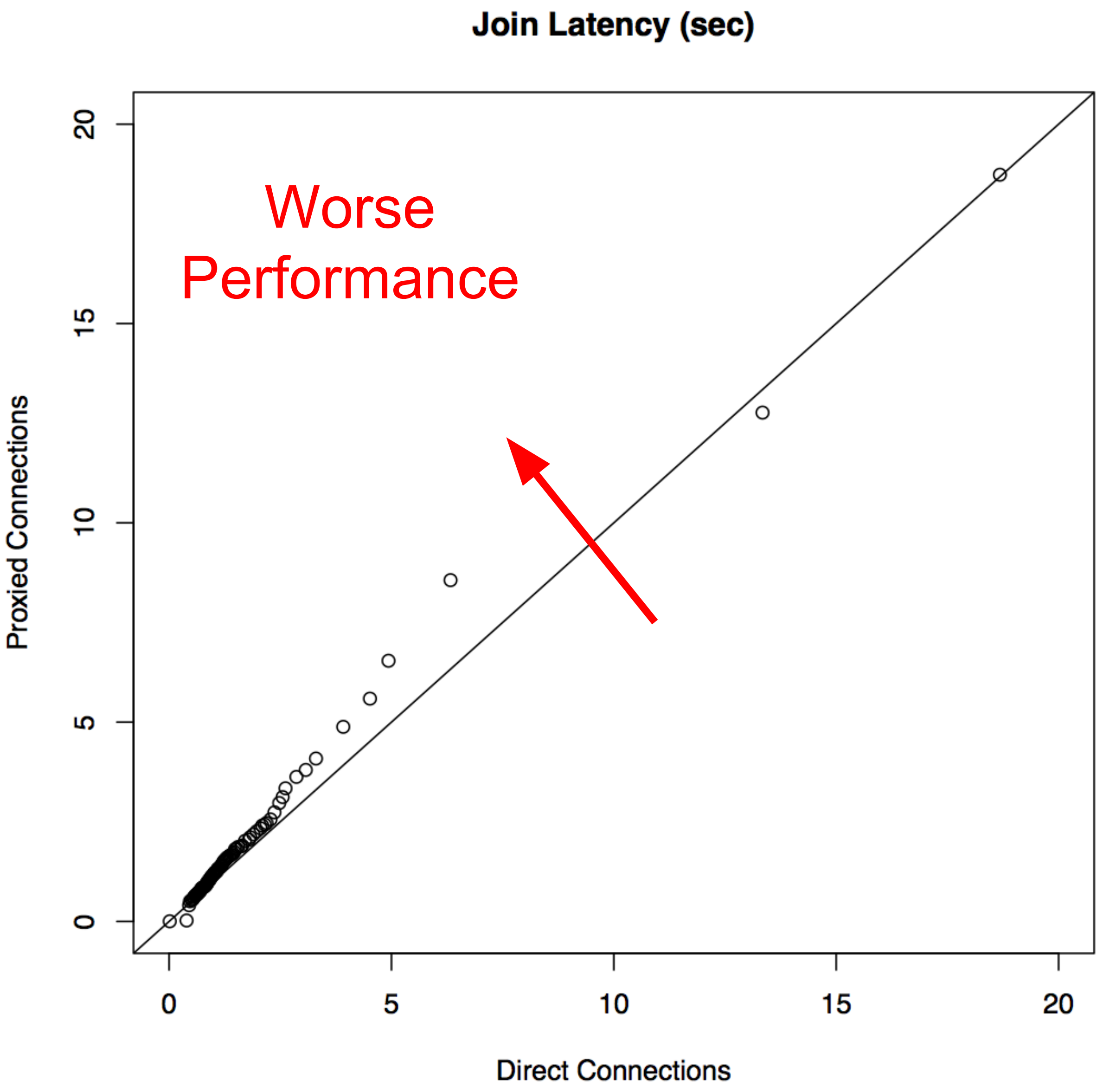


1. Join Latency: $T_1$-$T_0$
2. Playback time: $T_P = (T_2 - T_1) + (T_4 - T_3) + (T_7 - T_6) + (T_9 - T_8)$
3. Total Rebuffer time: $T_3$-$T_2$
4. Battery Lifetime (Power consumed during $[T_0, T_9]$)

## Total Rebuffer Time (sec)



*Direct connections rebuffer for less time*

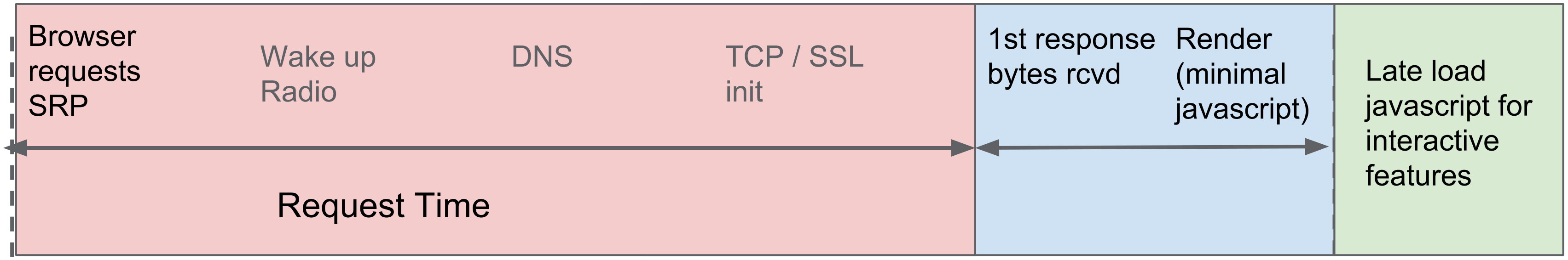


Worse Performance

*Direct connections have lower join latency*

# Decreasing web search latency on 2G networks

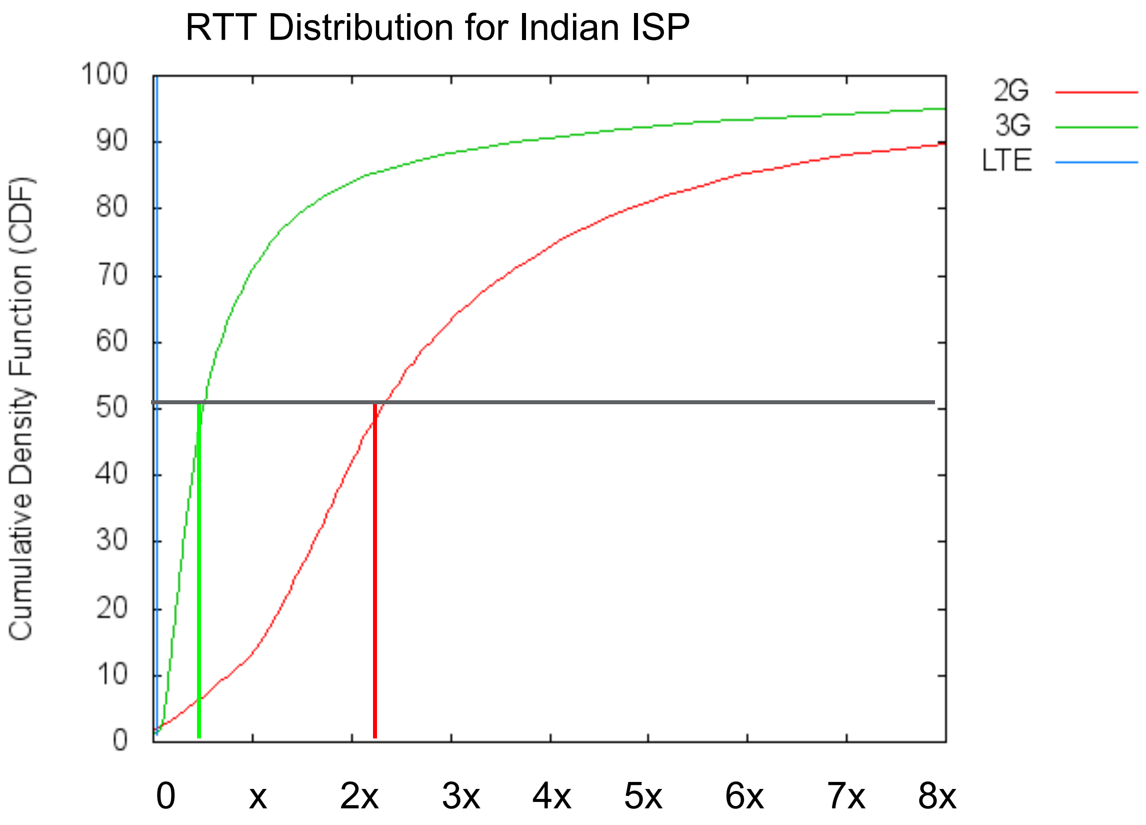
- Large portion of users in emerging markets access the Internet over 2G networks
- End-to-end Latency is 2 components
- Byte reduction can only improve Response Receipt / Render



- Request time is driven by RTT to closest Google front end (= 4* RTTs for HTTPS)

# RTT as a function of network type

- RTT between UEs and closest Google server
- Considerable variation in RTT
- *Where is the variation coming from?*


RTT Distribution for Indian ISP
Cumulative Density Function (CDF)
x
2x
3x
4x
5x
6x
7x
8x
2G
3G
LTE