

Behind the Curtain

The importance of replica selection in next generation cellular networks

John P. Rula Fabián E. Bustamante
Northwestern University

1. INTRODUCTION

Smart devices are becoming the primary or only Internet point of access for an ever larger fraction of the population. Nearly a quarter of current web traffic is mobile, and recent industry studies have estimated a fourfold increase on global mobile data traffic by 2018, mainly driven by the content demands and growing number of smart phones and tablets [2]. The most recent CISCO VNI report estimates that by 2018, the majority of North America devices and connections will have 4G capability and, while 4G will be 15% of world-wide connections then, these connections will be responsible for 51% of traffic.

Cellular networks pose a challenge to content delivery networks (CDNs) given their opaque network structure, limited number of ingress points, and obfuscated DNS infrastructure. Previously, large cellular radio latencies meant CDN replica selection had little impact on the total end-to-end latency. However, the advancement of 4G networks such as LTE has lowered mobile device access latency to make it comparable with many existing broadband services, making the choice of content replica server a significant contributor to end-to-end performance.

In general, but particularly in cellular networks, CDNs have limited signals for locating clients. Mobile IPs have been shown to be dynamic for mobile end hosts [1], and external entities such as CDNs are prevented from probing their mobile clients or their infrastructure by NAT and firewall policies implemented by cellular operators.

In this poster, we present preliminary work looking at the impact of replica selection in next generation cellular networks. Using a collection of over 250 mobile end-hosts over a two-month period, we explore CDN replica selection in cellular networks measuring the latency to content replicas for a selection of popular mobile websites. We find that clients in next generation radio technologies can see up to 400% differences in latency to selected replicas. We discover that, in large part, these poor selections are due to current localization approaches employed by CDNs such as DNS redirection which, while fairly effective in wired hosts, performs rather poorly within cellular networks mainly due to cellular DNS behavior.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). Copyright is held by the author/owner(s).

SIGCOMM'14, August 17–22, 2014, Chicago, IL, USA.

ACM 978-1-4503-2836-4/14/08.

<http://dx.doi.org/10.1145/2619239.2631465>.

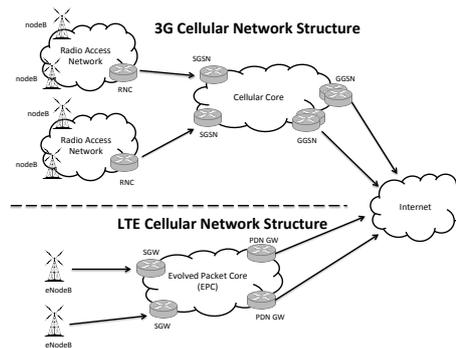


Figure 1: Network architecture evolution from 3G to LTE.

2. MOTIVATION

CDNs are responsible for delivering most of today's Internet content, replicating popular content on servers worldwide. CDNs direct users to "nearby" replicas based on the location of users' DNS resolver.

The significantly better performance of next generation cellular networks, like LTE, compared with 2G/3G networks have made content replica selection a significant factor of a mobile user's experience. When compared with 3G networks, 4G LTE offers a flatter network structure and significantly lower radio access latency and variance. For instance, median DNS resolution times for clients on LTE ranged from 30-50ms, while median resolution times for the same clients on 3G networks ranged from 100-200ms. Figure 1 illustrates the difference in network structures between 3G and 4G network architectures, including the elimination of the Radio Network Controller (RNC) which allows a direct link from each cellular tower to the network core.

We observe a wide range of performance diversity of content replicas seen by each client. We measured the HTTP time-to-first-byte to replicas for a collection of websites. By aggregating all replica servers seen by clients in our experiments, and taking the average client latency to each replica, we found the lowest latency replica for each website. All other replicas are shown as a percentage difference between their latency and the "best" seen replica. The cumulative distribution of these ratios is shown in Fig. 2.

While the degree of replica differential performance varies per operator, we consistently found replica latency increases ranging from 50% to 100% in all networks. In an extreme case, we observed replicas with 400% increased latency over the closest observed replica server for clients in Sprint.

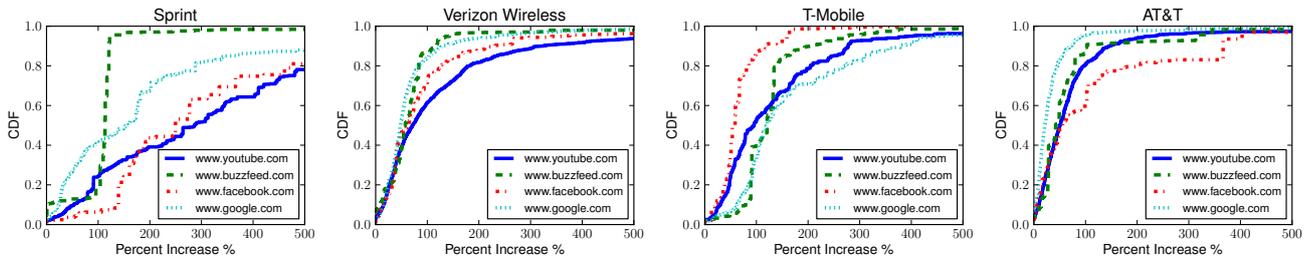


Figure 2: Performance comparison of HTTP time-to-first-byte for replica servers selected for mobile clients. We see cellular users directed toward a wide range of replica servers, with certain servers seeing 400% increased HTTP latency over the closest observed replica.

3. CELLULAR DNS BEHAVIOR

Many of the largest CDNs rely on DNS for client location/replica selection in cellular networks. Our results show that cellular DNS makes a poor indicator for client location, due to the opaqueness of cellular resolvers to external services, and the inconsistency between mobile clients and their visible LDNS resolvers.

CDNs typically aggregate client resolvers behind traceroute divergence points and map clients based on measurements to these points. Unlike the majority of DNS resolvers, cellular DNS resolvers are unable to be probed and measured by CDNs. The inability to traceroute the cellular DNS resolvers invalidates this approach.

We discovered inconsistency between mobile clients and their cellular DNS resolvers over time. In other words, a CDN will see the same client originating from different locations, and depending on the extent of DNS load balancing, users can see high variability in the quality and performance of selected content replicas.

The movement of a client's visible DNS resolver is shown in Fig. 3 for users in T-Mobile and AT&T. The figure plots the enumerated LDNS resolvers observed by mobile clients. The blue marks on the bottom plot represent individual IP addresses seen and the red marks on the top represent resolver /24 prefixes observed by the mobile client over time. Surprisingly, many of the different DNS resolvers observed exist in separate /24 prefixes. This shuffling between resolvers in different prefixes is important because we found several CDNs mapping replica servers to resolver /24 prefixes.

4. CELL NETWORKS HELP

Our work reveals several hurdles for cellular content delivery. Without the cooperation of cellular operators, CDNs have limited options for locating clients within cellular networks. Previous work has found the temporal inconsistency with cellular client IP addresses [1], and the previous section outlined our findings showing the inconsistency between mobile clients and their visible resolver. Additionally, our findings uncovered inconsistency in anycast addresses, limiting the appeal of anycast CDN services.

5. REFERENCES

- [1] M. Balakrishnan, I. Mohamed, and V. Ramasubramanian. Where's that phone?: geolocating ip addresses on 3g networks. In *Proc. IMC*, 2009.
- [2] CISCO. CISCO visual networking index: Global mobile data traffic forecast update 2013-2018. Technical report, CISCO Systems Inc., 2014.
- [3] Q. Xu, J. Huang, Z. Wang, F. Qian, and A. G. Z. M. Mao. Cellular data network infrastructure characterization and implication on mobile content placement. In *Proc. ACM SIGMETRICS*, 2011.

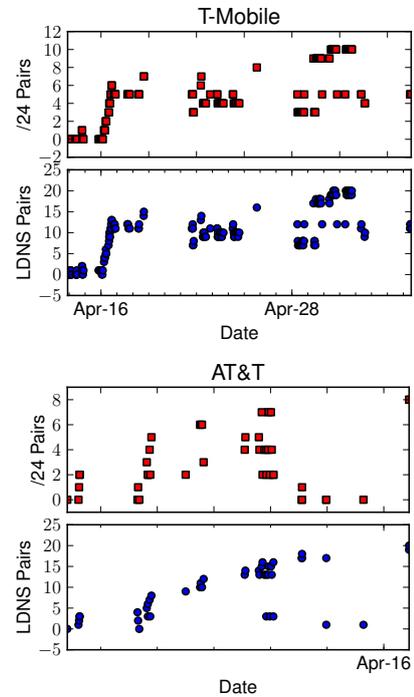


Figure 3: Number of external resolvers observed by a client in AT&T and T-Mobile. Bottom: number of external resolver IP addresses. Top: number of unique /24 prefixes observed by resolvers. Client DNS resolvers change not just within localized clusters, but span multiple /24 prefixes over time.