Contents lists available at ScienceDirect
Vehicular Communications
www.elsevier.com/locate/vehcom

The impact of vehicular traffic demand on 5G caching architectures: A data-driven study

Francesco Malandrino [a], Carla-Fabiana Chiasserini [a,b], Scott Kirkpatrick [c]

[a] Politecnico di Torino, Italy
[b] CNR-IEIIT, Italy
[c] The Hebrew University of Jerusalem, Israel

ARTICLE INFO

Article history:
Received 8 June 2016
Received in revised form 17 October 2016
Accepted 11 November 2016
Available online xxxx

Keywords:
???

ABSTRACT

The emergence of in-vehicle entertainment systems and self-driving vehicles, and the latters' need for high-resolution, up-to-date maps, will bring a further increase in the amount of data vehicles consume. Considering how difficult Wi-Fi offloading in vehicular environments is, the bulk of this additional load will be served by cellular networks. Cellular networks, in turn, will resort to caching at the network edge in order to reduce the strain on their core network – an approach also known as *mobile edge computing*, or "fog computing".

In this work, we exploit a real-world, large-scale trace coming from the users of the We-Fi app in order to (i) understand how significant the contribution of vehicular users is to the global traffic demand; (ii) compare the performance of different caching architectures; and (iii) studying how such a performance is influenced by recommendation systems and content locality. We express the price of "fog computing" through a metric called *price-of-fog*, accounting for the extra caches to deploy compared to a traditional, centralized approach. We find that "fog computing" allows a very significant reduction of the load on the core network, and the price thereof is low in all cases and becomes negligible if content demand is location specific. We can therefore conclude that vehicular networks make an excellent case for the transition to mobile-edge caching: thanks to the peculiar features of vehicular demand, we can obtain all the benefits of "fog computing", including a reduction of the load on the core network – reducing the disadvantages to a minimum.

1. Introduction

Back in 2010, the traffic demand of newly-introduced iPhones briefly disrupted some cellular networks [1]. It is uncertain whether such disruptions are likely to happen again; however, there is no doubt that *if* they do happen, vehicular users will be among the main culprits.

The reason for this trend is multifold. First, vehicles carry people, and people carry multiple, data-hungry mobile devices. Second, vehicles themselves are increasingly often equipped with entertainment devices, which only add to the problem. Third, vehicles download navigation data, e.g., map updates: while this is a minor component of the overall traffic today, it is expected to increase by orders of magnitude with the introduction of self-driving vehicles, which will need much more detailed and up-to-date maps.

To make things worse, virtually *all* such data demand will be served by cellular networks. Indeed, most offloading solutions target pedestrian users, because their position changes relatively slowly over time and because they are more likely to be covered by such networks as Wi-Fi.

Caching is a primary way in which cellular network operators plan to react to this demand surge. One of the most popular solutions is to move caches as close as possible to the users, in the context of an approach known as *fog computing* (a term created by Cisco [2]). It is expected that doing so will increase the cache hit ratio while reducing the service latency as well as the traffic load on the cellular core network. On the negative side, it will require deploying multiple, smaller caches. Additional help is expected from recommendation systems, whose effect is to shape the demand concentrating it around the most popular content items. Intuitively, having fewer, popular items to serve will improve caching performance.

In this context, our paper targets three main questions.

E-mail address: francesco.malandrino@polito.it (F. Malandrino).

http://dx.doi.org/10.1016/j.vehcom.2016.11.007
2214-2096/

Table 1
The Los Angeles dataset.

Metric	Value
Time of collection	Oct. 2015
Total traffic	35 TByte
Number of records	81 million
Unique users	64,386
Unique cell IDs	47,928
Mobile operators	AT&T (16,992)
(number of cells)	Sprint (2,764)
	T-Mobile (24,290)
	Verizon (3,882)

Vehicular demand. What is the data demand generated by today's vehicular users? Which apps and services represent the most significant contributions thereto?
Caching architectures. Given a target hit ratio, what is the relationship between caching architecture and size of the caches we need to deploy? What is the impact of moving caches from core-level switches to individual base stations, on the total cache size, on the distance data must travel within the core network, and on the load thereof? What changes if a recommendation system is in place?
Location-specific content. Content items consumed by future vehicular networks are expected to strongly depend on the location – augmented maps for self-driving vehicles being the most obvious example. What will be the impact of this kind of content on caching?

We answer these questions using a set of real-world, large-scale measurement data, coming from users of the WeFi app [3]. Due to its crowd-sourced nature, our dataset includes data for: (i) multiple apps, including video (e.g., YouTube) and maps; (ii) multiple types of users, from pedestrian to vehicular ones; (iii) multiple network technologies, including 3G, LTE, and Wi-Fi; (iv) multiple operators.

We describe our dataset, as well as the additional processing we need to perform in order to enhance the information it provides, in Section 2. Then, in Section 3 we explain how we model caching and caching architectures in our vehicular scenario. Section 4 presents numerical results and some relevant insights we obtain from them. Finally, Section 6 concludes the paper and sketches future work directions.

2. Input data

We describe the WeFi dataset we have access to in Section 2.1. Then in Section 2.2 we detail the processing steps we need, in order to extract further information that is not directly included therein. Finally, Section 2.3 explains how we complement the available information using other datasets and well-known information.

2.1. The WeFi dataset

Our data comes from the users of an app called WeFi [3]. WeFi provides its users with information on the safest and fastest Wi-Fi access points available at the user's location. At the same time, it collects information about the user's location, connectivity and activity. WeFi reports over seven million downloads of the app globally, and over three billion daily records. In this work, we use the dataset relative to the city of Los Angeles – a vehicle-dominated environment. Its main features are summarized in Table 1.

Each record contains the following information:

- day, hour (a coarse-grained timestamp);
- anonymized user identifier and GPS position;
- network operator, cell ID, cell technology and local area (LAC) the user is connected to (if any);
- Wi-Fi network (SSID) and access point (BSSID) the user is connected to (if any);
- active app and amount of downloaded/uploaded data.

If the location of the user or the networks she is connected to change within a one-hour period, multiple records are generated. Similarly, one record is generated for each app that is active during the same period. The fact that location changes trigger the creation of multiple records allows us to assess whether, and how much, each user moves during each one-hour period. As we will see in Section 2.2, this is instrumental in distinguishing between static and vehicular users. Combining this knowledge with network technology information allows us to ascertain which types of traffic cellular networks ought to worry about.

Fig. 1 shows the cell deployment of the four main operators present in our trace. We can see that all operators cover the whole geographical area we consider, but using radically different strategies. T-Mobile and, to a lesser extent, AT&T, deploy a large number of cells, each covering a comparatively small area. Sprint and, especially, Verizon, follow the opposite approach: their networks are composed of relatively few cells, each covering a fairly large area.

This fundamental difference reflects on the topologies of each operator's core network, and potentially on the effectiveness of different caching architectures. It is worth to stress that using a real-world, crowd-sourced trace such as ours, we are able to properly account for these factors, which are typically neglected by more abstract models.

2.2. Further data processing steps

From the WeFi dataset we easily identify several types of users and the content they consume.
User type. The WeFi app can be installed on a variety of mobile devices. The users carrying them can be static (e.g., sitting in a café), pedestrian (e.g., walking or jogging), or vehicular. We discriminate among these cases by looking at the *distance* covered by each user during each one-hour period. Fig. 2(a) shows the distribution thereof: we have almost 40% of static users, which do not move at all, a large number of pedestrian users covering moderate distance, and some users covering larger ones.

Given our focus on vehicular applications, it is important not to include pedestrian or static users in our analysis. To this end, conservatively we label as vehicular those users that travel more than 5 km in any one-hour period.[1] Fig. 2(b) shows the fraction of vehicular users as a function of time, and exhibits the familiar morning and afternoon peaks.
Content type. As recalled in Section 2, records contain an `app` field, containing the class name of the active application, e.g., `COM.GOOGLE.ANDROID.APPS.YOUTUBE.KIDS`. However, we cannot use this information directly for two main reasons. First and foremost, different class names may correspond to the same app, e.g., both `COM.GOOGLE.ANDROID.APPS.YOUTUBE.KIDS` and `COM.GOOGLE.ANDROID.YOUTUBE` correspond to YouTube. Furthermore, we are not only interested in individual apps, but also in the *category* they belong to, as summarized in Table 2.

It is important to point out that different content categories lend themselves to caching to radically different extents. Caches are virtually useless for real-time streaming content (while LTE broadcasting [4] represents a more promising alternative). On-demand video content can be successfully cached, especially if popular. Sport and news content is even easier to cache, as there is a limited number of items that are likely to be requested (e.g., the

[1] Notice that the same user can be vehicular in some time periods and static in others.

Please cite this article in press as: F. Malandrino et al., The impact of vehicular traffic demand on 5G caching architectures: A data-driven study, Veh. Commun. (2016), http://dx.doi.org/10.1016/j.vehcom.2016.11.007