Streaming Content from a Vehicular Cloud

Luigi Vigneri, Thrasyvoulos Spyropoulos, Chadi Barakat

Streaming Content from a Vehicular Cloud

Luigi Vigneri
Institut EURECOM
450 Route des Chappes
Biot, France
name.surname@eurecom.fr

Thrasyvoulos Spyropoulos
Institut EURECOM
450 Route des Chappes
Biot, France
name.surname@eurecom.fr

Chadi Barakat
INRIA
2004 Route des Lucioles
Valbonne, France
name.surname@inria.fr

ABSTRACT

Network densification via small cells is considered as a key step to cope with the data tsunami. Caching data at small cells or even user devices is also considered as a promising way to alleviate the backhaul congestion this densification might cause. However, the former suffers from high deployment and maintenance costs, and the latter from limited resources and privacy issues with user devices. We argue that an architecture with (public or private) vehicles acting as mobile caches and communication relays might be a promising middle ground. In this paper, we assume such a vehicular cloud is in place to provide video streaming to users, and that the operator can decide which content to store in the vehicle caches. Users can then greedily fill their playout buffer with video pieces of the streamed content from encountered vehicles, and turn to the infrastructure immediately when the playout buffer is empty, to ensure uninterrupted streaming. Our main contribution is to model the playout buffer in the user device with a queuing approach, and to provide a mathematical formulation for the idle periods of this buffer, which relate to the bytes downloaded from the cellular infrastructure. We also solve the resulting content allocation problem, and perform trace-based simulations to finally show that up to 50% of the original traffic could be offloaded from the main infrastructure.

Keywords

Caching; Opportunistic networks; Vehicular networks; Mobile data offloading; Multimedia Streaming; Optimization

1. INTRODUCTION

The diffusion of low-cost handheld devices has led to a large increase in the mobile traffic demand in the past few years. Specifically, the vast majority of the traffic concerns videos, and new streaming services have been recently introduced in the market (e.g., Netflix, Amazon Prime). In practice, current and near-future architectures (e.g., 3G, LTE) cannot keep up with such increase which is already overloading the cellular infrastructure [12]. Small cell densification has been proposed to improve the data rates and Quality of Experience (QoE) for users. While this solution decreases the load on the backbone, it actually provokes two problems: (i) small cells (SCs) require high CAPEX/OPEX costs to provide a dense enough radio access and backhaul transport network; (ii) SC deployments move the traffic overload from the core of the network to the backhaul. In order to solve (ii), several studies propose to cache popular content at the edge of the network (e.g., femtocells, picocells) [15, 27]. While caching in femtocells is useful to alleviate the backhaul congestion, it requires additional installation and maintenance costs. Other studies consider using mobile user equipment (UE) as relays and storage points [9, 16]; while more affordable for an operator, this solution faces significant technology adoption concerns, as mobile devices have limited storage capacity and strict battery constraints.

In this work, we propose to use public or private transportation means (e.g., buses, cars) as data caches and mobile relays, directly accessible by nearby UE. Vehicles create a *vehicular cloud* controlled by mobile network operators (MNOs). End users requesting to stream a content, can download chunks of it from nearby vehicles when available, filling up their playout buffer while watching, or, alternatively, stream the video directly from the cellular infrastructure when their buffer is empty and no vehicle is in reach. We focus on streaming content because of its major contribution to Internet traffic according to recent measurement studies [12]. Also, while the number of content in the Internet is large, content popularity is known to exhibit strong Zipfian characteristics, suggesting that reasonable hit ratios for the cloud can be achieved with moderate storage capacity per vehicle. Vehicles bring mobility for free, allowing UE to quickly browse the caches of many different encountered vehicles (while consuming the content chunks already available in their buffer, at a slower rate), thus virtually increasing the size of accessible caches.

We believe that the current infrastructure can be easily turned into a working vehicular cloud in a cost-efficient way: vehicles of all types are usually widespread in urban locations even in developing countries, and a subset of them could be readily equipped with storage capacity, wireless communication capabilities and basic computational power [2] at a much lower cost than SC deployments [25, 28]. Our solution brings a twofold benefit: on the one hand, it reduces the number of potential interruptions during playback, as future content chunks are *prefetched* from the cloud, when available, leading to improved QoE for the user; on the other hand, it promotes existing and new MNOs to operate in emerging markets enabling them to offer users a 3G/4G-like experience at a much lower cost.

To summarize, our main contributions in this paper are:

- We model the buffer dynamics as a queueing system, and analyse the characteristics of its idle periods (during which access to the cellular infrastructure is required).
- Based on this model, we formulate the problem of optimal allocation of content in vehicles, in order to minimize the

total load on the cellular infrastructure, assuming a fixed catalogue and a content popularity distribution.
- We provide closed-form expressions for the optimal allocation for two interesting regimes of vehicular traffic densities, assuming a relatively generic setting.
- We validate our theoretical results, using real traces for content popularity and vehicle mobility, and show that our system can offload up to 50% of streamed data in realistic scenarios, even with modest technology penetration.

The rest of the paper is organized as follows: first, in Section 2 we present the network model and the system assumptions; next, we provide closed-form expressions for optimal allocation of content in different traffic regimes in Section 3; then, we show the results of the real-trace based simulations in Section 4; finally, we list related work in Section 5, and we conclude our paper in Section 6.

2. SYSTEM MODEL

2.1 Architecture

In our architecture, we consider three types of nodes:

- **Infrastructure Nodes ($\mathbb{I}$)**, e.g., cellular base stations or macro-cells. These nodes are directly connected to the Internet and they can obtain any content. Their task is to fill the cloud caches with popular content, and serve requests for content not found in the cloud.
- **Helper Nodes ($\mathbb{H}$)**, e.g., cars, buses, taxis, where $|\mathbb{H}| = \alpha$. These nodes cache popular content, and have storage capacity limitations. Furthermore, we assume for simplicity that the $\mathbb{H}$ nodes cannot communicate with each other[1].
- **End Users ($\mathbb{U}$)**, e.g., smartphones, tablets, netbooks. A content requested by a user is streamed using a playout buffer for that content in her device. Content chunks inside the playout buffer are consumed at the viewing (playout) rate r_P. In parallel, whenever an $\mathbb{H}$ node with the watched content is encountered (referred to as a *contact*), subsequent chunks not already in the device buffer can be downloaded at low cost, for the duration of that contact. This download takes place at rate r_V. If the buffer is (almost) empty, the $\mathbb{U}$ node immediately switches to the cellular interface and downloads content at a rate r_C, until another vehicle storing the content is encountered. Finally, during a contact, simultaneous connections are not allowed, i.e., a $\mathbb{U}$ node can download from one and only one $\mathbb{H}$ node at a time.

Previous work has confirmed the feasibility of opportunistic connections between vehicles and UE [10]. IEEE 802.11p [4], which has been developed for the specific context of vehicular networks, is considered as the de facto standard. This standard actually covers simplicity (uncoordinated access mechanism, no authentication) and low delay (few hundreds ms in crowded areas). What is more, it is possible to implement a low battery consumption version in modern UEs without compromising performance [11]. Recently, there was an increasing interest in adopting LTE to support vehicular network applications [7]. Differently from IEEE 802.11p, LTE improves performance throughput, reliability and scalability at the cost of higher latency. Still, [20] shows that LTE is capable of satisfying delay requirements for most of the vehicular applications.

[1]While such communication could help fetch content over multiple hops, MANET-like schemes are known to considerably increase the complexity of the approach, while bringing incremental benefits. We defer their study to future work.

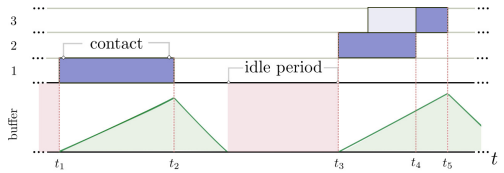

Figure 1: Sequence of contacts with three vehicles (above), and amount of data in end user buffer over time (below). When the buffer empties, the red region indicates that data are downloaded from the cellular infrastructure.

In Fig. 1 we sketch the basic communication protocol: during a contact, the user can fill her buffer while streaming the video (e.g., between t_1 and t_2 it will download from vehicle 1); when the buffer empties (red region), the user is redirected to the cellular infrastructure. Since we assume that a $\mathbb{U}$ node can download content from only a single $\mathbb{H}$ node at a time, the user will switch to vehicle 3 only at t_4, i.e., after she has finished downloading from vehicle 2.

2.2 Assumptions

We make the following system assumptions:

A.1 (Download rate) - We assume r_V and r_C to be larger than r_P in order to guarantee uninterrupted streaming[2]. We assume r_C is equal to $r_P + \epsilon$ (where $\epsilon > 0$ is small), in order to limit the access to the cellular infrastructure to the minimum required to ensure smooth playout (for simplicity, we will assume in our analysis that $\epsilon = 0$).

A.2 (Content popularity) - Let $\mathcal{K}$ be the set of content that can be requested by $\mathbb{U}$ nodes, such that $|\mathcal{K}| = k$. Each content $i \in \mathcal{K}$ is characterized by known size s_i and popularity ϕ_i, namely the expected number of requests during a given time window. Similar to a number of works on edge caching [15, 27], we assume this time window is a system parameter chosen by the MNO, during which estimates for the expected popularity per content are available. Every time period, the cellular operator updates its caches according to the new estimated popularity. Several studies have shown that it is possible to predict the popularity with good accuracy over relatively short periods (until few weeks for YouTube videos) [29]. Without loss of generality, we assume content is sorted by popularity as $\phi_1 \geq \phi_2 \geq \cdots \geq \phi_k$.

A.3 (Buffer capacity) - Each $\mathbb{H}$ node can store at most β bytes. While this assumption might seem to oversimplify the problem, we will show that our results can be applied to any buffer size. Moreover, we assume end user buffers to be large enough to store entirely any requested content.

A.4 (Inter-contact times) - Unless otherwise stated, pairwise inter-contact times between $\mathbb{H}$ and $\mathbb{U}$ nodes are drawn from a generic distribution with rate λ, and the contact duration is drawn from a generic distribution with mean $\mathbf{E}[S]$.

Given the above assumptions, the result of the policy is that (i) the user's video is never interrupted provided the infrastructure can guarantee at least the playout rate (if that

[2]Without this assumption, the streaming would require initial buffering in order to avoid interruptions, which are known to significantly degrade QoE [18]. While our framework could be extended to include initial buffering, due to space limitations we defer it to future work.

Table 1: Notation used in the paper.

α	Number of vehicles
β	Buffer size per vehicle
s_i	Size of content i
ϕ_i	Number of requests for content i
N_i	Number of copies stored for content i
k	Number of content in the catalogue
λ	Pairwise inter-meeting rate between $\mathbb{H}$ and $\mathbb{U}$
$\mathbf{E}[S]$	Mean contact duration
r_P	Playout rate
r_V	Download rate from $\mathbb{H}$ nodes

is not the case, then this is an issue of the infrastructure and not of the vehicular storage cloud); (ii) while the video plays out at the user, future parts of it are actually downloaded from locally encountered vehicles (in principle pre-fetched) thus offloading some of it from the infrastructure. As long as the playout buffer remains non-empty, the infrastructure does not need to be accessed. And when it does, we ensure that the minimum necessary amount of bytes is downloaded from the infrastructure ($r_C = r_P + \epsilon$).

The notation used in the paper is summarized in Table 1.

2.3 Problem formulation

Definition 1. *The number of bytes downloaded from $\mathbb{I}$ is given by $\sum_{i=1}^{k}\sum_{j=1}^{\phi_i} X_{ij}$, where X_{ij} is a random variable corresponding to the bytes of the j^{th} request for content i.*

Conditionally on the characteristics of the content (i.e., size and popularity) X_{ij} are IID random variables depending only on the mobility statistics (e.g., meeting rate with vehicles, contact duration). Thus, its expected value $\mathbf{E}[X_{ij}]$ is equal to $\mathbf{E}[X_{ij}|\phi_i, s_i] = \mathbf{E}[X_i],\ \forall j \in [1, \phi_i]$.

The goal of this paper is to find the optimal number of copies to allocate in $\mathbb{H}$ nodes in order to minimize the number of bytes downloaded from the cellular infrastructure. From Definition 1, we can derive the following theorem:

Theorem 2.1 (Optimization problem)**.** *The optimal number of copies for content i ($N_i \in \mathbb{N}$) is given by the solution of the following optimization problem:*

$$\underset{N_i}{\text{minimize}} \quad \sum_{i=1}^{k} \phi_i \cdot \mathbf{E}[X_i], \tag{1}$$

$$\text{subject to} \quad 0 \le N_i \le \alpha, \quad i \in \mathcal{K}, \tag{2}$$

$$\sum_{i=1}^{k} s_i N_i \le \alpha\beta, \tag{3}$$

where the objective function is the total number of bytes downloaded from the cellular infrastructure in a given period of time. The objective function is subject to two constraints: (i) the number of replicas is limited by the number of vehicles participating in the cloud and cannot be negative (Eq. (2)); (ii) the buffer capacity is limited as shown in Assumption A.3 (Eq. (3)). Due to space limitations, all proofs can be found in a tech report, available at [30], and we will provide instead only a proof sketch for a subset of results.

3. ANALYTICAL MODEL

In this section, we derive $\mathbf{E}[X_i]$ according to the vehicle density, i.e., the rate of vehicles met by $\mathbb{U}$ nodes, and we solve the respective optimization problem of Theorem 2.1 to find the optimal content allocation. We model the playout buffer at $\mathbb{U}$ nodes with a queue, where its number of "jobs" corresponds to the amount of bytes available in the playout buffer, i.e., the number of bytes prefetched from the cloud, but not yet consumed; thus, when the queue empties (*idle period*), the user will switch to the cellular interface.

3.1 Low Traffic (LT)

Let us assume first that contacts with vehicles are sparse, and do not overlap with each other, i.e., a $\mathbb{U}$ node meets one and only one vehicle with the requested content at a time. Assume that $\mathbb{U}$ node is streaming content i. We can model the playout buffer for content i as a bulk $\mathrm{G}^Y/\mathrm{D}/1$ queue (as shown in Fig. 2). A contact between $\mathbb{U}$ and a vehicle storing content i corresponds to a new (bulk) arrival in the buffer. Since there are N_i such vehicles, the total arrival rate into the queue is λN_i (while the actual inter-arrival distribution is generic, G). Each arrival brings a random amount of Y new bytes to be consumed (that depends on the random contact duration with the vehicle). Finally, bytes in the buffer are served (i.e., viewed by the user) at the constant playout rate r_P. We are interested in the idle periods of the queue, i.e., when the buffer is empty and the user is redirected to the cellular infrastructure. We can enunciate the following theorem:

Theorem 3.1 (Low Traffic)**.** *If $\lambda\mathbf{E}[S]\cdot\alpha \ll 1$ (Low Traffic), the optimal content allocation replicates the most popular content in any vehicle, until the buffer space is filled up:*

$$N_i^* = \begin{cases} \alpha, & \text{if } 1 \le i \le \gamma, \\ 0, & \text{otherwise,} \end{cases}$$

where $\gamma \triangleq \max\{\bar{\gamma} \in \mathbb{N} | \sum_{i=1}^{\bar{\gamma}} s_i \le \beta\}$.

Sketch of proof. Since the video length is much larger than the mean contact duration, namely $s_i/r_P \gg \mathbf{E}[S]\ \forall i \in \mathcal{K}$, the playout queue sees several idle/busy cycles. For this reason, we can apply stationary regime analysis to the proposed queue. Specifically, the fraction of time that the $\mathrm{G}^Y/\mathrm{D}/1$ queue is idle is $1 - \lambda\mathbf{E}[S]\cdot r_V/r_P \cdot N_i$. Then, the objective function of the optimization problem is equivalent to:

$$\min_{N_i} \sum_{i=1}^{k} \phi_i \cdot s_i(1 - \lambda\mathbf{E}[S]\cdot r_V/r_P \cdot N_i) \equiv \max_{N_i} \sum_{i=1}^{k} \phi_i \cdot s_i N_i. \tag{4}$$

We distinguish two cases according to the value of β: *large* β, when all content fits in a vehicle buffer (i.e., $\beta \ge \sum_{i=1}^{k} s_i$); and *small* β, otherwise:

- *Large* β: since the RHS of Eq. (4) is a strictly monotonic increasing function according to N_i, then $N_i^* = \alpha\ \ \forall i \in \mathcal{K}$, due to the constraint in Eq. (2).
- *Small* β: let $y_i \triangleq s_i N_i$ be the total number of bytes to allocate per content. Here, we solve a continuous relaxation for N_i. We rewrite the optimization problem as:

$$\max_{y_i} \sum_{i=1}^{k} \phi_i y_i \quad \text{s.t.} \sum_{i=1}^{k} y_i \le \alpha\beta. \tag{5}$$

  Additionally, from Eq. (2) we know that y_i ranges between 0 and $s_i\alpha$. The optimization problem can be read as if we had a budget of bytes (until $\alpha\beta$) to use in order to maximize the objective function. We can proceed iteratively in order to find the optimal value of y_i. Let $\epsilon > 0$ be a small number of bytes ($\epsilon \ll \alpha\beta$) to assign to one of the content. Initially, let further assume that $y_i = 0\ \ \forall i \in \mathcal{K}$. An optimal solution assigns the ϵ bytes to

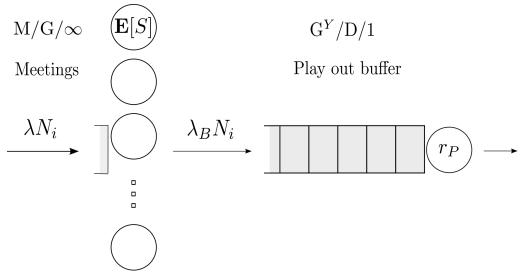

Figure 2: Queuing model proposed for generic traffic regime.

content 1 ($y_1 = \epsilon$), since ϕ_1 is maximum. We iterate this step by summing the marginal budget ϵ, until one of the two following sub cases arises:

(i) $y_1 = s_i\alpha$: due to Eq. (2), no more bytes can be assigned to y_1. In this case, the additional ϵ bytes will be assigned to y_2, and so on;

(ii) $\sum_{i=1}^{k} y_i = \alpha\beta$: there is not available space in the cloud. Then, N_i^* is equal to y_i/s_i.

As stated in Theorem 3.1, the optimal solution allocates copies to the content with the largest popularity. □

Corollary 1. *An optimal policy stores α copies for the (approximately) $\beta/\mathbf{E}[s_i]$ most popular content.*

Corollary 2. *Let β_j denote the buffer capacity for vehicle j. Then, for Low Traffic, the optimal content allocation replicates the most popular content, until the buffer space is filled up. Specifically, $N_i^* = \sum_{j=1}^{\alpha} \mathbb{1}_A(\beta_j)$, where $A = \{n \in \mathbb{N} | n \geq i\}$. If $\beta_j = \beta\ \forall j \in \mathbb{H}$, then the optimal allocation corresponds to Theorem 3.1.*

3.2 Generic Traffic (GT)

In busy urban environments, a user might be within range of many vehicles at the same time, a number of which storing the content of interest. If a user is downloading video i from car A, and the connection is lost (e.g., car A moves away), the user could just keep downloading from another car B storing i, also in range. Hence, as long as there is *at least* one vehicle with a copy within range for a duration T, then the user will be downloading content i at rate r_V for all of T. We are now thus interested in modelling the duration of periods with at least one vehicle with a copy in range.

Unfortunately, we cannot use the previous model here, as it assumes the absence of overlaps. In fact, an encountered car does not necessarily bring an amount of content $\mathbf{E}[S] \cdot r_V$ now. Consider the example above with car A and car B. When the user switches to downloading from car B, the contact with car B was already ongoing and thus does not have the same statistics as in Low Traffic regime. We choose to model these contact periods (with potential overlaps) with the busy periods of a queue with infinite servers. However, to derive the mean busy period, we have to assume that pairwise inter-contact times between users and vehicles are exponentially distributed with mean λ. While this is still subject of debate, several studies have shown that this exponential assumption is reasonable, especially in the tail of the distribution [13, 19]. Thus, we model the queue as an M/G/∞ (Fig. 2), whose output is feeding arrivals into the playout buffer as described in Section 3.1. The number of vehicles simultaneously inside the end user communication range, as a function of time, corresponds to the number of jobs in the M/G/∞ queue. Specifically, arrivals describe the vehicles met by a $\mathbb{U}$ node, and the service time is the contact duration. We will test the model in Section 4 against a trace with non-exponential contacts.

Lemma 3.2 (Busy period)**.** *Consider an M/G/∞ queue, with arrival rate λN_i and mean service time $\mathbf{E}[S]$. The mean busy period length is given by:*

$$\mathbf{E}[B] = \frac{e^{\lambda \mathbf{E}[S] \cdot N_i} - 1}{\lambda N_i}. \tag{6}$$

Proof. Consider the M/G/∞ queue previously described. Then, let B_n (resp. I_n) be the length of the n^{th} busy (resp. idle) period. Note that (B_n, I_n) forms an Alternating Renewal process. We can associate to each renewal a reward which corresponds to the duration of busy periods (i.e., the time during which there is at least one job in the queue). From the Renewal-Reward theorem [17], in the stationary regime, the average rate at which we earn reward is equal to the expected reward earned during a cycle, divided by the expected cycle length. It follows that, if $p(t)$ is the probability that the server is busy at time t, then:

$$p(t) = \frac{\mathbf{E}[B]}{\mathbf{E}[B] + \mathbf{E}[I]}. \tag{7}$$

In such M/G/∞ queue, the fraction of time that the system is busy (i.e., $p(t)$) is equal to $1 - e^{-\lambda N_i \cdot \mathbf{E}[S]}$; this result is directly derived from the stationary probabilities of an M/M/∞, a known insensitivity result [24]. Moreover, by the Markov property of arrivals, $\mathbf{E}[I] = 1/\lambda N_i$. Then, from Eq. (7), we deduce that $\mathbf{E}[B] = (e^{\lambda N_i \ \cdot \ \mathbf{E}[S]} - 1)/\lambda N_i$. □

Corollary 3. *Consider the M/G/∞ queue of Lemma 3.2. Then, its busy periods are generic distributed with mean $\lambda_B \ \cdot \ N_i$, where $\lambda_B \triangleq \lambda \ \cdot \ e^{-\lambda \mathbf{E}[S] \cdot N_i}$.*

As in the Low Traffic regime, we model the playout buffer with an G^Y/D/1 queue. Arrivals represent the beginning of a busy period of duration given by the M/G/∞ queue. We use Corollary 3 to define the arrival rate, and $\mathbf{E}[B] \cdot r_V$ as mean bulk size. We introduce the following result:

Theorem 3.3 (Generic Traffic)**.** *In the Generic Traffic regime, the optimal number of replicas splits content in three slots depending on the content popularity. Specifically:*

$$N_i^* = \begin{cases} 0, & \text{if } \phi_i < L \\ (\lambda \mathbf{E}[S])^{-1} \cdot \ln\left(\frac{\lambda \mathbf{E}[S] \cdot \phi_i}{m_C}\right), & \text{if } L \leq \phi_i \leq U \\ \alpha, & \text{if } \phi_i > U \end{cases}$$

where $L \triangleq m_C \cdot (\lambda \mathbf{E}[S])^{-1}$ and $U \triangleq m_C \cdot (\lambda \mathbf{E}[S])^{-1} \cdot e^{\frac{\alpha}{\lambda \mathbf{E}[S]}}$, and m_C is an appropriate Lagrangian multiplier.

Sketch of proof. According to Lemma 3.2 and Corollary 3, we can rewrite the optimization problem as follows:

$$\min_{N_i} \sum_{i=1}^{k} \phi_i \cdot s_i e^{-\lambda \mathbf{E}[S] \cdot N_i}. \tag{8}$$

The problem is convex since the objective function is sum of convex functions, the constraints are linear and the domain of feasible solutions is convex. Thus, it can be solved with the method of Lagrangian multipliers [6]. □

Table 2: Mobility statistics (λ and λ_B are in day^{-1}).

Range	λ	$\mathbf{E}[S]$	λ_B	$\mathbf{E}[B]$
Short range	1,680	29,53 s	1,238	34,53 s
Long range	4,391	57,77 s	0,924	139,11 s

4. SIMULATIONS

To validate our results, we perform simulations based on real traces for (i) mobility and (ii) content popularity:

- **Mobility**: We use the Cabspotting trace [26] to simulate the vehicle behaviour; this trace records the GPS coordinates for 531 taxis in San Francisco for more than 3 weeks with granularity of 1 minute. In order to improve the accuracy of our simulations, we increase the granularity to 5 seconds by linear interpolation.
- **Content popularity**: We infer the number of requests per day from a database with statistics for 100.000 YouTube videos [3]. The database includes static (e.g., title, description, author, duration) and dynamic information (e.g., daily and cumulative views, shares, comments). Data is worldwide, and we scale it linearly according to the estimated population of the centre of San Francisco.

In our simulations, we assume that end users can contact the vehicular cloud with either *short* (100m) or *long* (200m) range communications. We vary video qualities from 240p (which requires r_P = 400kbps) to 1080p (r_P = 4Mbps). What is more, current and near future vehicles can be equipped with 801.11 b/g/n mobile access points for wireless devices [1]. While these devices can provide much faster download rates, we set r_V to 5Mbps as a pessimistic scenario. Finally, we extrapolate the mobility statistics (λ and $\mathbf{E}[S]$) from the analysis of the Cabspotting trace to compute the optimal allocation. These values[3] are summarized in Table 2.

We build a MATLAB simulator as follows: first, we generate a set of requests, and we associate a random location (in GPS coordinates) to each one. The number of requests per content per day is given by the YouTube trace. Then, we store content in vehicles according to the policies shown in Section 3.1 and 3.2. For each request, we simulate the playout of the video; the end user buffer will be opportunistically filled when the vehicular cloud can be contacted, according to the mobility provided by the Cabspotting trace. In order to increase the number of simulations and to provide sensitivity analysis for content size and buffer capacity, we limit the number of content to 5000. We scale down the vehicle storage capacity β to ensure that 0,05%-1% of the total catalogue can be fit in each cache, which is an assumption that has also been used in [15, 22, 27]. Finally, content requests are generated over a period of 5 days.

Fig. 3 shows the fraction of bytes downloaded from the vehicular cloud according to different video qualities with Low Traffic regime. The plot reveals that long range communications can reduce the mobile traffic from 30% (full HD content) to 45% (low resolution content). What is more, short range communications can also provide gains between 15% and 35%. While in this simulation we are only storing the 0,2% of the total catalogue per vehicle, considerable gains can still be achieved. Not less important is that the number of vehicles participating in the cloud is small ($\alpha = 531$),

[3]The values of λ_B and $\mathbf{E}[B]$ consider $N_i = \alpha$.

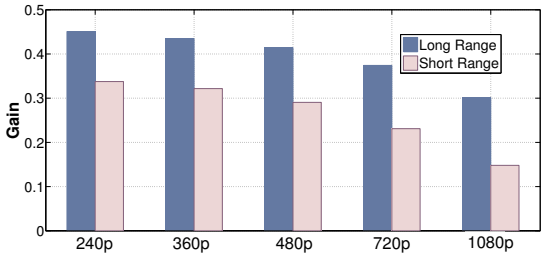

Figure 3: Data offloaded for different video quality ($\beta = 0,2\%$) for LT regime.

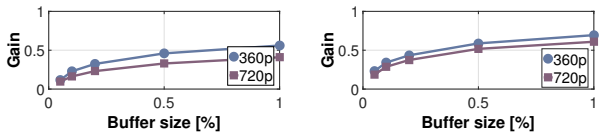

Figure 4: Data offloading gain with short (left) and long (right) range communications for GT regime.

approximately less than the 1% of estimated number of vehicles in the center of San Francisco. This provides some evidence on the advantages of offloading based on a vehicular cloud, compared to offloading using small cells or WiFi with deadlines, as for example in [8] or [21].

Fig. 4 depicts the fraction of data offloaded by the vehicular cloud for two video qualities (360p and 720p) according to different values of β for GT regime. On the one hand, small vehicle caches ($\leq 0,1\%$ of the catalogue) can provide interesting gains even for high quality videos (15-30%); on the other hand, using larger caches allows significant higher gains, up to 60% with long range communications.

Stationary regime is reasonable when the number of busy + idle periods is large enough such that the transitory phase becomes negligible. To have a high number of these periods, the average content size and the number of requests need to be large. Fig. 5 shows the fraction of data offloaded by our vehicular cloud for content of same size (with video quality 360p). We can note that when file size is larger than 100MB ($\approx$ 15 minutes), the number of bytes downloaded from the infrastructure becomes more stable, validating the stationary regime assumption. Furthermore, with short range communications, LT and GT policies provide almost the same gain: in fact, from Table 2 we can note that the value of λ_B (resp. $\mathbf{E}[B]$) is similar to λ (resp. $\mathbf{E}[S]$), revealing a limited number of contact overlaps. Differently, for long range communications the result in the GT regime can increase the offloaded traffic up to 5% more than LT regime, confirming the validity of the exponential assumption for the mobility.

5. RELATED WORK

The rapid increase in the mobile traffic demand has led to a large number of proposals to alleviate the load on the backbone. Related work can be split in caching at the edge and offloading on mobile devices. In the context of caching at the edge, traditional solutions concern adding storage capacity to small cells [15, 27] and/or to WiFi Access Points [14, 31]. However, a large number of small cells is required for an extensive coverage, which comes at a high cost [5]. By con-

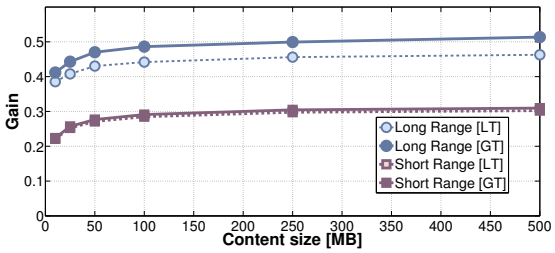

Figure 5: Data offloaded vs content size (video quality 360p).

trast, in an urban environment, the same area will contain thousands of vehicles. To alleviate this situation at a low cost, a number of works introduce delayed access [8, 21, 23]. Differently, in our work we guarantee QoE (since we provide uninterrupted streaming) with similar or better gains in terms of data offloaded than the aforementioned works.

Apart from small cells, researchers also proposed to use mobile devices to offload content through opportunistic communications [9, 16]. These works exploit the possibility of serving user requests from other nearby mobile devices. Moreover, a few works have also suggested to exploit vehicular networks to store content [32, 33]. Despite the obvious limited constraints in terms of resources (e.g., battery, storage), the majority of these works does not consider a common cloud maintained by the ISP for video streaming.

6. CONCLUSION

In this work, we have provided a low-cost alternative to decrease the load on the cellular network. Specifically, we have proposed to cache popular multimedia content in vehicles. Users can opportunistically fill up their buffer from nearby vehicles, in a transparent way, while streaming the requested content. We have modelled the user playout buffer with a queueing approach, and provided closed-form expressions for different vehicle densities. Finally, we have validated our results with real-trace based simulations. As future work, we will evaluate how the mobility patterns affect the system and if they can be used to improve performance. We also plan to study the transient phase of our queuing system.

7. REFERENCES

[1] Sierra wireless. http://www.sierrawireless.com.

[2] Your Car Is About To Get Smarter Than You. http://business.time.com/2014/01/07/your-car-is-about-to-get-smarter-than-you-are/.

[3] YouStatAnalyzer database. http://www.congas-project.eu/youstatanalyzer-database.

[4] IEEE Draft Standard for Amendment to Standard for Information Technology - Amendment 6: Wireless Access in Vehicular Environments. *IEEE Std P802.11p/D11.0*, 2010.

[5] N. Alliance. NGMN 5G White Paper. https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf, 2015.

[6] A. Antoniou and W.-S. Lu. *Practical optimization - algorithms and engineering applications*. Springer, 2007.

[7] G. Araniti et al. Lte for vehicular networking: a survey. *IEEE Communications Magazine*, 2013.

[8] A. Balasubramanian et al. Augmenting mobile 3g using wifi. In *ACM MobiSys*, 2010.

[9] X. Bao et al. Dataspotting: Exploiting naturally clustered mobile devices to offload cellular traffic. In *IEEE INFOCOM*, 2013.

[10] V. Bychkovsky et al. A measurement study of vehicular internet access using in situ wi-fi networks. In *ACM MOBICOM*, 2006.

[11] P. Choi et al. A case for leveraging 802.11p for direct phone-to-phone communications. In *IEEE ISLPED*, 2014.

[12] Cisco. Cisco visual networking index: Global mobile data traffic forecast update. 2014-2019.

[13] V. Conan et al. Characterizing pairwise inter-contact patterns in delay tolerant networks. Autonomics, 2007.

[14] S. K. Dandapat et al. Sprinkler: Distributed content storage for just-in-time streaming. In *Workshop on CellNet*, 2013.

[15] N. Golrezaei et al. Wireless device-to-device communications with distributed caching. *CoRR*, abs/1205.7044, 2012.

[16] B. Han et al. Mobile data offloading through opportunistic communications and social participation. *IEEE Trans. on Mobile Computing*, 2012.

[17] M. Harchol-Balter. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press, 2013.

[18] T. Hossfeld et al. Initial delay vs. interruptions: Between the devil and the deep blue sea. In *W. on QoMEX*, 2012.

[19] T. Karagiannis et al. Power law and exponential decay of intercontact times between mobile devices. *IEEE Trans. on Mobile Computing*, 2010.

[20] H.-Y. Kim et al. A Performance Evaluation of Cellular Network Suitability for VANET. *Journal of Electrical, Computer and Communication Engineering*, 2012.

[21] J. G. Lee et al. An approach to model and predict the popularity of online contents with explanatory factors. In *Web Intelligence*. IEEE, 2010.

[22] M. A. Maddah-Ali and U. Niesen. Fundamental limits of caching. *IEEE Trans. on Information Theory*, 2014.

[23] F. Mehmeti and T. Spyropoulos. Is it worth to be patient? analysis and optimization of delayed mobile data offloading. In *IEEE INFOCOM*, 2014.

[24] G. F. Newell. The M/G/∞ Queue. *SIAM Journal on Applied Mathematics*, 1966.

[25] V. Nikolikj and T. Janevski. A Cost Modeling of High-capacity LTE-advanced and IEEE 802.11ac based Heterogeneous Networks, Deployed in the 700 MHz, 2.6 GHz and 5 GHz Bands. *Procedia Computer Science*, 2014.

[26] M. Piorkowski et al. DAD data set epfl/mobility (v. 2009-02-24). http://crawdad.org/epfl/mobility/, 2009.

[27] K. Poularakis et al. Video delivery over heterogeneous cellular networks: Optimizing cost and performance. In *IEEE INFOCOM*, 2014.

[28] Senza Fili Consulting. The economics of small cells and wi-fi offload. 2013.

[29] G. Szabo and B. Huberman. Predicting the popularity of online content. *Comm. of the ACM*, 2010.

[30] L. Vigneri et al. List of proofs for "Streaming Content from a Vehicular Cloud". https://goo.gl/qFl9a1.

[31] F. Zhang et al. EdgeBuffer: Caching and prefetching content at the edge in the MobilityFirst future internet architecture. In *IEEE WoWMoM*, 2015.

[32] Y. Zhang et al. Roadcast: A popularity aware content sharing scheme in VANETs. In *IEEE ICDCS*, 2009.

[33] J. Zhao and G. Cao. VADD: Vehicle-Assisted Data Delivery in Vehicular Ad Hoc Networks. In *IEEE INFOCOM*, 2006.