Performance Evaluation of Video Transcoding and Caching Solutions in Mobile Networks

Salah Eddine Elayoubi*†, James Roberts†
*Orange Labs, Issy-Les-Moulineaux, France
†IRT SystemX, Paris, France

***Abstract*—The rapid growth in demand for video streaming applications is stressing the performance of wireless access networks. To alleviate congestion, vendors currently propose devices to be placed in the operator's network that transcode videos to a lower rate in order to reduce traffic volume in case of congestion. The devices are also able to cache popular videos, both to reduce the burden of transcoding and to alleviate backhaul load. The paper proposes a model of this augmented radio access network enabling an evaluation of the performance benefits for given transcoding and caching capacities. Our results show that a gain in cell capacity of 15% can be realized with moderate transcoding and cache capacities.**

I. Introduction

Video downloads represent a large and increasing proportion of the traffic handled in mobile access networks. Cisco forecasts a 69% annual growth rate for video demand with its share of traffic increasing from 53% in 2013 to more than 70% in 2018 [4]. This growth is stressing the performance of the wireless segment and, as any user knows, video streaming frequently suffers from stalling events that occur whenever the download rate is persistently less than the intrinsic video rate causing the playout buffer to empty.

If congestion cannot be avoided, network operators may consider it preferable to avoid stalling and other manifestations of demand overload by reducing the quality of downloaded content. Devices to perform "dynamic content optimization" are currently marketed by equipment vendors [13] [10]. Moreover, this function is highlighted as a use case to be considered by the recently launched Mobile-edge computing industry initiative [6]. Reducing image quality means reducing the traffic volume, thus alleviating congestion. From another point of view, since traffic is reduced, the required network capacity to meet user quality requirements is smaller leading to lower expenditure. The aim of the present paper is to investigate the tradeoff realized between the cost of the content optimization devices and the resulting savings in network infrastructure.

In view of the high proportion of video traffic, we restrict attention in this paper to an evaluation of the benefits of optimizing just video content. The considered devices reduce the volume of video demand, as necessary, by transcoding downloaded videos to a lower rate. The devices can also cache frequently requested items to limit the transcoding load. We consequently coin the generic name for such a device as, "video transcoder and cache" or VTC. Figure 1 shows the position of the VTC in the mobile access network. It is situated at some convenient point to act on the traffic of a number of base stations. Determining the precise location, implying the degree of traffic concentration and the device workload, is a design issue.

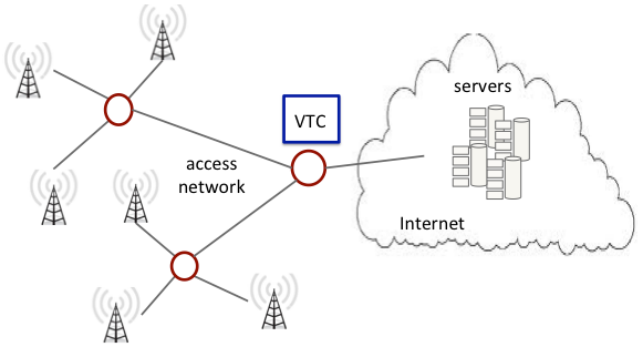

Fig. 1. The video transcoder and cache (VTC) concentrates the traffic of a number of base stations.

Other questions we address are, what capacity gain is possible through transcoding? and how should the device be dimensioned? i.e., how many videos can be transcoded in parallel and how many videos should the VTC be able to hold in its cache. It is impossible to provide definitive answers to these questions since the relative benefits of alternative solutions depend on many local parameters. Rather, we provide analytical tools to be used to evaluate relevant scenarios. These tools take account of the impact of different radio conditions, the dynamics of flow arrivals of different types, the traffic mix between these types, the distribution of video popularity and its impact on cache performance. We illustrate the use of the tools and how they can guide design and network planning decisions though a number of numerical applications.

Only certain downloads are susceptible to transcoding and the relative proportion of traffic in this category clearly significantly impacts the effectiveness of the VTC device. The current trend is towards greater proportions of content encryption reducing the potential advantage of transcoding and caching since these operations become ineffective. Nevertheless, it remains of interest to evaluate these advantages since, if they are significant, it might in fact modify the observed trends. Moreover, the emergence of so-called information-centric networks may solve the security and commercial issues that encryption is designed to resolve.

In the next section we more completely describe the consid-

ered VTC equipped mobile access network. We then proceed in Section III to build a performance model accounting for the random process of download arrivals and completions and the impact of diverse radio conditions within a cell, under the assumption that the VTC is "perfect" in always being able to supply a transcoded version of a requested video when necessary. In this section, we also propose performance criteria to be used when determining cell traffic capacity. Section IV further develops the performance model to account for limited VTC transcoding and caching capacity. The model is used in Section V to derive some numerical results illustrating possible system parameter choices and their impact on performance and, ultimately, on network traffic capacity.

II. System Description

We first broadly characterize downloads into three types depending on their adaptivity. We then describe the considered VTC device before proposing a simple model of the radio link and how its spectrum is shared by concurrent downloads.

A. Traffic types

Cell downlink capacity is shared by flows of three types. Type 1 flows are video downloads that can be identified by the considered VTC device through deep packet inspection. These video downloads are subject to transcoding and are the principal object of this evaluation. The other two types constitute background traffic.

Type 2 flows consist of regular downloads of documents of various kinds like web pages, emails, pictures as well as videos that are not compressible (because they are encrypted, for example). Type 3 flows are adaptive video streaming flows. Their download rate is adjusted by the application to approximately match the currently available bandwidth. This is achieved by the application requesting the video in segments of 2 to 4 seconds, say, where each segment exists in several versions corresponding to different rates. The application selects the segment version corresponding to its evaluation of the current path available bandwidth (see [9], for example).

B. Video transcoding and caching

The VTC can transcode and cache type 1 videos. These videos are assumed to be downloaded as fast as possible for subsequent viewing. "As fast as possible" means the download rate is determined as a fair share of cell capacity accounting for the number of competing flows and their particular radio conditions, as discussed in Section II-C below. With respect to transcoding, we envisage online and offline scenarios. In the online scenario, the original version is transcoded on-the-fly and a compressed version is sent to the user. The VTC can only perform online transcoding for a limited number of videos in parallel. In the offline scenario, a compressed version is created when absent from the cache, either systematically whenever the video is requested, or just when the video is requested and the user's cell has attained a certain degree of congestion. The user triggering transcoding always receives a copy of the original version.

The VTC can cache a limited number of videos. Several options are possible.

- it caches the version requested by the operator, original or compressed,
- it caches both versions systematically when either is requested and the video is not already cached,
- it only ever caches the compressed version, always referring to an upstream cache or server for the original version.

The adopted caching policy impacts the effectiveness of the VTC in increasing cell capacity.

The network operator determines which version of the video to download based on current cell traffic conditions. In periods of light traffic, or when the radio conditions of a requester are particularly favorable, the operator will forward the user's request for the original version. This might be found in the VTC cache or, if not, in some upstream cache or video server. Otherwise, the operator will seek to download a compressed version obtained from the cache or by online transcoding. The type of coding is determined by the operator at the start of transmission and does not change, even though the cell congestion status might evolve.

C. Download rates

We consider an isolated cell where the radio condition of a given user is assumed to remain constant for the duration of a download. Specifically, we assume a discrete set of M radio conditions (determined by the user's position relative to the base station, for instance) characterized by peak rates R_i bit/s for $i = 1, \ldots, M$. R_i is the rate a mobile terminal with conditions i would get if it were alone in the cell. When the overall number of users in the cell is n, users in radio class i get rate R_i/n. This is approximately the rate provided by the MAC layer of 3G and 4G cellular networks where cell capacity is shared by assigning roughly the same rate of equal length time slots to all cell users [2], [14].

The operator is assumed to determine which version of a downloaded video of type 1 to request from two parameters:

- the radio condition, as determined by the peak rate R_i, of the user in question,
- the number n of users currently active in the cell of any type.

The rate available to a new user is then $R_i/(n+1)$ and the requested coding rate should be chosen in consequence.

Type 2 downloads naturally proceed at the current fair rate for their class, R_i/n. The adaptive video streaming applications managing type 3 flows have several versions of each segment available, each corresponding to a different rate. The application in the user terminal chooses the version to request for each segment depending on its current vision of the network congestion state. Available rates vary from application to application and the algorithm used to choose the rate of each segment is typically quite complex (e.g., [9]). As our focus is on the performance of the VTC, we adopt a simplified model of type 3 rate adjustments in the following evaluation (see Sec. III-A).

III. Cell performance

We propose a Markov model to characterize the population of simultaneous downloads of all types. The model is first defined here under the simplifying assumption that a compressed coding is always available for download whenever requested. We then discuss performance criteria to be used to determine the maximum per-cell load attainable. We consider the additional impact of limited VTC capacity in Section IV.

A. Markov model

We adopt a simple Markovian traffic model for the sake of tractability. We note, however, that the derived results are typically more generally valid thanks to the insensitivity properties of the underlying processor sharing system [2]. Flows of type j with radio condition i are assumed to arrive as a Poisson process of rate λ_{ij} for $i = 1, ..., M$ and $j = 1, 2, 3$.

The size of a type 1 flow depends on the video coding rate. We assume the video when played back has an exponentially distributed duration of mean τ_1 seconds and that the original and compressed coding rates are C_o and C_c bit/s, respectively. The amount of data to download is thus either $C_o\tau_1$ or $C_c\tau_1$ bits. We consider here that the compressed version is always available, either through online transcoding or thanks to a very large cache yielding negligible miss rates. This enables us to quantify the maximum gain in capacity attainable by the use of the VTC. We assume compressed coding is requested for a new download under condition i whenever the population of users n is greater than R_i/C_o.

Each flow of type 2 requires the download of an exponentially distributed volume of mean σ_2 bits and a download with condition i proceeds at rate R_i/n. To model type 3 adaptive streaming flows, we suppose their rate is adjusted continuously, exactly like type 2 flows. However, the duration of the flow is now independent of the download rate. We assume the distribution of this duration is exponential of mean τ_3.

System state is defined by the number of flows in progress of each radio condition and each type, distinguishing the video coding rate of type 1 flows. Let a_{io} and a_{ic} be the number of type 1 flows with radio condition i downloading the original and the compressed version, respectively. Let b_i and c_i be the number of type 2 and 3 flows with radio condition i, respectively. Finally, denote by n the overall number of flows in progress, $n = \sum_{i=1}^{M}(a_{io} + a_{ic} + b_i + c_i)$.

Vector $\vec{n} = (a_{1o}, \ldots, a_{Mo}, a_{1c}, \ldots, a_{Mc}, b_1, \ldots, b_M, c_1, \ldots, c_M)$ is a Markov process with the following non-zero transition rates:

- a_{io} increases to $a_{io} + 1$ at rate λ_{i2} if $\frac{R_i}{n+1} \geq C_o$,
- a_{io} decreases to $a_{io} - 1$ at rate $a_{io}\frac{R_i}{nC_o\tau_1}$,
- a_{ic} increases to $a_{ic} + 1$ at rate λ_{i2} if $\frac{R_i}{n+1} < C_o$,
- a_{ic} decreases to $a_{ic} - 1$ at rate $a_{ic}\frac{R_i}{nC_c\tau_1}$,
- b_i increases to $b_i + 1$ at rate λ_{i1},
- b_i decreases to $b_i - 1$ at rate $b_i\frac{R_i}{n\sigma_2}$,
- c_i increases to $c_i + 1$ at rate λ_{i3},
- c_i decreases to $c_i - 1$ at rate $\frac{c_i}{n\tau_3}$.

The above transition rates for valid states $\vec{n}$ define a transition matrix Q and the steady state probabilities $\pi(\vec{n})$ can be obtained by solving $\pi Q = 0$. This system can be readily solved numerically as long as the state space is not too large.

B. Performance criteria

We consider the following three performance metrics.

a) Compression probability: This is the probability an arriving type 1 user should be sent the compressed version. It depends on the new user's radio condition and the total number of active users in the cell. The compression probability p_c^i for a flow starting with radio condition i is thus

$$p_c^i = 1 - \sum_{\vec{n} \in S^i} \pi(\vec{n}), \tag{1}$$

where the summation domain $\vec{n} \in S^i$ covers states such that $\sum_k (a_{ko} + a_{kc} + b_k + c_k + 1) < \frac{R_i}{C_o}$.

b) Rate deficit probability: A rate deficit occurs if the instantaneous download rate of a type 1 flow is less than the nominal coding rate, C_o or C_c for original and compressed versions, respectively. It is a measure of video quality in that a flow suffering persistent rate deficit will likely experience playout stalling. This metric is chosen for the sake of tractability since it is hardly possible to compute the stalling probability for this system [16].

The rate deficit probability is equal to the probability an ongoing flow shares cell capacity with a number of users greater than the threshold corresponding to its radio condition and coding version. When the system is in state $\vec{n}$, there are a_{ij} video users of radio class i downloading version j (original or compressed). These videos suffer rate deficit if the overall number of users n is greater than the ratio $\frac{R_i}{C_j}$. The proportion of class i flows in rate deficit in state $\vec{n}$ is thus equal to $\frac{a_{io}}{a_{io}+a_{ic}} \mathbb{1}_{\{n > \frac{R_i}{C_o}\}} + \frac{a_{ic}}{a_{io}+a_{ic}} \mathbb{1}_{\{n > \frac{R_i}{C_c}\}}$. The rate deficit probability for class i users, p_d^i, is thus

$$p_d^i = \sum_{\vec{n}} \left(\frac{a_{io}}{a_{io} + a_{ic}} \mathbb{1}_{\{n > \frac{R_i}{C_o}\}} + \frac{a_{ic}}{a_{io} + a_{ic}} \mathbb{1}_{\{n > \frac{R_i}{C_c}\}}\right) \pi(\vec{n}) \tag{2}$$

and the overall rate deficit probability is

$$p_d = \sum_{i=1}^{M} \sum_{\vec{n}} p_d^i. \tag{3}$$

c) Cell utilization: This metric is the proportion of time the cell is actively downloading data. It is equal to $1 - \pi(\vec{0})$. A classical pragmatic criterion used to dimension cell networks is to maintain utilization below some threshold, $1 - \pi(\vec{0}) \leq .8$, say.

IV. Impact of VTC capacity limitations

We now consider the end-to-end performance realized by the VTC device accounting for the impact on the model of Section III of its finite transcoding and caching capacities.

A. Cache hit rates

The VTC device is supposed to be situated at a backhaul node as depicted in Figure 1 and processes requests from a relatively large number of cells. The cache hit rate is evaluated under the following assumptions:

- the type 1 catalogue is of size N videos,
- the VTC contains a cache of size $N\delta$ bytes and applies the least recently used (LRU) replacement policy,
- the average size of a video is σ_o bytes in its original coding and σ_c bytes in its compressed version ($\sigma_c/\sigma_o = C_c/C_o$),
- we assume $N\delta \gg \sigma_o$ so that we can ignore edge effects and assume the cache is saturated if and only if $n_o\sigma_o + n_c\sigma_c \geq N\delta$ where n_o and n_c represent the number of videos stored in the original and compressed versions, respectively.
- we apply the independent reference model (IRM), equivalently assuming requests for both original and compressed versions arrive as a Poisson process (see [7], for example, for a discussion on the validity of this approximate model),
- video popularity is assumed to follow a Zipf law of parameter α, i.e., the arrival rate for requests for the r^{th} most popular video is proportional to $1/r^\alpha$,
- the probability p_c^i a compressed version is requested is given by (1), independently for all videos.

To evaluate hit rates we use the Che approximation [3] and, more explicitly, its Gaussian approximation for Zipf popularities derived by Fricker *et al.* [7]. We in fact extend the approximation somewhat to account for the different sizes of the original and compressed versions.

Let p_o and p_c denote the probability original and compressed versions are inserted in the cache on a request for the video (of either type). If only the requested version is cached, $p_c = \sum_i \lambda_i p_c^i / \sum_i \lambda_i$ and $p_o = 1 - p_c$ where p_c^i is given by (1). If both versions are systematically cached, $p_o = p_c = 1$. If only the compressed version is cached, $p_c = \sum_i \lambda_i p_c^i / \sum_i \lambda_i$ and $p_o = 0$.

By the Che approximation, the hit rate for version $j \in \{o, c\}$ of the r^{th} most popular video is

$$h_j^{(r)} = 1 - e^{-p_j t_C / r^\alpha}$$

where t_C is the so-called characteristic time for a cache of size C. A straightforward extension of the arguments in [7] shows that $t_{N\delta}$ is asymptotically equal to βN^α where β is the solution to the equation

$$\delta = \sigma_o \int_0^1 (1 - e^{-p_o \beta / x^\alpha}) + \sigma_c \int_0^1 (1 - e^{-p_c \beta / x^\alpha}).$$

In the following we need the overall hit rate for just the compressed version (cell performance does not depend on whether the original version comes from the VTC cache or an upstream server). Denoting this hit rate by h_c, we have

$$h_c = p_c \frac{\sum_r h_c^{(r)} / r^\alpha}{\sum_r 1/r^\alpha}.$$

B. Impact on cell performance

We now turn to the interaction between caching performance and cell performance, accounting for the fact that a requested compressed coded video is not always available as assumed in Section III. We temporarily ignore the possibility to perform on-the-fly video transcoding.

The transition rates of Section III-A remain the same except for the following:

- a_{ic} increases to $a_{ic}+1$ at rate $h_c\lambda_{i2}$ if $R_i/(n+1) < C_o$,
- a_{io} increases to $a_{io}+1$ at rate $(1-h_c)\lambda_{i2}$ if $R_i/(n+1) < C_o$.

The overall compressed version hit rate h_c is independent of the user radio condition i. Its value depends on the assumed caching policy. If both original and compressed versions are cached together, or if only the compressed version is ever cached, the value of h_c does not depend on the radio model (that determines the probability the compressed version is requested). This is so because the hit rates, computed as in Sec. IV-A, do not then depend on the values of p_o and p_c. The hit rate does depend on the radio model when the original and compressed versions are cached independently. For this case, we apply a fixed point approximation, as follows.

We first calculate the probabilities p_o and p_c for each version using an initial estimate of h_c. The hit rate is then calculated using the Che approximation with these relative request rates. The new value of h_c is used to refine the estimates of p_o and p_c, and so on until convergence.

C. Limited transcoding capacity

On-the-fly transcoding improves cell performance by satisfying some requests for the compressed version of videos absent from the VTC cache. Suppose the VTC is equipped to transcode a maximum of T videos in parallel. This capacity is to be shared by a number of cells as depicted in Figure 1. We assume there are K cells and, for the sake of simplicity, that these cells have the same radio and traffic characteristics.

The system operates as follows: on detecting a new type 1 flow, the network operator decides if it should request a compressed version; if so, it checks the VTC cache and downloads the compressed version if present; if not, and if the VTC is currently transcoding less than T videos, the requested video is transcoded on-the-fly; if there is no cached copy and all transcoding capacity is already in use, the cell downloads the original version (leading to rate deficit for some ongoing flows).

Let $\theta(k)$ for $k \geq 0$ be the distribution of the number of type 1 videos currently being downloaded in the compressed version. This marginal distribution is derived by summing probabilities $\pi(\vec{n})$ over states $\vec{n}$ such that $\sum_i a_{ic} = k$. Given k compressed version downloads, the probability that l of these are being satisfied by transcoding, for $0 \leq l \leq k$, can be approximated by the binomial distribution with "success probability" $1 - h_c$. It is then straightforward to compute the mean m and variance v of the number of videos currently

being transcoded. We have,

$$m = \sum_{k \geq 1} k(1 - h_c)\theta(k), \tag{4}$$

$$v = \sum_{k \geq 1} k(1 - h_c)(k - kh_c + h_c)\theta(k) - m^2. \tag{5}$$

We now approximate the aggregate transcoding load from K cells by a Gaussian distribution with mean Km and variance Kv. An estimate of the probability a transcoding request fails, f, is then given by the probability load exceeds capacity,

$$f = \frac{1}{2}\text{erfc}(\frac{T - Km}{\sqrt{2Kv}}) \tag{6}$$

where erfc is the standard error function.

Finally, for a given cell, the probability a request for a compressed type 1 video can be satisfied is given by the revised "hit rate",

$$h_c' = h_c + (1 - h_c)(1 - f).$$

This probability must be substituted for h_c in the transition rates of Sec. IV-B and applied in the fixed point approximation described therein.

V. Numerical results

We present a selection of numerical results intended to illustrate the tradeoffs realized by the considered VTC device and the possibility to optimize performance by appropriate dimensioning.

A. Radio conditions and traffic data

We consider downloads in a nominal 3G system with just two distinct radio conditions: users at the cell edge have a peak rate of 5 Mbps while users at the center have a peak rate of 15 Mbps. This is clearly a crude approximation that nevertheless allows us to capture the significant impact on performance of heterogeneous radio conditions. Half of downloads are assumed to be to edge users and half to center users for all three types.

We suppose 75% of flows are video downloads of which a variable proportion is of type 1 and can be transcoded. The original rate of type 1 videos is $C_o = 1$ Mb/s and we assume the VTC transcodes this, as necessary, bringing the rate down to $C_c = 0.25$ Mb/s. The mean duration of these videos is $\tau_1 = 5$ minutes. The mean size of a regular document download (type 2 flows) is $\sigma_2 = 10$ MB. The mean duration of adaptive streaming videos is $\tau_2 = 5$ minutes. Results are plotted below in terms of the overall flow arrival rate bearing in mind that the cell load is not an independent variable because of the use of transcoding (type 1) and adaptive streaming (type 3).

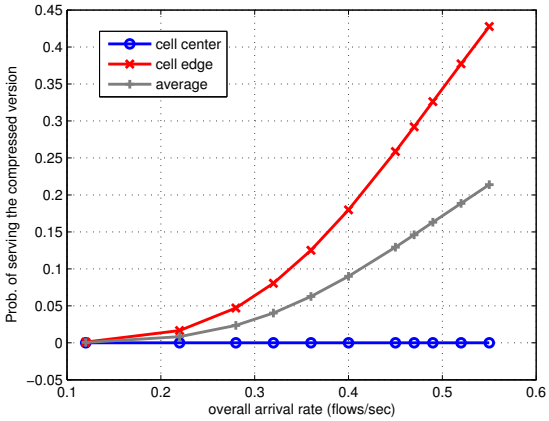

Fig. 2. Probability of being served with the compressed version.

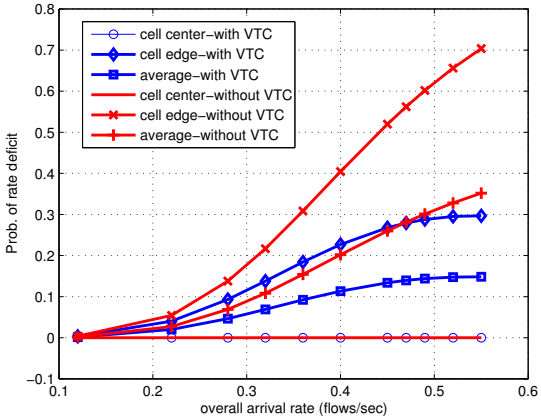

Fig. 3. Probability of rate deficit.

B. Maximum gains from the VTC

We first assume transcoding is not limited by the capacity of the VTC, and the model presented in Section III applies. This enables an evaluation of the maximum potential capacity gain as a function of the traffic mix. Figures 2, 3 and 4 plot the values of the performance criteria, compression probability, rate deficit probability and cell utilization, respectively, under the assumption that 70% of video traffic can be transcoded (i.e., is of type 1).

The results of Figure 2 show that transcoding is mainly performed for users with the worst radio conditions at the cell edge. The proportion is significant for moderate load and increases steadily with load. Figure 3 plots the rate deficit with and without transcoding, distinguishing edge and center users. Again, this performance degradation impacts only the cell edge users. For the particular configuration considered, the average rate deficit probability is reduced by half when transcoding is applied. Lastly, Figure 4 shows how cell utilization is reduced by about 15% by the use of transcoding. Note that utilization increases sub-linearly with the flow arrival rate even without

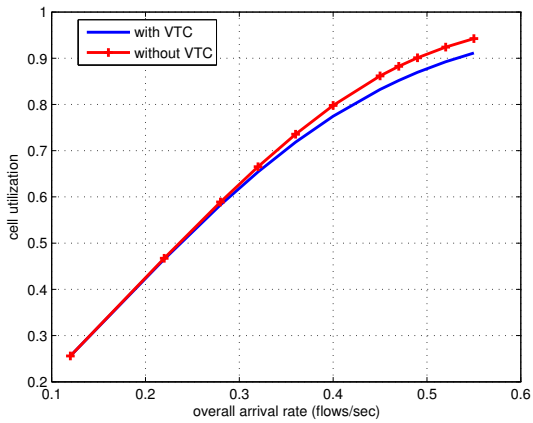

Fig. 4. Cell utilization.

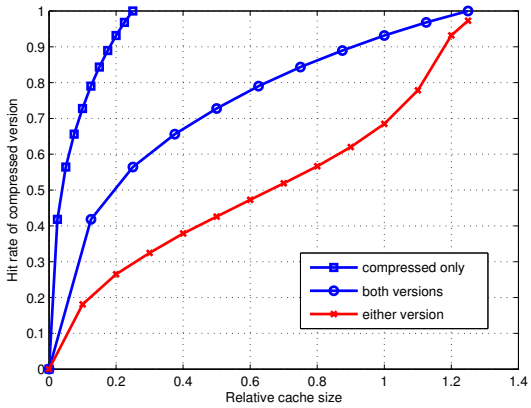

Fig. 5. Hit rate of compressed objects as a function of relative cache size δ.

TABLE I
CAPACITY GAIN FOR DIFFERENT TRAFFIC MIXES (TYPES 1:2:3).

Traffic mix (%)	52.5:22.5:25	37.5:37.5:25	22.5:52.5:25
Original capacity	0.31 flows/s	0.33 flows/s	0.35 flows/s
Gain with VTC	16%	15%	11%

transcoding due to the presence of type 3 adaptive streaming flows. The use of transcoding for type 1 accentuates this effect.

C. Cell capacity

The performance improvement brought by the use of transcoding leads to an increase in network capacity. By this we mean the network is dimensioned to meet performance targets that are reached at higher loads when the VTC is used. This implies for a given level of demand (in flows/sec/m^2) that the cell density can be so much smaller and the investment in infrastructure so much less. In this subsection we quantify the average gain in the capacity of one cell under the following performance thresholds:

- The proportion of users served with the compressed version should be less than 30%. We choose this value as it is commonly used as the threshold on the proportion of calls served at half-rate in a mobile telephone network.
- The probability of rate deficit is limited to a maximum of 10%. This threshold is intended to ensure a sufficiently small probability of stalling in video playouts.
- Cell utilization must be less than 80%. This is a threshold typically used to dimension mobile data networks.

Table I presents some results illustrating the potential capacity gain for different proportions of the three types of flow. The first column corresponds to the traffic mix used in Figures 2, 3 and 4 and is derived by applying the above dimensioning thresholds. It turns out that with this configuration the most severe constraint is the rate deficit probability. This limits cell capacity to 31 flows/s without transcoding, increasing by 16% to 36 flows/s with a perfect VTC. The other two columns show that this gain is naturally less significant as the proportion of type 1 flows in the traffic mix decreases.

This capacity gain is to be set against the cost of the VTC. One can naturally test the impact of alternative dimensioning criteria. If the threshold on the rate deficit probability is increased to 20%, the cell capacity increases but the gain brought by transcoding remains approximately the same.

D. Impact of the cache size and the online transcoding capacity

Recall that the results above are derived assuming the compressed version is always available when requested. In this section we relax this assumption and quantify the impact of cache misses on realizable capacity gains. We first discuss $h_c(\delta N)$ as a function of δ, giving the hit rate of requests for the compressed version for given cache size where the latter is expressed as a fraction of the amount of memory needed to cache the entire catalogue in the original version. We assume a catalogue size of $N = 10^4$ videos having a Zipf popularity distribution with exponent $\alpha = 0.8$. Figure 5 plots this function for the three caching options considered in Section II-B: cache the version requested, cache both versions systematically, cache only the compressed version.

To derive the first curve on the left, we assume the probability the compressed version is requested is $p_c = 0.1$ (cf. Fig. 2). Note the unfavorable impact of only caching the requested version. It is more effective for the hit rate of the compressed version to cache it systematically whenever either version is requested. This is because the popularity of the compressed version is otherwise 10 times less than that of the original version (since $p_c = .1$).

More significantly, comparison with the leftmost curve shows that the cache is much more efficiently used (for our objective of alleviating congestion in the radio part) if only the compressed version is ever cached. This curve is, in fact, identical to the second (caching both versions) with the x coordinate divided by five (since $(C_o + C_c)/C_c = 5$).

To evaluate the impact of the hit rate on cell capacity, we first integrate the effects of a limited transcoding capacity. We suppose the number of cells K is equal to 100 and use the

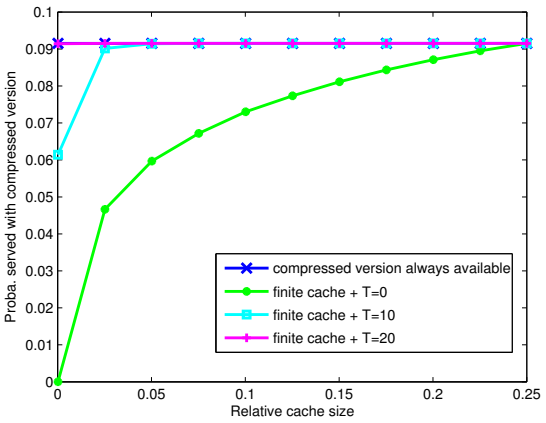

Fig. 6. Probability of being served with the compressed codec for different cache sizes and different online transcoding capabilities.

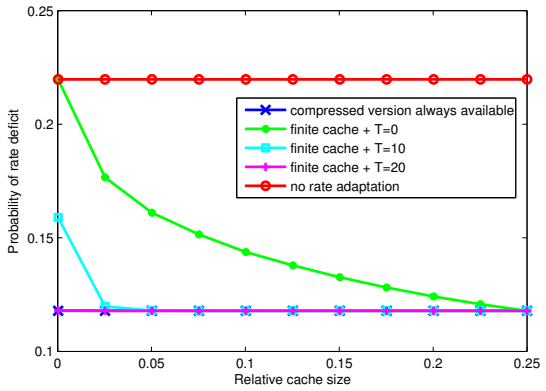

Fig. 7. Probability of rate deficit for different cache sizes and different online transcoding capabilities.

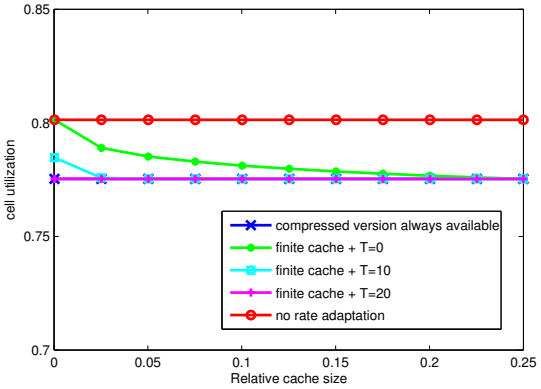

Fig. 8. Cell load for different cache sizes and different online transcoding capabilities.

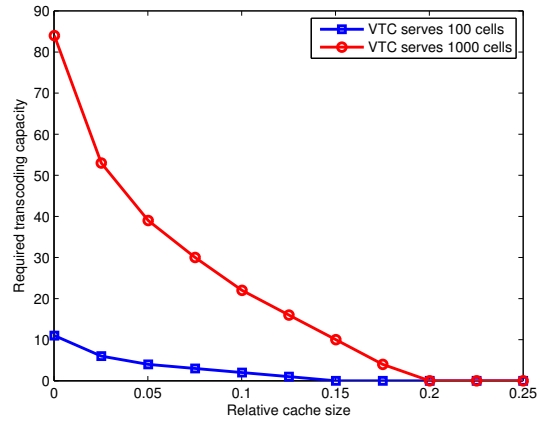

Fig. 9. Max capacity

formulas of Section IV-C to determine the modified hit rate h_c' with h_c evaluated for the option where the VTC only caches the compressed version. Our three performance criteria are then evaluated, as previously indicated.

Figures 6, 7 and 8 plot the compression probability, rate deficit probability and utilization, respectively, as functions of the relative cache size δ. The figures show one curve for each of three chosen values for the VTC transcoding capacity, 0, 10 and 20 videos in parallel.

Obviously, all three performance criteria improve as either the cache capacity or the transcoding capacity increase. More interestingly, the results of the figures suggest the maximum gains discussed in Section V-C can be attained with relatively small VTC capacities. If the transcoding capacity is 10 parallel downloads, a relative cache size of only $\delta = .025$ (i.e., 10% of the entire catalogue in compressed coding) is sufficient. Alternatively, with a coding capacity of 20, it is unnecessary to equip the VTC with a cache (in fact, a transcoding capacity of only slightly more than 10 is sufficient). On-the fly transcoding would only be unnecessary if the cache were sized to contain almost all of the catalogue.

The above results relate to a VTC concentrating the traffic of 100 base stations. It is interesting to understand how the VTC should be dimensioned to to serve a different number of base stations. According to the IRM model of caching, the cache size required for a given hit rate is independent of the concentration factor. On the other hand, the transcoding load and the resulting probability of failure (6) do depend on how much traffic the VTC receives.

Figure 9 shows the transcoding capacity required to realize maximum network capacity gains as a function of the relative cache size for a VTC concentrating 100 and 1000 base stations. These results reveal a transcoding scale economy in that the capacity for 1000 cells is less than 10 times the capacity required for 100. Scale economies are actually more significant in increasing the concentration factor from 10 to 100, say. Of course, scale economies are greatest for caching since cache size is independent of the traffic volume. On the other hand, larger VTCs placed to concentrate a greater number of cells lead to more traffic to be handled in the wired

network between the VTC and the base station. This particular cache-bandwidth tradeoff is significant but out of scope for the present work.

VI. Related work

We are not aware of other work specifically on evaluating the relationship between transcoding and the performance of the radio access network, even though video transcoding in this context has long been recognized as an interesting possibility (e.g., [15]) and is highly relevant to very recent proposals (e.g., [6]). Transcoding has, of course, already been widely used for voice calls through the use of half-rate coding, although the performance model for this service, as proposed by Ivanovich *et al.* [11], for example, cannot be applied to our system. The accepted use of half-rate coding usefully illustrates that a tradeoff of performance criteria, like reduced voice quality for lower call blocking or reduced image quality for less video stalling, is reasonable despite possible concerns over network neutrality issues.

There is currently a large amount of ongoing research into the performance of caching, mainly in the context of information-centric networking. Our simple model of Zipf popularity with the assumed independent reference model is clearly only an approximation. Recent developments, as discussed in the paper by Martina *et al.* [12], for instance, account more accurately for phenomena like catalogue and popularity dynamics. However, as results in the cited paper confirm, the IRM model may be considered to remain a reasonable choice for our intended broad discussion of design choices.

While the literature is sparse on the envisaged transcoding and caching solution, there is a growing body of work on the relation between caching and adaptive video coding. For example, the paper by Grandl *et al.* [8] considers how adaptive coding impacts the performance of a proposed information-centric network solution where variously coded video segments are cached in the access network. Caching is used in this context mainly to reduce traffic in the backhaul, however, and not in the radio link. The paper by Aouine *et al.* [1], on the other hand, is somewhat more relevant in that it envisages a network where the operator intervenes to limit the number of codings available to end users. The objective is to improve the efficiency of caching in reducing backhaul traffic but it is interesting to note that this type of intervention might additionally be made to alleviate radio congestion, like transcoding, by imposing lower rate segments when necessary.

In evaluating the transcoding option, it is important to account appropriately for the performance of the radio link. We have adopted here a relatively simple model linking radio conditions with dynamic, stochastic demand, inspired by the seminal work of Bonald and Proutière [2]. There is clearly scope for refining this model to account more precisely for the physical and link layer mechanisms of real 3G and 4G networks (e.g. [5]). However, for the present purpose of gaining insight and understanding tradeoffs, we believe the simple model is sufficient.

VII. Conclusions

We have proposed a performance model to evaluate the potential gains brought by proposed video transcoder and cache (VTC) devices to be introduced in the downstream mobile access network. The model accounts for heterogeneous radio conditions, the random process of flow arrivals of three distinct types, and the on-the-fly transcoding and caching capacities of the VTC. This model has been developed to enable an operator to appraise the economic interest of these devices, proposed either as stand-alone appliances or as part of a future software-defined radio access network.

The operator's appraisal must take account of multiple criteria defining the system dimensions and costs whose detailed discussion is beyond the scope of the present paper. We believe, however, that the model itself and the sample of numerical results presented here bring some useful insights into the design options.

The gains in network capacity are at first sight modest (no more than 16% in the considered scenarios) though the high cost of the mobile access network may still outweigh the cost of the devices. The 16% gain was evaluated assuming more than 50% of network flows are susceptible to transcoding to a relatively low compressed rate. Unfortunately, the trend is for this proportion to be diminished due to the increasing use by video content providers of encryption, preventing both transcoding and transparent caching. A new business model to align the interests of operator and content providers is perhaps necessary before the gains of the VTC device can be fully exploited.

The current trend to use adaptive streaming for video applications is also reducing the proportion of video demand that is susceptible to transcoding. On the other hand, adaptive streaming applications may be considered to obviate the need for transcoding since they naturally reduce the flow rate when the network is congested or when radio conditions are poor. A simpler device might be employed by an operator to ensure the rate adaptations requested by applications have the desired impact on cell performance.

Online, on-the-fly transcoding and caching are to some extent interchangeable. If the VTC can transcode offline and cache a large proportion of compressed videos using LRU replacement, on-the fly coding is unnecessary. On the other hand, if the device is capable of transcoding a moderate number of videos in parallel, the VTC can dispense with caching. The optimum mix of cache and transcoding capacity depend on relative costs and can readily be derived from the proposed model.

Acknowledgment

This research work has been partially funded by the Technological Research Institute SystemX, within the project "Network Architectures" hosted at LINCS.

REFERENCES

[1] Z. Aouini, M. T. Diallo, A. Gouta, A.-M. Kermarrec, and Y. Lelouedec. Improving caching efficiency and quality of experience with cf-dash. In *Proceedings of Network and Operating System Support on Digital Audio and Video Workshop*, NOSSDAV '14, pages 61:61–61:66, New York, NY, USA, 2014. ACM.

[2] T. Bonald and A. Proutière. Wireless downlink data channels: User performance and cell dimensioning. In *Proceedings of the 9th Annual International Conference on Mobile Computing and Networking*, MobiCom '03, pages 339–352, New York, NY, USA, 2003. ACM.

[3] H. Che, Y. Tung, and Z. Wang. Hierarchical web caching systems: modeling, design and experimental results. *IEEE JSAC*, 20(7):1305–1314, 2002.

[4] Cisco. Visual networking index: Global mobile data traffic forecast update, 20132018, 2014.

[5] R. Combes, S.-E. Elayoubi, and Z. Altman. Cross-layer analysis of scheduling gains: Application to lmmse receivers in frequency-selective rayleigh-fading channels. In *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2011 International Symposium on*, pages 133–139, May 2011.

[6] ETSI. Mobile-edge computing: Introductory technical white paper, 2014.

[7] C. Fricker, P. Robert, and J. Roberts. A versatile and accurate approximation for lru. cache performance. In *Proceedings of ITC 24*, 2012.

[8] R. Grandl, K. Su, and C. Westphal. On the interaction of adaptive video streaming with content-centric networking. In *Packet Video Workshop*, 2013.

[9] T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson. A buffer-based approach to rate adaptation: Evidence from a large video streaming service. In *Proceedings of the 2014 ACM Conference on SIGCOMM*, SIGCOMM '14, pages 187–198, New York, NY, USA, 2014. ACM.

[10] Huawei. Cloud-based media storage and processing data center. Huawei white paper, 2014.

[11] M. Ivanovich, M. Zukerman, P. Fitzpatrick, and M. Gitlits. Performance between circuit allocation schemes for half- and full-rate connections in gsm. *Vehicular Technology, IEEE Transactions on*, 47(3):790–797, Aug 1998.

[12] V. Martina, M. Garetto, and E. Leonardi. A unified approach to the performance analysis of caching systems. In *INFOCOM, 2014 Proceedings IEEE*, 2014.

[13] Nokia. Intelligent base stations. Nokia Networks white paper, 2014.

[14] L. Rong, S. Elayoubi, and O. Haddada. Performance evaluation of cellular networks offering tv services. *Vehicular Technology, IEEE Transactions on*, 60(2):644–655, Feb 2011.

[15] T. Warabino, S. Ota, D. Morikawa, M. Ohashi, H. Nakamura, H. Iwashita, and F. Watanabe. Video transcoding proxy for 3g wireless mobile internet access. *Communications Magazine, IEEE*, 38(10):66–71, Oct 2000.

[16] Y. Xu, E. Altman, R. El-Azouzi, M. Haddad, S. Elayoubi, and T. Jimenez. Probabilistic analysis of buffer starvation in markovian queues. In *INFOCOM, 2012 Proceedings IEEE*, pages 1826–1834, March 2012.