ABSTRACT

The increasing demand of mobile data traffic is starting to stress the networks of mobile network operators (MNOs). Techniques for offloading traffic (partially or totally) from the network of the MNO are currently being designed and deployed at various points of the network, which depend on the goal of each MNO. Such techniques divert traffic to offloading networks, such as Wi-Fi access networks, femtocells, etc., or directly to the Internet. This paper presents a generic analytical approach for studying the impact that offloading techniques might have on the networks of MNOs.

The model is generic in the sense that it is independent of the specific offloading technology used and may be of use to provide bounds on network dimensioning. Based on previous measurements found in the literature, our model assumes that user activity periods and periods characterizing offloading are heavy-tailed. We model them as strictly alternating independent ON/OFF processes. Therefore, the non-offloaded traffic (i.e., that traffic still being served by the MNO on a regular basis) is modeled as the product of these two processes. We prove that the resulting process is long-range dependent, with heavy-tailed ON/OFF-period durations, and its characteristic parameters can be derived from those of the initial processes. We also evaluate the network resources required to serve the traffic resulting from those of the initial processes. We conclude that offloading does not always mean reduction of resource consumption, and that the distribution of offloading periods turns out to be the main design parameter to deploy effective offloading strategies in the networks of MNOs. Extensive simulations, in which the Hill’s estimator was used to obtain the main parameters of the resulting process, confirm the results of our analytical study.

Categories and Subject Descriptors
C.2.1 [Wireless Communications]; I.6.m Simulation and Modeling: Miscellaneous

General Terms
Performance, Theory

Keywords
mobile networks; traffic offloading; modeling; ON/OFF process; heavy-tails; long-range dependence; network dimensioning

1. INTRODUCTION

Mobile data growth is starting to stress the infrastructure of mobile network operators (MNOs). Furthermore, all forecasts predict further increases in the near future [1]. The problem of handling all such data is further exacerbated by the increasing rates offered by new access technologies (e.g., LTE, LTE-A), as the core will have to handle very high volumes of aggregated traffic. A number of techniques are being conceived to offload the core network of MNOs by diverting packets to other networks at different points of the architecture or by confining local traffic to the local network (e.g., enterprise network). Some examples of such techniques are LIPA, SIPTO, I-WLAN, IFOM, and MAPCON [2], [2], [3], [4].

Up to now, the effort has been mainly devoted to designing and, in some cases, to measure the potential gains of using Wi-Fi for offloading. As for the characterization of offloading, [5] analyzes the Wi-Fi connectivity patterns of 100 iPhone users during two and a half weeks, and, based on this, estimates the potential benefits of using Wi-Fi for offloading traffic from the 3G core network. Wi-Fi connectivity patterns are shown to follow a heavy-tail distribution. In the same direction, [6] presents an experimental study of the availability of 3G and WLAN coverage from moving vehicles in three different cities. The authors also deploy...
a system in their testbed that exploits Wi-Fi for offloading when available and then compute the gains obtained.

However, up to our knowledge, there has been no prior work analytically studying such offloading gains. An eventual analytical framework is expected to help in taking network design decisions when deploying offloading techniques. This paper makes a contribution in this direction.

The theoretical framework proposed in this paper is based on traffic models found in the literature. In this sense, and based on previous measurements [5], we characterize the duration of offloading periods as heavy-tailed. On the other hand, [8] and references therein present well-known models of the behavior of single flow traffic, which describe traffic burstiness in terms of long-range dependence and heavy-tails. In the same way, we characterize user activity according to these models.

Therefore, the main contributions of this paper are:

- Theoretical framework for studying mobile traffic offloading, in which both user activity and offloading periods are modeled by means of strictly alternating ON/OFF processes whose ON and OFF durations are characterized by heavy-tailed distributions. Furthermore, the process describing the non-offloaded traffic (i.e., that traffic still needing to be served by the MNO on a regular basis) is modeled as the product of the two previous processes. This framework is agnostic of the specific offloading technique applied as well as the node where offloading is applied.

- Derivation of the exact parameters describing the asymptotic behavior of the non-offloaded traffic, which is shown to be long-range dependent with ON/OFF durations following a heavy-tailed distribution. We also obtain the parameters characterizing this process as a function of those of the original processes.

- Based on the characteristics of the aggregated non-offloaded traffic, this paper provides performance bounds of resource consumption in the network of the MNO when deploying offloading strategies. Furthermore, we illustrate the influence of the different design parameters on the dimensioning of the network before and after applying offloading. We conclude that offloading does not always entail a gain in terms of network resources and that the appropriate design of offloading periods is key in network dimensioning.

- Validation of the analytical results by means of simulations.

The paper is organized as follows. Fundamental concepts on network traffic modeling used in this paper are given in Section 2. In Section 3, we propose a model for the non-offloaded traffic of a single source that eventually reaches the network of the operator. In Section 4, we discuss on the applicability of our model in the context of offloading techniques. Section 5 is devoted to the main results of our study, in particular, the derivation of the parameters characterizing the asymptotic behavior of the non-offloaded traffic. In Section 6, we validate the main results by means of simulations. In Section 7, we study the implications of deploying an offloading strategy on the provisioning of resources in the network of the MNO. Finally, Section 8 concludes the paper.

2. PRELIMINARIES

Let $X(t)$ be a strictly alternating ON/OFF process where ON and OFF periods follow the cumulative distribution functions denoted by $F_{on}(t)$ and $F_{off}(t)$, respectively. Assume now that as $t \to \infty$ the complementary distribution function is

$$F_{on}^{c}(t) = c(1 - t^{-\alpha_{on}}) L_{1}(t)$$

and

$$F_{off}^{c}(t) = c(1 - t^{-\alpha_{off}}) L_{2}(t),$$

where $c > 0$ is a constant and $L_{1}(t) > 0$, is a slowly varying function at infinity [7] ($j = 1$ for ON- and $j = 2$ for OFF-periods), and $a(t) \sim b(t)$ means that $a(t)/b(t) \to 1$ as $t \to \infty$. The ON and OFF sequences are heavy-tailed whenever $1 < \alpha_{on} < 2$ and $1 < \alpha_{off} < 2$, respectively. In such a case, the process $X(t)$ presents heavy-tailed renewal cycles (i.e., one ON period and the following OFF period) with

$$1 < \alpha_{min} = \min(\alpha_{on}, \alpha_{off}) < 2.$$

The literature presents two fundamental results related to this type of processes. First, the authors of [17] show how this type of processes presents long-range dependence (LRD). Specifically, they prove how the asymptotic behavior of the covariance function $\gamma_{3}(t)$ of process $X(t)$ by time $t$ presents the following shape:

$$\gamma_{3}(t) \sim \sigma_{t}^{3} t^{-\alpha_{min}} L(t) \quad as \quad t \to \infty,$$

where $0 < \beta \leq \alpha_{min} - 1 < 1$.

Second, the authors of [7] provide a proof of the fundamental relationship between self-similarity of traffic and the aggregation of a large number of heavy-tailed ON/OFF sources. More specifically, the authors denote as $V(t)$ the variance of the process resulting from the aggregation of several heavy-tailed ON/OFF sources and they prove that

$$V(t) \sim \sigma_{t}^{2} t^{3-\alpha_{min}} L^{*}(t) \quad as \quad t \to \infty,$$

where $\sigma$ is a constant and $L^{*}(t)$ is a slowly varying function at infinity. Then, given this result, the authors show that the fluctuations from the mean level of the aggregated process behave like a fractional Brownian motion (FBM) process (i.e., self-similar traffic) with Hurst parameter $(H)$,

$$\frac{1}{2} < H = \frac{3 - \alpha_{min}}{2} < 1.$$

It is also relevant to the procedure that we follow in this paper to consider the following relations [7]. For the stationary process modeling the ON/OFF behavior of a single source $(X(t))$, the probability that at time $t$ it is in an ON-period is calculated as $P\{(time \; t \; is \; on)\} = \mu_{on} / (\mu_{on} + \mu_{off})$ for all $t$, where $\mu_{on}$, $\mu_{off}$ are the mean durations of an ON-period and an OFF-period respectively. Then, the covariance function of the process can be expressed as

$$\gamma_{3}(t) = \mu_{on}(\tau_{11}(t) - \mu_{on}),$$

where $\tau_{11}(t) = P\{(time \; t \; is \; on) \; | \; time \; 0 \; is \; on\}$. The renewal equation for $\tau_{11}(t)$ is:

$$\tau_{11}(t) = G_{on}(t) + \int_{0}^{t} F_{on}^{c}(t-u)dH_{12}(u),$$

with $G_{on}(t) = P\{(remaining \; life \; of \; the \; first \; ON \; interval > t) \; | \; time \; 0 \; is \; ON\}$, and where $H_{12}(u)$ is the function corresponding to the inter-renewal (the pair of ON- and OFF-periods) distribution $F_{on}F_{off}$. $(H_{12} = \sum_{k=1}^{\infty}(F_{on}F_{off})^{k})$. 


Its density \( h_{12}(u) \), when defined, is the probability density that the end of an OFF-period occurs at time \( u \) given that at time \( t = 0 \) the process is in ON state. To understand (6) note that if at \( t = 0 \) the process is in ON state, then \( t \) can be ON either if it belongs to the same ON-period as time 0 or, if there is a subsequent OFF-period, that there is an OFF/ON transition at some time \( u \) and \( t \) belongs to this subsequent ON-period. The asymptotic behavior of the function \( G_{on}^c(t) \) when \( t \to \infty \) can be derived as [7]

\[
G_{on}^c(t) \sim l_1 t^{-c_{on} + 1} L_1(t),
\]

where \( l_1 \) is a constant.

For a detailed study of heavy-tails, LRD, and self-similar traffic modeling the reader is also referred to [12], [15], [16].

3. TRAFFIC MODEL FOR OFFLOADING

In the same way as in previous literature, the activity of a single source is modeled as a strictly alternating ON/OFF process [7]. In our model, we introduce the effect of offloading over such source by means of an additional strictly alternating ON/OFF process. The resulting traffic sent to the network of the MNO in a regular way (i.e., the non-offloaded traffic) from a single source is then modeled as the product of the two above processes. Therefore, the aggregation of several such resulting processes models the traffic still needing to be handled by the operator in the conventional way, which is the one to be used for the purposes of network dimensioning. The following subsections introduce the details and notation of all above processes.

3.1 Model of user activity

As in previous literature, the activity of a single source/user is modeled as a strictly alternating ON/OFF process, where ON periods are i.i.d., OFF periods are i.i.d., and ON and OFF periods are independent. Furthermore, previous measurements have also shown that such periods follow heavy-tailed distributions (e.g., due to file size distributions, web pages) [8] (and references therein). This is also assumed in our model. Therefore, the process \( Y(t) \) is a stationary binary time series \( \{Y(t), t \geq 0\} \) such that [7]

\[
Y(t) = \begin{cases} 
1 & \text{activity period} \\
0 & \text{idle period} 
\end{cases}
\]

A representation of such a process is presented in Figure 1. During the activity period, the user is transmitting and receiving packets. On the other hand, no traffic is exchanged with any network during idle periods, e.g., user reading time of downloaded content. Notice that \( Y(t) \) only describes source/user behavior and is independent from the network through which the traffic is sent, i.e., no offloading considerations have been made in this process.

3.2 Model of offloading periods

Previous measurements carried out in real networks have shown that smartphone connection and disconnection periods to Wi-Fi access points follow heavy-tailed distributions [5]. This observation led us to also model offloading periods for a single source as strictly alternating ON/OFF process. In the same way as before, we also assume that ON periods are i.i.d., OFF periods are i.i.d., and ON and OFF periods are independent. Therefore, the process \( X(t) \) is a stationary binary time series \( \{X(t), t \geq 0\} \) such that

\[
X(t) = \begin{cases} 
1 & \text{user flow sent through MNO network} \\
0 & \text{user flow sent via offloading network} 
\end{cases}
\]

A representation of such a process is presented in Figure 2. During ON periods, the traffic generated by the source (if any) would be sent towards the network of the MNO in a regular way (i.e., traffic is not offloaded). On the other hand, during OFF periods all traffic generated by the source would be routed through the offloading network (e.g., smartphone under the coverage of a Wi-Fi access point or a femtocell).

3.3 Model of non-offloaded traffic from a single source

The traffic generated by a single source that is treated in a conventional way by the MNO (i.e., non-offloaded traffic) can be modeled as the product of the previously defined processes. That is,

\[
Z(t) = X(t)Y(t).
\]

Figure 3 represents such a process, which is also a strictly alternating ON/OFF process, with ON and OFF periods following heavy-tailed distributions and whose characteristic parameters can be derived from those of the original ones, as shown below. During ON periods, the traffic being generated by the source (i.e., user activity in ON state) is forwarded to the network of the MNO as usual. On the other hand, during OFF periods, either there is no activity from the source or traffic is being sent through the offloading network (e.g., through Wi-Fi, femtocells).

3.4 Problem formulation and high-level view of steps followed to solve it

The goal of this paper is to evaluate the potential benefits that offloading techniques could bring to operators. In other words, their interest is to compare what is the resource consumption in their network without applying offloading and when applying offloading on a large scale (see Section 7). Hence, operators are interested in the behavior of the process modeling the aggregation of several \( Z(t) \) processes. Therefore, our problem is to obtain an analytical model describing the behavior of such an aggregated traffic and the resource consumption it entails.

![Figure 1: User activity is modeled as a strictly alternating ON/OFF process, \( Y(t) \)](image)

![Figure 2: Offloading periods are modeled as a strictly alternating ON/OFF process, \( X(t) \)](image)

![Figure 3: Model of non-offloaded traffic from a single source](image)
A high level view of the steps needed for solving the problem follows. Since, as explained above, $Z(t)$ has ON and OFF periods that follow heavy tailed durations, this process is long-range dependent. Therefore, the aggregation of several such processes is self-similar and can be characterized by means of the Hurst parameter $H$. This can be done by means of the same techniques presented in [7]. In turn, parameter $H$ can be derived from the parameters characterizing the heavy tailed behavior of the ON and OFF periods of $Z(t)$. Besides, such parameters can be obtained from those of processes of the original processes $X(t)$ and $Y(t)$. Therefore, once the parameters of the original processes are known, by following the above steps, we can characterize the behavior of the aggregated non-offloaded traffic, and hence, the resources needed in the network of the MNO to serve it.

4. APPLICABILITY OF THE MODEL

There are various ways in which offloading could be deployed in a network. Currently, the most popular one is the use of Wi-Fi (instead of 3G), when available, for all the traffic exchanged by the user [2]. With the advent of femtocells, some parts of the network of the operator could also be offloaded when the user is at home or at an enterprise, as her traffic would be routed through the femtocell.

In this paper, we are interested in modeling the traffic that eventually reaches the network of the operator in a regular way (i.e., the non-offloaded traffic). Our goal is to analytically study the benefits (if any) that offloading generates in the network of the MNO in terms of resource consumption. In this sense, the traffic pattern generated towards the network of the MNO in the two above scenarios (i.e., Wi-Fi or femtocells) is the same. That is, all traffic generated by the user when outside coverage of a Wi-Fi access point (or a femtocell) is sent to the network of the MNO as usual. On the other hand, traffic is fully offloaded when under the coverage of the access point. Our model captures this traffic pattern disregarding the technology in use for offloading, and in this sense, it is agnostic from the specific technology in use.

As mentioned in Introduction, there are many other offloading techniques (e.g., I-WLAN, MAPCON, IFOM). As a group, the main qualitative difference with respect to those explained in the previous paragraph is that offloading is done on a per-flow basis, i.e., some flows are offloaded and some other flows are not when under the coverage of a Wi-Fi access point or a femtocell. In a more generic case, there may be intermediate nodes in which offloading decisions are taken (e.g., the gateway of a company towards the network of the MNO). However, in terms of traffic pattern generated towards the network of the MNO, the resulting flow would behave in the same way as in previous cases. Therefore, our model might still be used for those flows that are offloaded, no matter where the offloading point is in the network (e.g., terminal, femtocell/access point, intermediate node). Hence, the total aggregated traffic toward the network of the MNO would result from the aggregation of offloaded flows (output of our model) with regular flows that are not offloaded at all.

5. MAIN RESULTS

**Lemma 1.** Let $Z(t)$ be the process resulting from the product of processes $X(t)$ and $Y(t)$. Let these two latter processes have i.i.d. ON periods, i.i.d. OFF periods, and independent ON and OFF periods with heavy-tailed distributions. The complementary distribution function for the ON-periods of process $Z(t)$ satisfies

$$F_{on}^{cz}(t) \sim c_1^z t^{-\alpha_{on}} \bar{L}_1(t),$$

where

$$\alpha_{on} = \alpha_{X} + \alpha_{Y} - 1,$$

and $\alpha_{on}$ and $\alpha_{on}'$ are the parameters characterizing the heavy tails of the ON and OFF periods of processes $X(t)$ and $Y(t)$, respectively. Hence, as $\alpha > 1$,

$$\alpha_{on} > \max\{\alpha_{on}', \alpha_{on}''\}. \quad (13)$$

Additionally, the complementary distribution function for the OFF-periods of process $Z(t)$ satisfies

$$F_{off}^{cz}(t) \sim c_2^z t^{-\alpha_{off}} \bar{L}_2(t),$$

where

$$\alpha_{off} = \min\{\alpha_{X,off}, \alpha_{Y,off}\}. \quad (15)$$

In (11), (14) $c_1^z > 0$ is a constant and $\bar{L}_1(t) > 0$ is a slowly varying function at infinity.

**Proof.** (Lemma 1) First, we prove the first part of Lemma 1 concerning the complementary distribution function of the ON-periods of process $Z(t) = X(t)Y(t)$.

The complementary distribution function of the duration of the ON periods of process $Z(t)$, $F_{on}^{cz}(t)$, can be expressed as

$$F_{on}^{cz}(t) = 1 - F_{on}(t) = P(ON > t) = P(ON_X > t)P(ON_Y > t). \quad (16)$$

Let us consider all possible options for the transitions from OFF to ON periods of process $Z(t)$. There are three possible situations. First, processes $X(t)$ and $Y(t)$ were in an OFF period and at a certain instant they both switched to ON period (we refer to this case as OFF/ON to ON/ON transition). Second, $X(t)$ was in an ON period and $Y(t)$ in an OFF period, and at a given random instant $Y(t)$ switches to an ON period (OFF/ON to ON/ON). And third, $X(t)$ was in an OFF period and $Y(t)$ in an ON period, and at a given instant $X(t)$ switches to an ON period (OFF/ON to ON/ON). Then, for the first type of transition, the duration of the ON-period of process $Z(t)$ is the minimum of the ON durations of $X(t)$ and $Y(t)$, i.e., the intersection of the ON periods of both processes. When the first type of transition.

Figure 3: Non-offloaded traffic from a single source is modeled as the product of two strictly alternating ON/OFF processes, $Z(t) = X(t)Y(t)$
happens, the probability that the duration of the ON period of process $Z(t)$ is greater than $t$, can be expressed as the product of the complementary distribution functions of $X(t)$ and $Y(t)$, i.e., $F^{cx}_{on}(t)F^{cy}_{on}(t)$. As for the second type of transition, the duration of the ON-period of process $Z(t)$ is calculated as the minimum between the ON duration of process $X(t)$ and the remaining lifetime of the ON period of process $X(t)$, and the probability of this duration being greater than $t$, can be expressed as $G^{cx}_{on}(t)F^{cy}_{on}(t)$. Similarly, for the third type of transition, we have $F^{cx}_{on}(t)G^{cy}_{on}(t)$.

By the law of total probability, the function $F^{cx}_{on}(t)$ can then be expressed as

$$F^{cx}_{on}(t) = P[off/off \rightarrow on/on]F^{cx}_{on}(t) + P[on/off \rightarrow on/on]G^{cx}_{on}(t)F^{cy}_{on}(t) + P[off/off \rightarrow on/on]G^{cx}_{on}(t)G^{cy}_{on}(t),$$

where $P[off/off \rightarrow on/on]$, $P[on/off \rightarrow on/on]$, and $P[off/off \rightarrow on/on]$ are the probabilities of occurrence of each of the above transitions.

Considering that the ON periods of $X(t)$ and $Y(t)$ are heavy-tailed (see eq. (1)), the complementary distribution functions of their ON-periods are

$$F^{cx}_{on}(t) \sim c_1t^{-\alpha_0}L_1(t); F^{cy}_{on}(t) \sim c_1t^{-\beta_0}L_1(t),$$

respectively. Furthermore, and according to (7), function $G^{cx}_{on}(t)$ for the ON-periods of the processes $X(t)$ and $Y(t)$, respectively, is

$$G^{cx}_{on}(t) \sim t_1^{1-\alpha_0}L_1(t); G^{cy}_{on}(t) \sim t_1^{1-\beta_0}L_1(t).$$

Taking into account (18) and (19), expression (17) can be expressed as:

$$F^{cx}_{on}(t) \sim c_1t^{-(\alpha_0 + \alpha_0)}L_1(t) + c_1t^{-(\alpha_0 + \beta_0)}L_1(t) + c_1t^{-(\alpha_0 + \gamma_0 + 1)}L_1(t),$$

where $c_1$, $c_2$, and $c_3$ are positive constants and $L_1(t)$, $L_1''(t)$ are positive slowly varying functions at infinity. The asymptotic behavior of terms with $t^{-(\alpha_0 + \beta_0 + 1)}$ decrease significantly slower than that with $t^{-(\alpha_0 + \gamma_0 + 1)}$ as $t \rightarrow \infty$. In other words, since $t^{-(\alpha_0 + \gamma_0 + 1)} = o(t^{-(\alpha_0 + \gamma_0 + 1)})$ as $t \rightarrow \infty$, the asymptotic behavior of $F^{cx}_{on}(t)$ can be expressed as

$$F^{cx}_{on}(t) \sim c_1t^{-(\alpha_0 + \gamma_0 + 1)}L_1(t).$$

Hence, $\alpha_0^z = \alpha_0 + \alpha_0 - 1$, and, as $\alpha_0^z$ and $\gamma_0^z$ are greater than 1, $\alpha_0^z > max(\alpha_0, \gamma_0)$. This concludes the proof of the first part of Lemma 1.

The second part of Lemma 1 concerning the complementary distribution function of the OFF-periods of process $Z(t)$ is proved in a similar way. But, in this case, we characterize transitions of $Z(t)$ from an ON period to an OFF period. Thus, the same three possible transitions from ON to OFF depending on processes $X(t)$ and $Y(t)$. We refer to each of them as $on/on \rightarrow off/off$, $on/on \rightarrow off/on$, and $on/on \rightarrow on/off$, respectively.

Let us now focus on the complementary distribution function for the OFF period of $Z(t)$. The main difference with that of its ON period is that when one of the two processes ($X(t)$ and $Y(t)$) is in an OFF period, the state of the other process does not matter, as $Z(t)$ will be in an OFF period. Therefore, there may be multiple combinations that generate an OFF period of $Z(t)$. So, for the first type of transition the probability that the OFF period lasts more than $t$ is $F^{' off}(t) + F^{off}(t) + o(t^{-\alpha_0^0}) + o(t^{-\gamma_0^0})$, where the small $o$ terms represent the less likely combinations that eventually generate an OFF period. As for the second type of transition, we have $F^{off}(t) + o(t^{-\alpha_0^0}) + o(t^{-\gamma_0^0})$. And for the third one, such probability is $F^{off}(t) + o(t^{-\gamma_0^0})$.

By the law of total probability, function $F^{cx}_{off}(t)$ can be expressed as follows

$$F^{cx}_{off}(t) = P[on/off \rightarrow off/off] \times (F^{cx}_{off}(t) + F^{off}(t) + o(t^{-\alpha_0^0}) + o(t^{-\gamma_0^0}) + P[on/on \rightarrow off/on]F^{cx}_{off}(t) + o(t^{-\alpha_0^0}) + \sigma_{on/off}^2).$$

Since the OFF periods of processes $X(t)$ and $Y(t)$ have heavy-tailed durations (and in accordance with eq. (1)), their complementary distribution functions satisfy

$$F^{off}(t) \sim c_2t^{-\alpha_0^0}L_2(t); F^{off}(t) \sim c_2t^{-\gamma_0^0}L_2(t).$$

Then, taking into account (23), expression (22) can be expressed as

$$F^{cx}_{off}(t) \sim c_2t^{-\alpha_0^0}L_2(t) + c_2t^{-\gamma_0^0}L_2(t),$$

where $c_1$, $c_2$, and $L_2(t)$, $L_2''(t)$ are positive slowly varying functions at infinity.

Hence $\alpha_0^0 = \min\{\alpha_0^0, \alpha_0^0, \alpha_0^0\}$. This concludes the proof of the second part of Lemma 1.

Theorem 1. Let $Z(t)$ be the process resulting from the product of two strictly alternating ON/OFF processes $X(t)$ and $Y(t)$ with i.i.d. ON periods, i.i.d. OFF periods, and independent ON and OFF periods with heavy-tailed distributions. Then, $Z(t)$ is long-range dependent with covariance function:

$$\gamma_z(t) \sim \sigma_z^2t^{-\alpha_{min}}$$

where

$$\alpha_{min} = \min\{\alpha_{on}, \alpha_{off}\} = \min(\alpha_0^0 + \alpha_0^0 - 1, \alpha_0^0, \gamma_0^0).$$

Proof. (Theorem 1) The proof goes from the expression (2) and Lemma 1 (the expressions (12),(15)).

6. VALIDATION OF MAIN RESULTS

Taking into account the fact that both processes $X(t)$ and $Y(t)$ are heavy-tailed, we use the Pareto distribution with parameters $\alpha$ and $K$ (the scale parameter) to model both the ON- and OFF-durations of the processes. In our simulations we use the Hill’s estimator [10] that provides a good way to estimate tail-index $\alpha$ of Pareto-like tails [11]. In accordance with the methodology presented in [11], the ON- and OFF-durations were simulated with $10^{10}$ states and 15000 largest sequences were considered out of them in the estimation procedure. For a case study initial data were formed on a basis of statistical information provided in the experimental studies on the performance of 3G mobile data offloading through Wi-Fi networks [5] (for the process $X(t)$) and the burstiness of web traffic caused by accesses of HTTP clients.
Figure 4: Estimation of the tail-index \( \alpha_{min}^x \) by means of Hill’s estimator when \( \alpha_{min}^x = \alpha_{off}^x \)

Figure 5: Estimation of the tail-index \( \alpha_{on}^x \) by means of Hill’s estimator

to the heterogeneous contents (text, images, PDF files, etc.) [9] (for the process \( Y(t) \)).

Thus, for the process \( X(t) \) we use the following initial data: \( \alpha_{on}^x = 1.6, \alpha_{off}^x = 1.1, K_{on}^x = 12.5, K_{off}^x = 14.5, \rho_{on}^x = 0.25, \mu_{on}^x = 40 \text{ min}, \mu_{off}^x = 120 \text{ min}. \)

For the process \( Y(t) \) the input data are: \( \alpha_{on}^y = 1.51, \alpha_{off}^y = 1.27, K_{on}^y = 6.1, K_{off}^y = 5.96, p_{on}^y = 0.39, \mu_{on}^y = 18 \text{ s}, \mu_{off}^y = 28 \text{ s}. \)

In accordance with the simulations the tail-index of the process \( Z(t) \) is \( \alpha_{min}^z \approx 1.1 \) as illustrated in Figure 4. This agrees with Theorem 1 proved in Section 5 according to which \( \alpha_{min}^z = \min\{\alpha_{on}^z, \alpha_{off}^z, \alpha_{off}^y\} = \min\{2.11, 1.1, 1.27\}. \) In this case (26), \( \alpha_{min}^z = \alpha_{off}^y. \)

Figure 5 shows simulations to estimate the parameter \( \alpha_{on}^z \), which in accordance with Lemma 1 (12) should be equal to 2.11 (\( \alpha_{on}^z + \alpha_{off}^y - 1 \)). The result of the simulations matches well with the analytical result.

The extensive simulations using Hill’s estimators were carried out for different sets of initial data and in all cases values of the tail-index \( \alpha_{min}^z \) were estimated correctly in accordance with the analytical study presented in Section 5. As an example, for the data set \( \alpha_{on}^z = 1.17, \alpha_{off}^z = 1.8, \alpha_{on}^y = 1.23, \alpha_{off}^y = 1.55 \), we have a case when in accordance with (26) \( \alpha_{min}^z = \alpha_{on}^y = 1.4. \) As seen from Figure 6 it is perfectly agree with the result of the simulations.

7. PERFORMANCE BOUNDS IN OFFLOADING SCENARIOS

This section adopts the perspective taken in [13] and [14] in order to evaluate the system-wide implications of the findings in Section 5. Specifically, we study which are the large-scale implications of implementing an offloading strategy on the provision of resources in a Mobile Network.

In particular, Norros [13] derives an approximation of the capacity that a system requires under fractional Brownian traffic in order to provide a target level of QoS, expressed as a bound (\( \epsilon \)) on the amount of traffic lost. Let us consider an arrival traffic model such as

\[
A(t) = \lambda + \sqrt{a\lambda B_H(t),} \quad (27)
\]

where \( \lambda \) is the average amount of traffic that arrives to the network, \( a \) is a variance coefficient and \( B_H(t) \) is a normalized FBM process with Hurst parameter \( H \).

Given \( A(t) \), the capacity needed to guarantee that the probability of system buffer overflow is bounded \( P(W > w) < \epsilon \) can be approximated by \([13] \),

\[
C \approx \lambda + (H^H(1 - H)^{1-H})(2\ln(2\epsilon))^{\frac{1}{H}} w^{-\frac{1}{H}} \frac{1}{\lambda^H}, \quad (28)
\]

As Theorem 1 reveals, when applying offloading to a heavy-tailed ON/OFF source, the distribution of the resulting sequence of ON/OFF durations may still keep heavy-tail behavior but with a different behavior of the tail. At large scale, when aggregating a large number of sources, the result is that the burstiness of the arrival flow has changed.

Let us denote as \( H_x \) and \( H_y \), the Hurst parameters of the aggregate arrival process to a system when we apply an offloading strategy \( (H_x) \) and when it is not applied \( (H_y) \). Taking into account Theorem 1 and the relation \( H = (3 - \alpha_{min}^y)/2 \) we make the following observations:

\[
\begin{align*}
H_x > H_y, & \quad \text{if } \alpha_{min}^x = \alpha_{off}^x \text{ and } \alpha_{off}^x < \alpha_{min}^y, \\
H_x \leq H_y, & \quad \text{otherwise}
\end{align*}
\]

(29)
From a network dimensioning perspective relation (29) has a high relevance. Specifically, the expression (29) states that special care should be taken to control the distribution of the duration of offloading periods. In particular, when disconnection (offloading) durations present higher heavy-tailness than the original system (i.e., $\alpha_{\text{off}} < \alpha_{\text{min}}$), the net result is that the offered load to the Mobile Network will present a higher degree of burstiness (i.e., $H_z > H_y$).

In order to illustrate this fact let us consider the following example. Assume we have a network without the offloading service dimensioned to support an offered load of $\lambda_y = 200$ Mbps with an outage probability of $\epsilon = 10^{-5}$ given a buffering capacity of $w = 100$ kbytes. Assume further that the system aggregates heavy-tailed ON/OFF sources with pareto-like ON and OFF durations with parameters $\alpha_{\text{on}} = 1.3$ and $\alpha_{\text{off}} = 1.8$.

Now assume that we introduce an offloading strategy in the previous system. Taking into account relation (29) we can define a worst case scenario when $\alpha_{\text{off}} \rightarrow 1$. We can also define a best case scenario when both $\alpha_{\text{on}}$ and $\alpha_{\text{off}}$ do not have a heavy-tailed distribution (i.e., $\alpha_{\text{on}} = 2$ and $\alpha_{\text{off}} = 2$). Applying (28), Figure 7 plots the amount of resources that the system needs to offer the quality of service specified ($\epsilon = 10^{-5}$). Specifically, the figure shows the capacity needed in the original system together with the worst and best case scenarios when we offload a 50% of the traffic.

The figure illustrates two important observations. First, as explained above, the parameter $\alpha_{\text{off}}$ has a capital importance as we can end up in an extreme case where we start offloading data but we need to increase the amount of resources of the network to maintain a certain level of QoS (due to an increase of data burstiness). Second, introducing a fine-grained control of the distribution of offloading periods can be highly beneficial as it can also lead to decreasing the burstiness of data traffic to the core of the network.

An alternative way to present this last observation is computing the amount of overprovisioning that the system needs to keep up with a target QoS. In this context, overprovisioning is understood as the resources beyond the average load that the operator must put in place to serve bursts of traffic. Figure 8 plots this for the same scenario as in the previous figure. In particular, it plots, for a fluid model, the excess of resources over the total capacity that the network operator needs to implement to maintain the objective probability of loss ($\epsilon$). As the figure shows, in the worst case scenario even though we reduce the amount of load to the system, we need to increase the overprovisioning margin to keep up with the QoS. The contrary happens in the best case scenario where in addition to data traffic reduction the operator can also relax overprovisioning needs.

Finally, let us consider Figure 9, which plots the relation $C_z/C_y$ between the resources needed after and before offloading with respect to the tail index of offloading periods ($\alpha_{\text{off}}$). The figure illustrates the influence of the different design parameters on the dimensioning of the network after applying offloading. In particular, the figure focuses on illustrating the strong dependence of the effectiveness of offloading on the actual distribution of offloading periods. Taking as a reference the dashed line in the figure indicating the amount of traffic offloaded (i.e., $P(X(t)$ is in ON state), it can be seen that when the distribution of offloading periods exhibits large tails (i.e., $\alpha_{\text{off}}$ is low) there is a high probability that the network operator needs to increase resources to maintain QoS. Another interesting observation is that the reduction on the number of resources is best (and cannot do better than) when $\alpha_{\text{off}} \geq \min(\alpha_{\text{off}}, \alpha_{\text{on}})$.

As a summary, this section illustrates that the main consequence of Theorem 1 is that the distribution of offloading periods turns out to be the main design parameter in order to implement effective offloading strategies in Mobile Networks.

8. CONCLUSION

This paper presents an analytical framework for modeling traffic in mobile networks implementing offloading. The framework is agnostic of both the specific offloading technique and the point of the network in which it is applied (e.g., terminal, access point). The non-offloaded traffic generated by a single source ($Z(t)$), i.e., the traffic served in a regular way by the operator, is modeled as the product of two strictly alternating ON/OFF processes with ON and OFF independent heavy-tailed durations. One of these processes models the activity of the user ($Y(t)$) and the other one models the offloading periods ($X(t)$).
By studying its asymptotic behavior, we prove that the ON/OFF process $Z(t)$ has ON and OFF periods with durations that follow a heavy-tailed distribution, hence making it long-range dependent. We also derive the tail-indices describing the heaviness of the ON and OFF periods of $Z(t)$ as a function of tail-indices of the two original processes. The analytical results were validated by means of simulations.

By studying the characteristics of the aggregation of several sources of non-offloaded traffic, we provide performance bounds of the resource consumption in the network of the mobile operator, which can be used when dimensioning the network. Furthermore, we quantitatively evaluate the benefits of offloading by comparing the required resources before and after deploying offloading for providing a given quality of service.

One of the main conclusions of our research is that offloading does not necessarily entail less resource consumption in the network of the operator. Under certain conditions, and due to an increase of the burstiness of the non-offloaded traffic, the amount of network resources to offer a given level of QoS is increased. In this context, the tail-index of offloading periods ($\alpha_{off}$) is the most important design parameter to make non-offloaded traffic less heavy-tailed, hence reducing the resources needed in the network of the operator.

9. REFERENCES


