Robust admission control for streaming and elastic services in cellular networks

Elena Bernal-Mor, Vicent Pla and Jorge Martinez-Bauset
Dept. Comunicaciones, Universidad Politecnica de Valencia, UPV
Camino de Vera s/n, 46022, Valencia, Spain
Email: elbermo@upvnet.upv.es, vpla@dcom.upv.es, jmartinez@upvnet.upv.es

***Abstract*—The specific features of cellular networks and especially terminal mobility make the session admission control (SAC) in such networks more complex. This paper studies the robustness of the Virtual Partitioning (VP) admission policy in connection with multiservice cellular networks and considering streaming and elastic traffic in scenarios that must support high overloads. The VP policy is compared with the Multiple Fractional Guard Channel (MFGC) policy. The main contributions of the paper are the study of a new design method, the integration of streaming and elastic traffic, the study of the sensitivity to the channel holding time distribution and the use of absorbing Markov processes to calculate the probability that a handover occurs.**

I. Introduction

The specific features of cellular network and especially terminal mobility make the session Amission Control (SAC) in such networks more complex. Quality of Service (QoS) of all ongoing sessions and new ones accepted must be provided. Furthermore, future mobile systems are expected to support a large variety of services that carry different types of traffic. The traffic can be broadly categorized as streaming or elastic [1].

This work is focused on networks that support high levels of congestion, where there are high priority classes that can generate high demands. The main problem is to guarantee the needs of all classes under high overloads. We can find these scenarios, for instance, in public cellular networks that support emergency services after a disaster applied for broadband wired networks [2] . This paper studies a new policy based on Virtual Partitioning (VP) policy [3], in connection with multiservice cellular networks and considering both streaming and elastic traffic.

The efficiency of several SAC policies in cellular networks have widely been studied in previous papers [4], [5]. For single service scenarios trunk reservation policies like the Guard Channel (GC) [6] and Fractional Guard Channel (FGC) [4] are optimal for common QoS objective functions [4]. But those studies are made considering scenarios with only streaming traffic.

Robustness is the ability to respond to statistical fluctuations and also the adaptability in an overloaded scenario where arrival rates are higher than the expected values. Robustness has been studied in the literature [3] but considering only streaming traffic.

This paper considers both streaming and elastic traffic. Elastic flows are generally transported over TCP which takes care of rate adaptation and bandwith sharing among the different flows. If the total traffic demand of elastic flows exceeds the available capacity some flows might be aborted due to impatience. Flow impatience can arise from human impatience or because TCP or higher layer protocols interpret that the conexion is broken. Abandonments are useful to cope with overload and serve to stabilize the system but, on the other hand, this phenomenon will have a negative impact on the efficiency because capacity is wasted by non-completed flows. That fact leads to think that SAC should also be enforced for elastic traffic [1]. There are previous works that study streaming and elastic traffic but in cellular networks is rare. In [7] elastic and streaming is studied but impatience and mobility are not considered.

The rest of the paper in structured as follows. In Section II the system model is described. Section III describes the design of VP parameters without mobility and in Section IV VP in an cellular scenario is studied. In Section V VP and other policies are compared. Finally, Section VI concludes the paper.

II. Model description

We consider a single cell, which has a total of C resource units where each resource unit has a capacity of R bits per second. The system offers N_s differents classes that carry streaming traffic and N_e classes with elastic traffic. Thus, the total number of classes in the system is $N = N_s + N_e$. Streaming classes use the resource units they need and the resource units that are not occupied by streaming traffic are used by elastic traffic. To avoid starvation the system reserves 1 resource unit for elastic traffic. For each type of service new and handover arrivals are distinguished, so that there are N types of service and $2N$ arrival types. Arrivals are numbered in such manner that for service i new arrivals are referred to as arrival type i, whereas handover arrivals are referred to as arrival type $N + i$.

For the sake of mathematical tractability we make the common assumptions of Poisson arrival processes. The arrival rate for new (handover) sessions of service i is λ_i^n (λ_i^h) where for $i = 1, \ldots, N_s$ are streaming classes and for $i = 1+N_s, \ldots, N$ are elastic classes.

The system state is described by the N-tuple $\boldsymbol{x} = (x_1, \ldots, x_N)$, where x_i ($i = 1, \ldots, N$) represents the number of type-i sessions regardless they were initiated as new or handover arrivals.

For streaming classes, a request of service i consumes b_i resource units, $b_i \in \mathbb{N}$. The service duration of service

$i = 1, \ldots, N_s$ is exponentially distributed with rate μ_i^c. The cell residence time of a service $i = 1, \ldots, N_s$ is exponentially distributed with rate μ_i^r. Hence, the resource holding time in a cell for service $i = 1, \ldots, N_s$ is exponentially distributed with rate $\mu_i = \mu_i^c + \mu_i^r$ and the mean number of handovers per session, when the number of resource units is infinite is $N_i^h = \mu_i^r/\mu_i^c$. We consider the system is homogeneous and in statistical equilibrium. Since the blocking and forced termination probabilities are close to the objectives and they have low values, the handover arrival rates are related to the new session arrival rates through the expression, $\lambda_i^h = \lambda_i^n N_i^h$.

It is assumed that each elastic flow is rate limited by terminal capabilities, i.e. it will receive its fair share of the ratio link bandwidth up to a maximum b_i^M. Moreover, every elastic service must receive at least a minimum bandwidth represented by b^m. The average rate at which clients are served depends on the flow size (given in bytes) since the average rates are given in flows/s. We assume here that the flow size has a exponential distribution with mean L. A client in the system might become impatient and decide to leave the system, this implies an abandonment and if there are a high number of abandonments the capacity of the system is wasted. To model that behavior it is considered the impatience rate $\beta_i(\boldsymbol{x})$. Moreover we define the following parameters: The maximum number of flows in the system $n_M = \lfloor C/b_m \rfloor$, and the maximum service rate of class i is $\mu_i^M = b_i^M/L$.

The amount of resources occupied at state $\boldsymbol{x}$ by streaming traffic is represented by $b(\boldsymbol{x}) = \sum_{i=1}^{N_s} x_i b_i$ and the total number of elastic flows in the system in state $\boldsymbol{x}$ is defined as $c(\boldsymbol{x}) = \sum_{i=N_s+1}^{N} x_i$.

For elastic classes we consider, without loss of generality, that flows are sorted in increasing order of their rate-limits b_i^M. Then, the service rate $\mu_i^{c,e}(\boldsymbol{x})$ and the impatience rate $\beta_i(\boldsymbol{x})$ per flow when the system is in state $\boldsymbol{x}$ are defined as:

$$\mu_i^{c,e}(\boldsymbol{x}) = \begin{cases} \min\left(\mu_i^M, \frac{(C-b(\boldsymbol{x}))R/L}{c(\boldsymbol{x})}\right) & i = 1 \\ \min\left(\mu_i^M, \frac{(C-b(\boldsymbol{x}))R/L - \sum_{j=1}^{i-1}(x_{j+N_s}\mu_j^{c,e}(\boldsymbol{x}))}{\sum_{j=i}^{N_e} x_{j+Ns}}\right) & i \neq 1 \end{cases} \quad (1)$$

$$\beta_i(x) = \beta_i^1 \left(\frac{\mu_i^M}{\mu_i^{c,e}(\boldsymbol{x})} - 1\right) + \beta_i^0.$$

The resource holding rate and the impatience rate of elastic classes in state $\boldsymbol{x}$ are respectively $\mu_i^e(\boldsymbol{x}) = x_{i+N_s}(\mu_i^{c,e}(\boldsymbol{x}) + \mu_i^{r,e})$ and $\beta_i^T(\boldsymbol{x}) = x_{i+N_s}\beta_i(\boldsymbol{x})$, where $\mu_i^{r,e}$ is the cell residence rate for elastic classes.

The aggregated arrival rate is $\lambda^T = \sum_{1 \leq i \leq N} \lambda_i^n$ and f_i represents the fraction of λ^T that correspond to service $i = 1, \ldots, N$, i.e. $\lambda_i^n = f_i \lambda^T$, where $0 \leq f_i < 1$ and $\sum_{1 \leq i \leq N} f_i = 1$.

Let us denote by $P_i^{b,n}$ the blocking probabilities for new arrivals and $P_i^{b,h}$ the blocking probabilities for handover arrivals. For elastic traffic we define P_i^a the abandonment probability. If all flows were let into the system, the abandonment probability may be considered a good and sufficient performance indicator. However, if there is some type of access restriction, both abandonment and blocking should be taken into account for characterizing system performance. Therefore, we define the success completion probability P_i^c which represents the probability that a flow is not blocked and it does not leave the system before being served due to impatience.

For streaming traffic, the QoS requirements are given in terms of upper-bounds for the new arrival blocking probabilities (B_i^n) and the handover failure probabilities (B_i^h), and for elastic traffic are given by lower-bounds of success completion probabilities (B_i^e).

The model of the system is a multidimensional birth and death process whose set of feasible states is

$$W := \left\{ \boldsymbol{x} : x_i \in \mathbb{N}; \quad \sum_{i=1}^{N_s} x_i b_i \leq C - 1; \quad \mu_i^{c,e}(\boldsymbol{x}) \geq b_m/L \right\}.$$

The coeficients $a_i^n(\boldsymbol{x})$ and $a_i^h(\boldsymbol{x})$, which depend on the SAC, denote the probabilities of accepting a new and handover arrival of service i respectively and $\pi(\boldsymbol{x})$ is the state stationary probablity. Then, the blocking probabilities for class i, where $i = 1, \ldots, N$, are obtained as

$$P_i^{b,n} = \sum_{\boldsymbol{x} \in W} (1 - a_i^n(\boldsymbol{x}))\pi(\boldsymbol{x}), \quad P_i^{b,h} = \sum_{\boldsymbol{x} \in W} (1 - a_i^h(\boldsymbol{x}))\pi(\boldsymbol{x}).$$

Likewise, the abandonment probabilities for class i, where $i = 1 + N_s, \ldots, N$, are obtained as

$$P_i^a = \frac{1}{\lambda_i^n(1 - P_i^{b,n}) + \lambda_i^h(1 - P_i^{b,h})} \sum_{\boldsymbol{x} \in W} \beta_{i-Ns}^T(\boldsymbol{x})\pi(\boldsymbol{x}).$$

To calculate the success completion probability, let us define P_i^h as the probability that the service time is higher than the cell residence time for service i, i.e. the mobile terminal i leave the current cell and continue its service in another cell. Furthermore, we define P' as the probability that the service of a session that has arrived in the cell after a handover finishes successfully. Let $\bar{P}$ be the complementary probability of P, $\bar{P} = 1 - P$, then we define the succes probability as:

$$P_i^c = \bar{P}_i^{b,n} \bar{P}_i^a (\bar{P}_i^h + P_i^h P').$$

where

$$P' = \bar{P}_i^{b,h} \bar{P}_i^a (\bar{P}_i^h + P_i^h P') = \bar{P}_i^{b,h} \bar{P}_i^a \bar{P}_i^h + \bar{P}_i^{b,h} \bar{P}_i^a P_i^h P'.$$

$$P' = \frac{\bar{P}_i^{b,h} \bar{P}_i^a \bar{P}_i^h}{1 - \bar{P}_i^{b,h} \bar{P}_i^a P_i^h}.$$

and finally

$$\begin{aligned} P_i^c =& \bar{P}_i^{b,n} \bar{P}_i^a \bar{P}_i^h + \frac{\bar{P}_i^{b,h} \bar{P}_i^a \bar{P}_i^h}{1 - \bar{P}_i^{b,h} \bar{P}_i^a P_i^h} P_i^h \bar{P}_i^{b,n} \bar{P}_i^a \\ =& \bar{P}_i^{b,n} \bar{P}_i^a \bar{P}_i^h \left(1 + \frac{\bar{P}_i^{b,h} \bar{P}_i^a P_i^h}{1 - \bar{P}_i^{b,h} \bar{P}_i^a P_i^h}\right) \\ =& \frac{(1 - P_i^{b,n})(1 - P_i^a)(1 - P_i^h)}{1 - (1 - P_i^h)(1 - P_i^a) P_i^h}. \end{aligned}$$

The handover probability P_i^h cannot be calculated in a simple way since the service time distribution is not exponential (see expresion (1)) but a phase type distribution [8]. A phase type distribution is the distribution until absorption in an absorbing Markov process. We can represent a phase type distribution by a pair $(\boldsymbol{\alpha}, \boldsymbol{S})$ where the $n \times n$ matrix $\boldsymbol{S}$ delineates the transition rates between the transient states in the absorbing Markov process, and the $1 \times n$ vector $\boldsymbol{\alpha}$ the probabilities that the process is started in state i. Remember that transition rates between the transient states and the absorbing state $n+1$ defined by the $n \times 1$ vector $\boldsymbol{\tau}$ satisfy $\boldsymbol{\tau} = -\boldsymbol{S}\boldsymbol{e}$, where $\boldsymbol{e}$ is defined to be a $n \times 1$ dimensioned column vector of 1s.

Therefore, it is necessary to work out the parameters that define the phase type distribution that models the service time distribution in the system under study. For that the initial birth and death process, whose system state is described by $\boldsymbol{x}$, is modified to accomodate the phase type distribution. A new absorbing Markov process is defined for each type of service i. Then, to define the states of the absorbing Markov process, we keep track of the path of a labeled session of class i through the states of the initial birth and death process until a call termination or an abandonment occurs. Therefore, in those new Markov processes a new absorbing state is added which represents the end of the service and the states that have none sessions of type i are removed since in the system there is always at least one session of type i which we are keeping track of.

To clarify this point we display an example of a system with only $N = N_e = 2$ elastic services. In Fig. 1 the transition rates between states for the initial birth and death process are shown, where $i = 0, \ldots, n_M$ and $j = 0, \ldots, n_M$. Notice that for states with $i = 0$ or $j = 0$ the transition to $i-1$ or $j-1$ respectively does not exist. The notation has been simplified as $a(\boldsymbol{x}) = a$, $\beta(\boldsymbol{x}) = \beta$ and $\mu(\boldsymbol{x}) = \mu$. In Figure 2 the transition rates for the absorving Markow process are shown when it is modeled the service time distribution for service 1. The state A is the absorbing state and then $i = 1, \ldots, n_M$ and $j = 0, \ldots, n_M$.

The vector $\boldsymbol{\alpha}_i$ is calculated from the steady state probabilities in the initial birth and death process. But as there are states that do not exist in the absorbing Markow process for class i, those who have $i = 0$, these probabilities are removed and the vector is normalized. The matrix $\boldsymbol{S}_i$ is the generator matrix of the absorbing Markov process for class i.

Then, for service i we have a exponential distribution for cell residence time with rate $\mu_i^{r,e}$ and a phase type distribution with paramenters $(\boldsymbol{\alpha}_i, \boldsymbol{S}_i)$ for service time distribution. The probability density functions are respectively $f_r(t) = \mu_i^{r,e} e^{-\mu_i^{r,e} t}$ and $f_c(t) = \boldsymbol{\alpha}_i e^{t\boldsymbol{S}_i}\boldsymbol{\tau}_i$ and the distribution functions are $F_r(t) = 1 - e^{-\mu_i^{r,e} t}$ and $F_c(t) = 1 - \boldsymbol{\alpha}_i e^{t\boldsymbol{S}_i}\boldsymbol{\tau}_i$.

Therefore, the probability that the service time T_c finishes before than the residence time T_r, i.e. a handover does not occur, is:

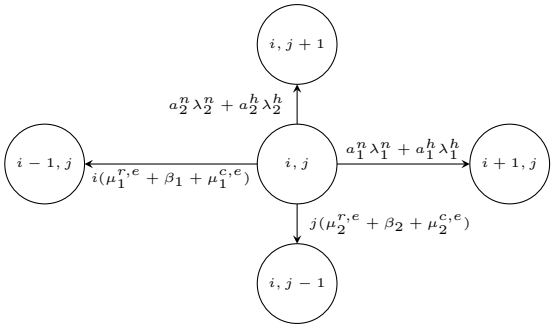

Figure 1. Transitions between states in the initial Markov process.

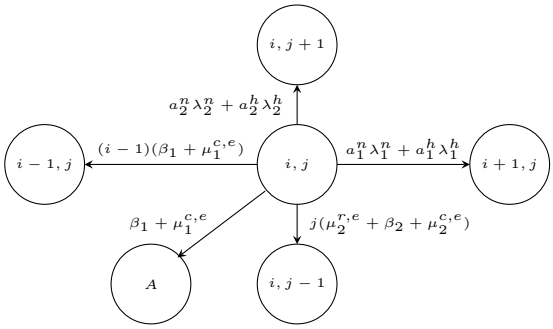

Figure 2. Transition between states in the absorbing Markow process for service 1.

$$
\begin{aligned}
1 - P_i^h = P(T_c < T_r) &= \int_0^\infty \int_{t_c}^\infty f(t_r, t_c) dt_r dt_c = \\
&= \int_0^\infty f_c(t_c) \int_{t_c}^\infty f_r(t_r) dt_r dt_c = \\
&= \int_0^\infty F_r^c(t_c) f_c(t_c) dt_c = \int_0^\infty e^{-\mu_i^{r,e} t} \boldsymbol{\alpha}_i e^{t\boldsymbol{S}_i} \boldsymbol{\tau}_i dt_c = \\
&= \boldsymbol{\alpha}_i (\mu_i^{r,e} I - \boldsymbol{S}_i)^{-1} \boldsymbol{\tau}_i.
\end{aligned}
$$

III. Design of VP parameters

In this section the performance of the VP policy dealing with streaming and elastic traffic is studied. We first define how VP operates, and next VP parameters are designed in a simple way considering a static system where mobility does not exist.

At the time of design, each streaming class is allocated a nominal capacity C_i, where $\sum_{i=1}^N C_i \geq C$ and each elastic class a nominal number of flows n_i^m that will be detailed below. Streaming classes that are using less than their C_i and elastic classes that have less than their n_i^m flows in the system are given higher priority. While all the classes are underloaded, the resources are shared without any restriction, but when a class is overloaded it is forced to back off if an underloaded class needs its allocated resources.

Let $\boldsymbol{x}'$ represent the state that will achieve the system if an arrival of type i is accepted. Then, with streaming traffic VP takes the following decisions: a session is accepted if $\sum_{i=1}^N b_i x_i + b_i \leq C - t_i^s(x_i)$ and the new service rates for all elastic classes are higher than the minimum after accepting the new session, i.e. $\mu_j^{c,e}(\boldsymbol{x}') \geq b^m/L, j = 1+N_s, \ldots, N$ and

rejected otherwise. The parameter $t_i^s(x_i)$ of service i may be interpreted as the parameter that introduces the trunk reservation mechanism but in this case this parameter is dynamic. $t_i^s(x_i) = s_i^s$ when $b_i x_i + b_i \leq C_i$ and $t_i^s(x_i) = t_i^s$ when $b_i x_i + b_i > C_i$, where $s_i^s \leq t_i^s$. Henceforth we refer to the VP policy with streaming traffic as VPS. With elastic traffic, VP takes the following decisions: The flow of class $i = 1 + N_s, \ldots, N$ is accepted if $\mu_j^{c,e}(\boldsymbol{x}') \geq b^m/L + t_i^e(x_i), j = 1 + N_s, \ldots, N$ and rejected otherwise. The parameters are $t_i^e(x_i) = s_i^e$ when $x_i + 1 \leq n_i^m$ and $t_i^e(x_i) = t_i^e$ when $x_i + 1 > n_i^m$, where $s_i^e \leq t_i^e$. Henceforth we refer to the VP policy with elastic traffic as VPE. From now on, to simplify the design we consider $s_i^s = s_i^e = 0 \ \forall i$ [3].

The nominal values C_i or n_i^m, depending on whether the class is streaming or elastic, allocated to each class are calculated considering that each class is isolated from the others in the system. Thus, given the parameters of the system, we search a minimum value of C_i or n_i^m so that $P_i^{b,(n,h)} \leq B_i^{b,(n,h)}$ or $P_i^c \geq B_i^c$ respectively. The minimum capacity C_i for the isolated service i is calculated using a binary search in one dimension until the QoS requirements are fullfilled. The nominal values n_i^m are the maximum number of flows of elastic class i that can be served at b_i^M in the isolated system: $n_i^m = \lfloor C_i/b_i^M \rfloor$. After that, all classes are considered in a aggregated system where the total capacity will be $C = \sum_i^N C_i$.

The trunk reservation levels t_i^s and t_i^e of service i need to be determined. Our goal is to find values of t_i^s and t_i^e in a simple manner and low computational cost. Thus, we have studied a system model without mobility. We study the blocking probabilities $P_i^{b,(n,h)}$ or the success completion probabilities P_i^c of each service, depending on whether the service is streaming or elastic, when the other services are overloaded. We search a set of t-parameters that provides a trade-off between robustness against overloads and efficiency. If the $t_i^{s,e}$ are high the VP policy tends to a Complete Partitioning (CP) policy, and although it presents high robustness, when the overall traffic is light the resources are underutilized since each service has its nominal allocation and share a low amount of resources. If the $t_i^{s,e}$ are low, it is the opposite case, it tends to a Complete Sharing (CS) policy and when some services are overloaded can overwhelm all the others since a high amount of resources are shared without restrictions. To cope with that, the behavior of the system is studied by varying λ_i, μ_i^r, μ_i^c and b_i for streaming classes and λ_i, b_i^M and b^m for elastic classes. By using heuristics, the chosen expression for parameter t_i^s and t_i^e are:

$$t_i^s = \sqrt{\frac{3}{2}C}\frac{1}{C_i\mu_i}\sum_{j\neq i}\lambda_j b_j \quad ; \quad t_i^e = 2\frac{b^m}{L}\sqrt{\frac{Cf_i}{b_i^M n_M}}. \qquad (2)$$

Notice that t_i^s is expresed in amount of units of resources and t_i^e in flows/s.

IV. VP in Multiservice Cellular Networks

The operation of VP has been described in networks without mobility and determined an expresion for its parameters $t_i^{s,e}$. Now, we consider the problem of extending a robust and efficient VP scheme applied to cellular networks.

The system can handle streaming and elastic traffic, moreover for each class new and handover arrivals are distinguished, thus we have to decide whether to combine new and handover sessions in a unique flow class or not. The failure of a handover session is highly undesirable but reserving channels for handover traffic could increase blocking probabilities for new requests. Hence, for streaming traffic a trade-off between the two QoS measures is needed. As new and handover arrivals of the same service have different QoS requirements ($B_i^{b,h} << B_i^{b,n}$), aggregating both type of arrivals into the same flow would be highly inefficient. But at the same time they cannot be managed as independent flows since λ_i^n and λ_i^h are related and undergo the same overloads. However, for elastic traffic, new and handover arrivals are considered as an unique flow since the abandonment probability does not depend on whether the flow was arrived at the system as new or handover request.

We propose a new VP scheme based on a combination of VPS, VPE and FGC. From now on we refer to it as VPC. The nominal capacity C_i and nominal flows n_i^m allocated to each class are calculated by isolating each class. For streaming classes, we consider N_s single service scenarios and for each of them, new and handover request are distinguished. Hence, a SAC is needed to provide the QoS requirements of these different type of arrivals. The SAC chosen is FGC. In this case handover requests are always admitted and new requests have an associated parameter $h_i \in \mathbb{R}$ that controls their acceptance. Considering that x_i is the amount of resources occupied in the system before the arrival of the new request, the following decisions can be taken. If $b_i x_i + b_i \leq \lfloor h_i \rfloor$, accept; if $b_i x_i + b_i = \lfloor h_i \rfloor + 1$, accept with probability $h_i - \lfloor h_i \rfloor$; if $b_i x_i + b_i > \lfloor h_i \rfloor + 1$, reject.

We consider all classes in an aggregated system with total resource units $C = \sum_{i=1}^{N_s} C_i$.

The VPC policy takes differents decisions depending on whether the arrival is either streaming or elastic traffic. If the arrival is a streaming class, the complete scheme VPC is represented in Fig. 3 and works as follows: handover arrivals of classes that do not fulfill VPS restriccions are rejected; otherwise they are accepted. New arrivals of classes that do not fulfill VPS restrictions are rejected. Note that for new arrivals we add $d_i = C_i - h_i$ to the VPS parameter t_i^s. If the new arrival passes VPS restrictions the system verifies the following conditions:

- According to the VP policy if the class i is using more than its nominal capacity (C_i), we are in the case where $\sum_{i=1}^N b_i x_i + b_i \leq C - (t_i^s + d_i)$. We say that we are in condition Cd_1. Therefore, the system has enough resources as the overall traffic in the system is light and so, the new arrivals of class i is accepted. This last decision is the reason why d_i is added. In this case, accepting all the new arrivals could be harmful for handover arrivals of the same class and hence, d_i resources are reserved for handover arrivals of service i.
- If this class is using less than its nominal capacity (C_i) we say that we are in condition Cd_2 and FGC policy

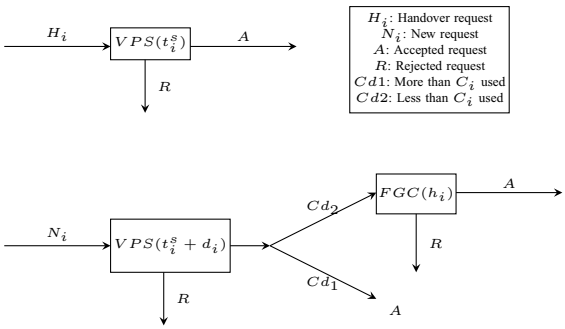

Figure 3. Acces control for VPS applied to cellular networks for streaming traffic.

is applied with the corresponding parameter h_i, i.e. if $[b_i n_i + b_i \leq C_i]$ and $[\sum_{i=1}^{N} b_i x_i + b_i \leq C]$ we apply FGC policy to decide on the acceptance of the new arrival of service i.

The values t_i^s are calculated with Eq. (2) where $\lambda_j = \lambda_j^n + \lambda_j^h$. If the arrival is of an elastic class, VPE is applied considering new and handover arrivals as a unique flow.

V. Analysis and Comparison between VP and other Policies

In this section, the performance of VPC is compared with that of the Multiple Fractional Guard Channel (MFGC) [9], an extension to the multiservice paradigm of the FGC. The MFGC policy has been chosen because its flexibility and because it has been well studied in the literature.

In MFGC the policy parameters t_i^n (t_i^h) control the amount of system resources that each new (handover) session type i can access. The policy MFGC is designed to maximize the offered traffic that the system can handle while meeting the QoS requirements, but that design requires high precision and its computational cost can be prohibitive for some practical systems. As it was explained in section III, VPC policy that we propose is configurated with low computational cost.

To deal with elastic traffic MFGC has to be redefined. We propose the following points: the policy parameters t_i^n (t_i^h) control the amount of system resources that each new (handover) streaming session type i can access as it is pointed in [10]. The elastic traffic has a CS policy and use the resources that streaming sessions do not use. Upon the arrival of a streaming or elastic flow it has to be checked that all flows of elastic traffic still receive a service rate higher than the minimum if the request is accepted, otherwise the request is rejected.

A. Overload

The behavior of VPC and MFGC policies is compared in a system with streaming and elastic traffic when the system is overloaded with different degrees of overload, considering that when a class i is overloaded, both arrival rates for new and handover arrivals are overloaded in the same degree. For service i, the overload is defined as the percentage of the sum of the arrival rates of services $j \neq i$ that exceed the forecasts.

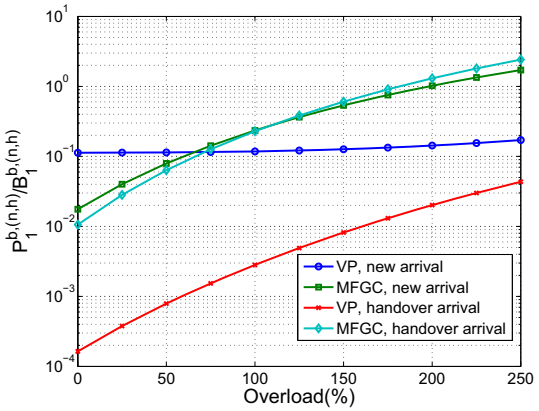

Figure 4. Ratios for streaming service 1

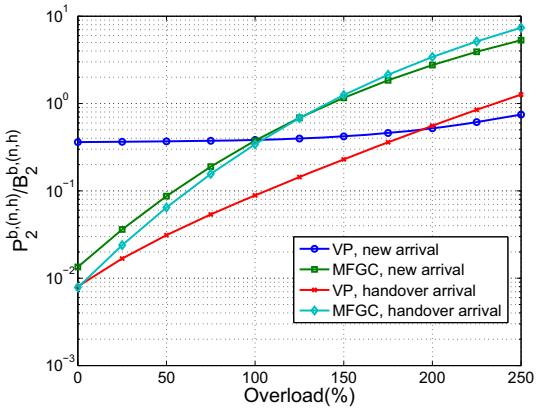

Figure 5. Ratios for streaming service 2

For the numerical examples we consider a system with 3 classes $N = 3$, where 2 classes carry streaming traffic $N_s = 2$ and 1 carries elastic traffic $N_e = 1$. The parameters are $\lambda_T = 3$, $f = [0.6, 0.3, 0.1]$, $b = [1, 2]$, $u^c = [0.5, 1]$, $u^r = [0.5, 0.5]$, $\mu^{r,e} = [0.25]$, $R = 100$, $b^m = 200$, $b^M = [500]$, $L = 100$, $\beta_0 = [0]$, $\beta_1 = [1]$, $\lambda_3^h = 0.5\lambda_3^n$, $B^{b,n} = [0.02, 0.01]$, $B^{b,h} = [0.004, 0.002]$ and $B^c = [0.99]$. The obtained nominal capacities needed to fulfill objectives for the scheme based on VP yields $C_i = [10, 10]$, $C_i^e = 800$ and therefore $C = 20$. The obtained parameter setting of the scheme based on VP policy has these values for FGC parameters: $h_1 = 8.2598$ and $h_2 = 9.1879$, and for VP parameters: $t_1^s = 1.6816$, $t_2^s = 1.4414$ and $t_1^e = 1.2649$. The parameter setting of MFGC policy is determined to be optimal. For a total capacity $C = 20$ the optimal configuration is $t^n = [15.5339, 17.3973]$ and $t^h = [17.2266, 18.9688]$.These results have been obtained using the algorithm proposed in [10].

In Fig. 4, it is shown the ratio of blocking probabilities of service 1 for new and handover arrivals achieved to the objectives ($P_1^{n,h}/B_1^{n,h}$) when services 2 and 3 are overloaded every one with the same overload and between 0% and 250% degree of overload. In Fig. 5 it is shown the ratio of blocking probabilities of service 2 for new and handover arrivals achieved to the objectives ($P_2^{n,h}/B_2^{n,h}$) when services 1 and 3 are overloaded. The ratio is calculated for both policies, VPS and MFGC. And in Fig. (6) it is shown the

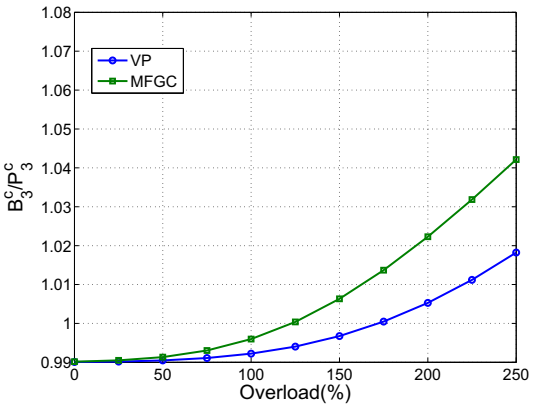

Figure 6. Ratios for elastic service 3

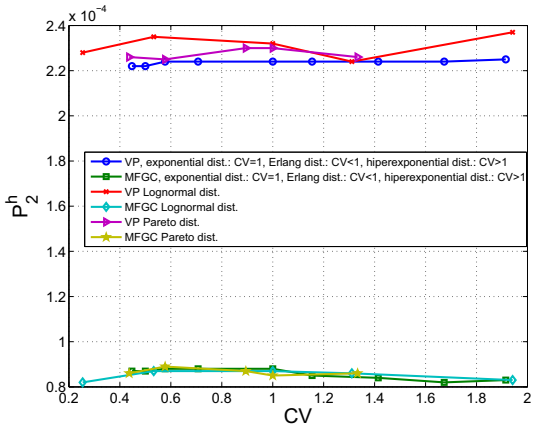

Figure 7. Sensitivity to resource holding time distribution for handover arrivals of service 2

ratio of success completion probabilities for service 3 when services 1 and 2 are overloaded.

Those figures confirm that the VPC is more robust than the MFGC Under overload conditions, the handover arrivals type of service 2 are the arrival types with worst behavior. For VPC, handover arrivals of service 2 have a better behavior than for MFGC policy. For low overload both ratios are lower than 1 and for high overloads VPC achieves lower ratios than MFGC policy. The VPC policy can support overloads of 230% fulfilling objectives, while MFGC policy can support 140% of overload. If we put our attention in overloads of 250% and handover arrivals of service 2, the achieved ratios for MFGC are 5.83 times the achieved ratios for VPC and for handover arrivals of service 1, the achieved ratios for MFGC are 55 times the achieved ratios for VPC. Other results not shown here, due to lack of space, lead to similar conclusions.

B. Sensitivity to resource holding time distribution

Up until now it has been assumed that the resource holding time is exponentially distributed. In this section we study by simulation in an Intel Core TM 2 Quad 4GB, other distributions such as the hiperexponential, the Erlang, the lognormal and the Pareto distribution to study the sensibility of both VPS and MFGC policies to resource holding time distribution. For that purpose we observe the behavior of the system varying the coefficient of variance (CV) of the distributions when all distributions have the same mean. The simulations have been made considering only streaming trafic and without overloads.

In Fig. 7 the blocking probabilities for handover requests for VPS and MFGC policies are shown for service 2. For a confidence level of 95%, the confidence intervals (CI) for each blocking probability, in the worst case is $\pm 10^{-5}$. As it can be seen, the results maintain a rather constant trend over the CV and hence over the resource holding time distribution.

VI. Conclusions

In this paper we have extended the VP policy to improve the behavior of multiservice cellular networks integrating both streaming and elastic traffic under overload conditions. The performance of VP for session admission control is studied and compared to multiple fractional guard channel (MFGC) policy. We have studied a new design method and a generalization of a mobile environment, the integration of streaming and elastic traffic in cellular systems, the sensitivity to the channel holding time distribution and we have presented a method to calculate the probability that a handover occurs for elastic flows.

The results show that VPC is more robust than MFGC. It was also studied the sensititity to resource holding time distribution showing that the behavior of the system does not depend on that distribution neither for MFGC policy nor for VP policy.

References

[1] T. Bonald and J. Roberts, "Congestion at flow level and the impact of user behaviour," *Computer Networks*, vol. 42, pp. 521–536, 2003.

[2] C. C. Beard and V. S. Frost, "Prioritized resource allocation for stressed networks," *IEEE/ACM Transactions on Networking*, vol. 9, no. 5, pp. 618–633, Oct. 2001.

[3] S. Borst and D. Mitra, "Virtual partitioning for robust resource sharing: computational techniques for heterogeneous traffic," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 5, pp. 668 – 678, Jun. 1998.

[4] R. Ramjee, R. Nagarajan, and D. Towsley, "On optimal call admission control in cellular networks," *Wireless Networks Journal (WINET)*, vol. 3, no. 1, pp. 29–41, 1997.

[5] D. García, J. Martínez, and V. Pla, "Admission control policies in multiservice cellular networks: Optimum configuration and sensitivity," *Lecture Notes in Computer Science*, vol. Wireless Systems and Mobility in Next Generation Internet, Gabriele Kotsis and Otto Spaniol (eds.), no. 3427, pp. 121–135, 2005.

[6] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Transactions on Vehicular Technology*, vol. VT-35, no. 3, pp. 77–92, Aug. 1986.

[7] T. Hwee-Pink, R. Nuñez-Quejia, A. F. Gabor, and O. J. Boxma, "Admission control for differentiated services in future generation CDMA networks," *Performance Evaluation*, vol. 66, no. 9-10, pp. 488–504, 2009.

[8] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM, 1999.

[9] H. Heredia-Ureta, F. A. Cruz-Pérez, and L. Ortigoza-Guerrero, "Multiple fractional channel reservation for optimum system capcity in multiservice cellular networks," *Electronics Letters*, vol. 39, no. 1, pp. 133–134, Jan. 2003.

[10] V. Pla, J. Martínez, and V. Casares-Giner, "Algorithmic computation of optimal capacity in multiservice mobile wireless networks," *IEICE Transactions on Communications*, vol. E88-B, no. 2, pp. 797–799, Feb. 2005.