Admission Control for Realtime Traffic: Improving Performance of Mobile Networks by Operating on Actual Throughput

Andreas Brandt*, Manfred Brandt†, Stefan Rugel‡ and Dietmar Weber‡

* Institut für Operations Research, Humboldt-Universität zu Berlin, Spandauer Str. 1, D-10178 Berlin, Germany,
Email: brandt@wiwi.hu-berlin.de

† Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB), Takustr. 7, D-14195 Berlin, Germany, Email: brandt@zib.de

‡ Siemens AG, Mobile Communication Networks, Hofmannstr. 51, D-81359 Munich, Germany
Email: stefan.rugel@siemens.com and dw.weber@siemens.com

***Abstract*— Deployment of realtime services will be important for the economic success of the 3GPP packet domain. Within this paper, a simple approach of admission control for realtime traffic is introduced, which takes its decisions on base of actual throughput. We present a fluid flow model, which allows to calculate the key performance indicators regarding packet loss, session request denial and plannable throughput. The formalism can be applied with good numerical accuracy and performance within the large range of traffic aggregation, which is characteristic for a mobile network. Exemplarily we apply the formalism to the traffic characteristics of Voice over IP and Push to Talk and show in principle, that our approach is promising in particular for Push to Talk with its high percentage of idle times.**

I. Introduction

From GPRS experience it comes more and more apparent that the offering of conversational and streaming services will be a crucial issue for the economic success of the 3GPP packet domain. New services will be deployed which may require realtime (RT) QoS, but differ from nowadays realtime applications as they are explicitly tailored to the prospects and limits of mobile networks. In particular, the bandwidth limitations of the radio access are decisive. On the one hand, this will drive technical development like data compression or the integration of WLAN hotspots into the 3GPP framework. On the other hand, this may favor such kind of services and applications, which do not use their full bandwidth all the time, like Push to Talk (PoC) which shows a considerably higher percentage of idle periods during the duration of a call than "classical" Voice over IP (VoIP) does. However, these characteristics can only be exploited if the admission control (AC) is made aware of the opportunity, that more sessions can be accepted than it seems to be possible when looking at the requested peak rates.

Our paper deals with this approach. Unlike previous investigations, which did a lot of detailed work on admission control for mobile networks, cf. [8] and the references herein, and also paid special attention to the 3GPP radio networks, cf. [12], we concentrate on the opportunity of overbooking. In order not to be limited by simulation time, we tried to develop a model which is simple enough to be tractable analytically, with a mathematical formalism being applicable to the wide range of traffic aggregation which is a main characteristic of mobile networks and extends from a few ten virtual sessions per cell up to some thousands of sessions which are handled in parallel by the components of the mobile core.

The paper is organized as follows: Section II gives an overview of our model from the technical point of view. Section III explains the mathematical framework, which is worth to take the main part of the document: even being more or less standard mathematics, the AC feedback mechanism itself and the large numbers of accepted sessions are highly non-trivial issues for numerical treatment. Section IV then applies the formalism to the problem under consideration, and we end up with the conclusion given in Section V.

II. Model Overview

The model under consideration here acts within the framework of the 3GPP end-to-end QoS concept, cf. [1] and [2]. As depicted in Fig. 1, the RT traffic rate is limited (by a token bucket meter) to a level somewhere below the maximum dynamic capacity, in order to guarantee e.g. maintenance traffic to be processed in time.

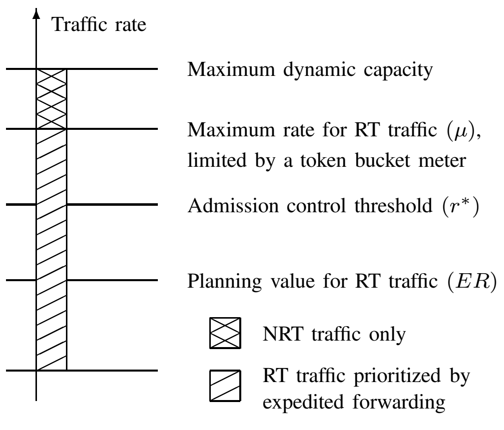

Fig. 1. Admission control embedded into RT traffic engineering.

We assume the following token bucket algorithm for the RT traffic rate: At constant rate μ tokens arrive at a token bucket of capacity b. Arriving tokens which find the bucket full get lost. An arriving packet is paired with a token from the token bucket – if there is any – and then served by the system. An arriving packet which cannot be paired with a token is discarded, i.e. gets lost.

Quite below the maximum rate μ for RT traffic, cf. Fig. 1, we find its planning value ER which results from the success ratio for incoming session requests. In between, we see the proposed admission control threshold r^* which is used to dynamically reject new incoming session requests if the current traffic rate is above r^*.

We consider an accepted incoming session request as a session which consists of subsequent on-off cycles generating packets at constant traffic rate r during the on periods, cf. Fig. 2. An incoming session request will be accepted if and only if the actual packet arrival rate, generated by the accepted session requests which are in an on period, i.e. active, is not larger than the above mentioned threshold r^*.

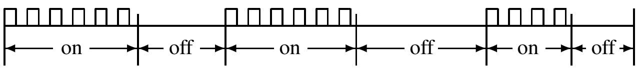

Fig. 2. Packet stream generated by an accepted session request consisting of three on-off cycles, i.e., the session request induces three on periods.

Interesting quantities for this system are: the probability $p_{\ell,r}$ that an incoming session request will not be accepted by the system, the mean rate ER of packets generated by session requests accepted by the system (planning value) and the packet loss probability $p_{\ell,p}$ due to an empty token bucket.

The engineering process, which is related to the technical approach described above, is built up in two steps:

1. r^* has to be determined on base of requirements for the packet loss rate.
2. ER has to be determined considering the required success rate for the incoming session requests.

For comparison, a simple static approach for AC only allocates packet or bit rate if it is available and has not been requested so far. So the maximum number of accepted sessions for the static case is μ/r. As with the dynamic approach, the mean number of accepted sessions and thus the mean traffic rate ER' results from the required success rate for the incoming session requests. The efficiency of the dynamic approach compared to static AC may be defined as $\omega := ER/ER'$. If $\omega > 1$, it pays for dynamic AC to exploit the fact that even realtime services tolerate packet loss below a given threshold. If $\omega < 1$, r^* is too low for an acceptable ER, in this case the static approach should be favored. A detailed quantitative approach will be given within Section III.

In general, the model calculations presented within the following section may be applied to both, bit and packet rate, as long as the fluid flow model is a good approximation. Within mobile networks, bandwidth and CPU capacity are dynamic resources to be protected by AC, the CPU capacity being mainly an issue for the core network due to the small size of the realtime packets.

III. Theoretical Calculation

For a mathematical treatment we assume that the incoming session requests, shortly called requests in the following, form a Poisson process of intensity λ. Further, assume that the on and off periods are exponentially distributed with mean $1/\alpha$ and $1/\beta$, respectively, and that the positive number of on-off cycles is geometrically distributed with parameter $q \in [0, 1)$.

Let $N_1(t)$ and $N_2(t)$ be the number of requests in the system at time $t + 0$ which are in an on and off period, respectively. The admission control strategy for the requests reads as follows: an arriving request at time t is accepted if and only if $N_1(t-0) \le n^*$, where $n^* := \lfloor r^*/r \rfloor$. The process $(N_1(t), N_2(t))$, $t \in \mathbb{R}$, is a Markov process with state space $\mathbb{Z}_+^2$ and corresponds to the dynamics of an infinite capacity two-node network with state-dependent arrival rate at the first node, cf. Fig. 3.

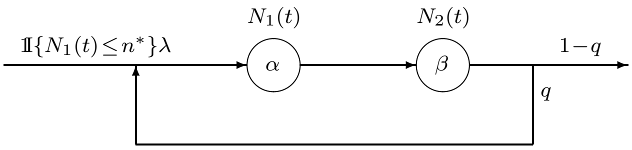

Fig. 3. Markov-modulated packet source as a two-node network.

In the following assume that $(N_1(t), N_2(t))$, $t \in \mathbb{R}$, is a stationary process. Having in mind real life applications where $EN_1(t) \approx EN_2(t) \approx 10^3$, the balance equations for the occupancy distribution in this network are too complex to be solved numerically. Therefore we are interested in approximations for the occupancy distribution $p(n) := P(N_1(t) = n)$ in Node 1 which can be computed much more efficiently.

A. Approximation by a binomial model

For the mean sojourn time EV of an accepted request in the system it holds

$$EV = 1/(\alpha(1-q)) + 1/(\beta(1-q)) . \tag{1}$$

We consider the approximate model where the sojourn time is exponentially distributed with parameter $\gamma := 1/EV$ and where any request in the two-node network is – when looking at an arbitrary time instant at the system – with probability

$$p := \beta/(\alpha+\beta) \tag{2}$$

in Node 1, i.e. active, and with probability $1 - p$ in Node 2.

Further, we approximate the dynamics of the total number $J(t)\,(= N_1(t) + N_2(t))$ of requests in the network at time $t + 0$ by a stationary birth-death process $J^{(1)}(t)$ with death rates $j\gamma$, $j \in \mathbb{Z}_+ \setminus \{0\}$, and birth rates

$$\lambda_j^{(1)} := \lambda \sum_{n=0}^{n^*} \binom{j}{n} p^n (1-p)^{j-n} , \quad j \in \mathbb{Z}_+ , \tag{3}$$

where we have in mind that in the approximate model a request is active with probability p and that hence the sum on the r.h.s. of (3) is the probability that an arriving request from outside will be accepted. Its distribution $\pi^{(1)}(j) := P(J^{(1)}(t) = j)$ is given by

$$\pi^{(1)}(j) = \pi^{(1)}(0) \prod_{\ell=1}^{j} \frac{\lambda_{\ell-1}^{(1)}}{\ell\gamma}, \quad j \in \mathbb{Z}_+ , \tag{4}$$

where $\pi^{(1)}(0)$ is determined by the normalizing condition. As with probability p a request is active in the approximate model, for the stationary distribution of the number of active requests we find

$$p^{(1)}(n) = \sum_{\ell=n}^{\infty} \pi^{(1)}(\ell) \binom{\ell}{n} p^n (1-p)^{\ell-n}, \quad n \in \mathbb{Z}_+ . \tag{5}$$

Note that in the limiting case $q \to 1$, where λ, $\alpha(1-q)$, $\beta(1-q)$, n^* are fixed, it holds: the sojourn time of an accepted request is exponentially distributed with parameter γ, $J(t)$ is a birth-death process with rates $\lambda_j^{(1)}$, $j\gamma$ and $p^{(1)}(n) = p(n)$, $n \in \mathbb{Z}_+$, cf. [7]. Hence the binomial model should provide a good approximation if the mean number of on periods generated by an accepted request is sufficiently large.

B. Approximation of the dynamics of the number of active requests by a birth-death process

For approximating the dynamics of $N_1(t)$ we define a stationary birth-death process $N_1^*(t)$ with birth rates λ_n^*, which have to be determined, and death rates[1] $\mu_n := n\alpha$ such that its distribution $p^*(n) := P(N_1^*(t) = n)$, $n \in \mathbb{Z}_+$, fits $p^{(1)}(n)$, $n \in \mathbb{Z}_+$. It follows

$$\lambda_n^* = (n+1)\alpha \frac{p^{(1)}(n+1)}{p^{(1)}(n)}, \quad n \in \mathbb{Z}_+ , \tag{6}$$

and after some algebra

$$\lambda_n^* = \frac{\lambda}{1-q} \frac{\sum\limits_{\ell=0}^{\infty} \Big(\prod\limits_{j=0}^{\ell+n} \frac{\lambda_j^{(1)}}{\lambda} \Big) \frac{1}{\ell!} \Big(\frac{\lambda}{\beta(1-q)} \Big)^{\ell}}{\sum\limits_{\ell=0}^{\infty} \Big(\prod\limits_{j=0}^{\ell+n-1} \frac{\lambda_j^{(1)}}{\lambda} \Big) \frac{1}{\ell!} \Big(\frac{\lambda}{\beta(1-q)} \Big)^{\ell}}, \quad n \in \mathbb{Z}_+ . \tag{7}$$

Because of (3), (7), it holds $0 < \lambda_n^* \le \lambda/(1-q)$, $n \in \mathbb{Z}_+$.

C. Modification of the birth-death process

Note that the birth-death process $N_1^*(t)$ only depends on the parameters $\lambda/(1-q)$, α, β and n^* while the Markov-modulated process $N_1(t)$ additionally depends on q. Thus modifying the birth rates λ_n^* of $N_1^*(t)$ by an additional parameter could improve the approximation. In the limiting cases $\beta \to 0$ and $\beta \to \infty$ the distribution of $N_1(t)$ corresponds to the distribution of the number of requests in an infinite server system with two arrival streams, where the first arrival stream in both cases is a state-dependent arrival process of intensity $\mathbb{I}\{n \le n^*\}\lambda$ and the second process is a Poisson process of intensity $\lambda(1-p_{\ell,r})q/(1-q)$ and a state-dependent arrival process of intensity $\mathbb{I}\{n \le n^*\}\lambda q/(1-q)$, respectively. Moreover, $N_1(t)$ is a birth-death process with birth rates $\mathbb{I}\{n \le n^*\}\lambda$ and death rates $n\alpha$ in case of $q = 0$. These observations suggest to replace the birth rates λ_n^* of $N_1^*(t)$ with

$$\lambda_n := \mathbb{I}\{n \le n^*\}\lambda + c\lambda_n^*, \quad n \in \mathbb{Z}_+ , \tag{8}$$

where $c \in \mathbb{R}_+$ is a parameter which has to be determined. Let $X(t)$ be the corresponding stationary birth-death process with birth rates (8) and death rates $\mu_n = n\alpha$, and let us denote by $p^{(a)}(n) := P(X(t) = n)$, $n \in \mathbb{Z}_+$, its distribution.

Since also for an approximation of $N_1(t)$ the rate of accepted requests should be equal to the rate of requests leaving the system, we claim for $X(t)$ the conservation law

$$\lambda \sum_{n=0}^{n^*} p^{(a)}(n) = \alpha(1-q) \sum_{n=1}^{\infty} n p^{(a)}(n), \tag{9}$$

which determines the parameter $c \in \mathbb{R}_+$ in (8) uniquely.

Lemma 3.1: Let λ, α, $\beta > 0$, $q \in [0, 1)$ and $n^* \in \mathbb{Z}_+$ be given. Then (9) has a uniquely determined solution $c \in \mathbb{R}_+$.

Proof: It follows easily that (9) is equivalent to

$$\begin{aligned} f(c) &:= \frac{\lambda}{1-q} \sum_{n=0}^{n^*} \frac{1}{n!\,\alpha^n} \prod_{j=0}^{n-1} (\lambda + c\lambda_j^*) \\ &= \sum_{n=0}^{\infty} \frac{1}{n!\,\alpha^n} \prod_{j=0}^{n} (\mathbb{I}\{j \le n^*\}\lambda + c\lambda_j^*) =: g(c) . \end{aligned} \tag{10}$$

As $f(0) \ge g(0)$ and $f(c) \le g(c)$ for sufficiently large c, there exists a solution of (9). In the following it will be shown that $g(c)/f(c)$ is strictly monotonically increasing for $c \in \mathbb{R}_+$, which provides the uniqueness. It holds

$$\frac{\lambda}{1-q} \frac{g(c)}{f(c)} - \alpha h(c) = \frac{\sum\limits_{n=n^*}^{\infty} \frac{1}{n!\,\alpha^n} \prod\limits_{j=n^*+1}^{n} (c\lambda_j^*)}{\sum\limits_{n=0}^{n^*} \frac{1}{n!\,\alpha^n} \prod\limits_{j=n}^{n^*} (\lambda + c\lambda_j^*)^{-1}}, \tag{11}$$

where

$$h(c) := \frac{\sum\limits_{n=0}^{n^*} \frac{n}{n!\,\alpha^n} \prod\limits_{j=0}^{n-1} (\lambda + c\lambda_j^*)}{\sum\limits_{n=0}^{n^*} \frac{1}{n!\,\alpha^n} \prod\limits_{j=0}^{n-1} (\lambda + c\lambda_j^*)}, \quad c \in \mathbb{R}_+ .$$

As the r.h.s. of (11) obviously is strictly monotonically increasing in $c \in \mathbb{R}_+$, it is sufficient to prove that $h'(c)$ is nonnegative. Using $\xi_j := \lambda + c\lambda_j^*$, after some algebra it follows

$$h'(c) = \Big(\sum_{0 \le m < n \le n^*} \frac{n-m}{n!\,\alpha^n\, m!\,\alpha^m} \Big(\prod_{j=0}^{n-1} \xi_j \Big) \Big(\prod_{j=0}^{m-1} \xi_j \Big) \Big(\sum_{j=m}^{n-1} \frac{\lambda_j^*}{\xi_j} \Big) \Big) \Big(\sum_{n=0}^{n^*} \frac{1}{n!\,\alpha^n} \prod_{j=0}^{n-1} \xi_j \Big)^{-2} \ge 0 .$$

□

[1] If n requests are active then $n\alpha$ is the rate that an on period will end.

As $p^{(a)}(n)$ should be an approximation of $p(n)$, we approximate $p_{\ell,r} = P(N_1(t) \geq n^*+1)$ and $ER = rEN_1(t)$ by $p_{\ell,r}^{(a)} := P(X(t) \geq n^*+1)$ and $ER^{(a)} := rEX(t)$, respectively. Note that the conservation law (9) yields

$$ER^{(a)} = r\lambda(1-p_{\ell,r}^{(a)})/(\alpha(1-q))\,. \tag{12}$$

The following theorem – the proof, cf. [7], is too long to be given here – tells us that in several limiting cases the proposed approximation $p^{(a)}(n)$, $n \in \mathbb{Z}_+$, of $p(n)$, $n \in \mathbb{Z}_+$, is exact. This gives some evidence that $p^{(a)}(n)$, $n \in \mathbb{Z}_+$, $p_{\ell,r}^{(a)}$ and $ER^{(a)}$ will provide good approximations for $p(n)$, $n \in \mathbb{Z}_+$, $p_{\ell,r}$ and ER, respectively, in the general case.

Theorem 3.1: In the limiting cases
(i) $q \to 1$, where λ, $\alpha(1-q)$, $\beta(1-q)$, n^* are fixed,
(ii) $q \to 0$, (iii) $\beta \to 0$, (iv) $\beta \to \infty$, (v) $n^* \to \infty$, where the remaining parameters are fixed, respectively,
it holds

$$p(n) = p^{(a)}(n)\,, \quad n \in \mathbb{Z}_+\,,$$

$$p_{\ell,r} = p_{\ell,r}^{(a)}\,, \quad ER = ER^{(a)}\,. \tag{13}$$

D. Approximation of the packet loss probability

According to the fluid flow approximation, arriving packets during an on period are approximated by a fluid with rate r and arriving tokens are approximated by a fluid with rate μ, i.e., the discrete nature of the packets and tokens is ignored.

The Markov process $(X(t), B(t))$, $t \in \mathbb{R}$, where $B(t)$ denotes the token bucket content at time $t+0$, is a birth-death fluid queue. Birth-death fluid queues with infinite buffer capacity are studied and surveyed in [9] Section 5. Having in mind the complexity of the birth rates λ_n, $n \in \mathbb{Z}_+$, for $X(t)$, which is already an approximation for $N_1(t)$, and that $EX(t) \approx EN_1(t) \approx 10^3$, already determining the zeros of the characteristic polynomial for the corresponding system of ordinary differential equations would lead to serious problems. In view of these difficulties, we are interested in approximations for the packet loss probability $p_{\ell,p}$.

Let $\tau := \mu/r\,(>0)$ be fixed and consider the stationary birth-death process $X(t)$, $t \in \mathbb{R}$, with birth rates λ_n and death rates μ_n approximating $N_1(t)$. For fixed $i \in \{\lceil\tau\rceil, \lceil\tau\rceil+1, \ldots\}$ let $X_i(t)$, $t \in \mathbb{R}$, be the birth-death process with distribution $P(X(t) \in (\cdot)\,|\,X(0-0) = i-1, X(0) = i)$, i.e., $X_i(t)$, $t \in \mathbb{R}$, has the Palm distribution of $P(X(t) \in (\cdot))$ with respect to the jump epochs from $i-1$ to i. Further, let T_i the first epoch after $t=0$ where $X_i(t)$ jumps from $\lceil\tau\rceil-1$ to $\lceil\tau\rceil$. The integrals

$$A_i := \int_0^{T_i} r(X_i(t)-\tau)_+\,\mathrm{d}t\,, \quad i \geq \lceil\tau\rceil\,, \tag{14}$$

correspond to the random amount of arriving fluid during $[0, T_i)$ which has to be paired from the token buffer. Let $A_{\lceil\tau\rceil-1} := 0$ and $A := A_{\lceil\tau\rceil}$ for notational convenience.

The intensity η of jumps of $X(t)$ from $\lceil\tau\rceil-1$ to $\lceil\tau\rceil$ is given by

$$\eta = \lambda_{\lceil\tau\rceil-1}p^{(a)}(\lceil\tau\rceil-1) = \mu_{\lceil\tau\rceil}p^{(a)}(\lceil\tau\rceil)\,. \tag{15}$$

Using the cycle formula for regenerative processes with respect to $r(X_{\lceil\tau\rceil}(t)-\tau)_+$, we obtain that

$$EA = E\,T_{\lceil\tau\rceil}\,E[r(X(t)-\tau)_+]\,, \tag{16}$$

where in view of (15) for the mean cycle length it holds

$$E\,T_{\lceil\tau\rceil} = \frac{1}{\eta} = \frac{1}{\mu_{\lceil\tau\rceil}p^{(a)}(\lceil\tau\rceil)}\,. \tag{17}$$

Consider the model with respect to the jump epochs from $\lceil\tau\rceil-1$ to $\lceil\tau\rceil$, i.e. $X_{\lceil\tau\rceil}(0-0) = \lceil\tau\rceil-1$, $X_{\lceil\tau\rceil}(0) = \lceil\tau\rceil$, and let B the random token bucket content at time $t=0$. Then $E[(A-B)_+]$ is the mean amount of packet fluid during $[0, T_{\lceil\tau\rceil})$ which gets lost. The mean total fluid EF of arriving packets during $[0, T_{\lceil\tau\rceil})$ is given by the cycle formula:

$$EF = E\,T_{\lceil\tau\rceil}\,E[rX(t)]\,. \tag{18}$$

The fraction $E[(A-B)_+]/EF$ is the packet loss probability within the birth-death process approximation of the fluid flow model. Assuming that at the beginning of the cycle at time $t=0$ the token bucket is full, we obtain the approximation

$$p_{\ell,p}^{(a)} := \frac{E[(A-b)_+]}{EF} = \frac{EA}{EF} - \frac{E[\min(A,b)]}{EF} \tag{19}$$

for the packet loss probability $p_{\ell,p}$.

By the Stone-Weierstrass theorem the function $\min(x,1)$, $x \in \mathbb{R}_+$, can be uniformly approximated by an exponential polynomial $\varphi(e^{-x}) = c_0+c_1e^{-x}+\ldots+c_me^{-mx}$ with arbitrary accuracy. Choosing $x = A/b$ provides the approximation

$$\min(A/b,1) \approx \sum_{k=0}^{m} c_k\,e^{-kA/b}\,.$$

In view of $E[e^{-kA/b}] = A^*(k/b)$, multiplying by b and taking expectation yields

$$E[\min(A,b)] \approx b\sum_{k=0}^{m} c_k A^*(k/b)\,. \tag{20}$$

Specifically, we have chosen

$$\varphi(y) := 1 - y^n\sum_{j=n}^{\infty}\binom{j-1}{n-1}\left(1-\sum_{i=n}^{j-1}\frac{1}{i}\right)_+(1-y)^{j-n} \tag{21}$$

for $n=8$, which is a polynomial of degree $m=20$. Note that $\varphi(e^{-x})$ converges uniformly on $\mathbb{R}_+$ to $\min(x,1)$ for $n \to \infty$, cf. [7].

E. Representation of $A^*(s)$ as continued fraction

Applying the law of total probability with respect to the first jump epoch of $X_i(t)$ after $t=0$ and the properties of birth-death processes provide the following renewal type equation for the LST $A_i^*(s) = E[e^{-sA_i}]$:

$$(\lambda_i+\mu_i+(i-\tau)rs)A_i^*(s) = \lambda_iA_{i+1}^*(s) + \mu_iA_{i-1}^*(s)\,,$$
$$s \in \mathbb{R}_+\,, \quad i \geq \lceil\tau\rceil\,. \tag{22}$$

Moreover, $A_{\lceil\tau\rceil-1} = 0$ yields

$$A_{\lceil\tau\rceil-1}^*(s) = 1\,, \quad s \in \mathbb{R}_+\,. \tag{23}$$

In the following let $\tau \in \mathbb{R}_+ \setminus \{0\}$, $s \in \mathbb{R}_+$ be fixed and

$$x_i := A_i^*(s)/A_{i-1}^*(s)\,, \quad i \geq \lceil \tau \rceil\,, \tag{24}$$

where because of (23) we have

$$x_{\lceil \tau \rceil} = A^*(s)\,. \tag{25}$$

Since A_i, $i \geq \lceil \tau \rceil - 1$, is a stochastically monotonically increasing sequence of r.v.'s, $A_i^*(s) = E[e^{-sA_i}]$ is a decreasing sequence, and hence it holds $x_i \in (0,1]$, $i \geq \lceil \tau \rceil$. Replacing in (22) $A_{i+1}^*(s)$ and $A_{i-1}^*(s)$ with $A_i^*(s)x_{i+1}$ and $A_i^*(s)/x_i$, respectively, cf. (24), we obtain

$$x_i = \frac{\mu_i}{\mu_i + (i-\tau)rs + \lambda_i(1-x_{i+1})}\,, \quad i \geq \lceil \tau \rceil\,. \tag{26}$$

Thus the sequence $(x_i)_{i \geq \lceil \tau \rceil}$ satisfies a continued fraction equation, and $A^*(s)$ is represented as a continued fraction because of (25). Since the r.h.s. of (26) defines a Lipschitz function of $x_{i+1} \in (0,1]$ into $(0,1]$ with Lipschitz constant λ_i/μ_i and as $\lim_{i\to\infty} \lambda_i/\mu_i = 0$, it follows that for any sufficiently large integer $n (> \lceil \tau \rceil)$, any $x_n \in (0,1]$ and any $s \in \mathbb{R}_+$ the iteration (26) defines a value $x_{\lceil \tau \rceil}$ which approximates $A^*(s)$ with arbitrary accuracy. Hence (16)–(21), (25), (26) provide an efficient algorithm for computing the approximation $p_{\ell,p}^{(a)}$ of the packet loss probability $p_{\ell,p}$

Remark 3.1: Note that the arguments used for our approximation of the performance measures remain valid for generally distributed off periods (with mean $1/\beta$).

The above algorithms for computing the approximate performance measures $p_{\ell,r}^{(a)}$, $ER^{(a)}$ and $p_{\ell,p}^{(a)}$ are implemented in a C-program. The computation is numerically stable over a rather wide range of parameters, which is not nearly reachable by an exact solution. For a set of parameters as used in the next section the computation takes only a few seconds on a 1 GHz P4 processor.

Simulation of the fluid flow model has shown that the performance measures $p_{\ell,r}$, ER and $p_{\ell,p}$ are well approximated for real life parameters, as expected in view of Theorem 3.1. For comparing performance measures of packet models obtained by simulation and solution of fluid flow models cf. [10].

IV. Results

The theoretical model sketched above will now be applied to two different kinds of realtime traffic models, the classical Voice over IP (VoIP) and Push to Talk (PoC) as a currently emerging application, cf. [3]. Table I shows their traffic parameters in so far as they are decisive for our investigations. The main difference among both models is the throughput utilization, which is defined here as the ratio of mean throughput and peak throughput of a single session. Even if PoC traffic statistics are not so well confirmed by practical experience as those for VoIP (for VoIP traffic parameters cf. e.g. [14]), there is some confidence that PoC, which is roughly speaking a way to use the mobile phone as a walky-talky, has longer silence periods due to its non-duplex mode of operation and due to blocking signaling periods during the communication phase.

TABLE I
Traffic Models.

	VoIP	**PoC**
Mean on time $(1/\alpha)$	2.5 s	6 s
Mean off time $(1/\beta)$	1.5 s	18 s
Mean number of on-off cycles $(1/(1-q))$	30	5
Throughput utilization $(p = (1/\alpha)/(1/\alpha+1/\beta))$	0.625	0.25
Mean call duration $(EV = (1/\alpha+1/\beta)/(1-q))$	120 s	120 s
Theoretical maximum efficiency $(1/p)$	1.6	4.0

With the traffic examples from above we now illustrate the engineering method which is sketched in Section II: first the proper admission control threshold r^* has to be found and second the planning value ER for the aggregated mean throughput has to be determined. This is done for given QoS requirements, i.e. the required values of the packet loss ratio, $p_{\ell,p}$, and $p_{\ell,r}$, the probability, that an incoming session request will not be accepted. For $p_{\ell,r}$, we choose a maximum value of 10^{-2}. In order to find a proper condition for $p_{\ell,p}$, we consider the asymptotic behavior of the packet loss rate with increasing rates of total requested load, as shown within Fig. 4. The requested load is given in multiples of μ and calculates to $\lambda r/(\mu\alpha(1-q))$. If not stated otherwise, calculations are done with the bucket size b set to zero in the fluid flow approximation $p_{\ell,p}^{(a)}$ of $p_{\ell,p}$ defined by (19).

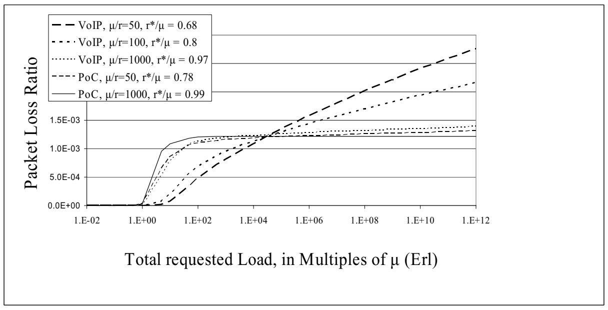

Fig. 4. Asymptotics of the packet loss ratio, $p_{\ell,p}$, as it depends on the rate of incoming session requests. The offered traffic rate is given in multiples of the maximum rate, μ, which is limited by the token bucket. The figures are shown for both kinds of traffic, VoIP and PoC, at different levels of aggregation $\tau = \mu/r$, where r is the packet rate of the individual sessions during an on phase. For each curve, the admission control threshold, r^*, is already within a reasonable range, i.e. with $p_{\ell,p}$ being of the order of 10^{-3} for very large values of the rate of incoming session requests.

The figure shows the efficiency of the dynamic approach of admission control: even at very high rates of incoming session requests – within Fig. 4 they extend up to values which would generate the 10^{12}-fold traffic with regard to the maximum rate μ – the packet loss $p_{\ell,p}$ does not increase considerably. For the sake of robustness of our dimensioning approach, we set r^* to values that limit $p_{\ell,p}$ to 10^{-3} and 10^{-4} at total requested traffic rates of the 10^4-fold of μ. Thus we expect the realtime traffic quality to be still acceptable even under extreme conditions. Please note, that for the rate of incoming session requests being within the planning region, $p_{\ell,p}$ will be well below 10^{-6}.

As shown in Fig. 5, r^* keeps nearly constant for very high values of the requested total traffic. The level of r^*/μ depends on traffic aggregation as expected, and differs considerably among both of the traffic models. This may be due to the differences regarding throughput utilization shown in Table I.

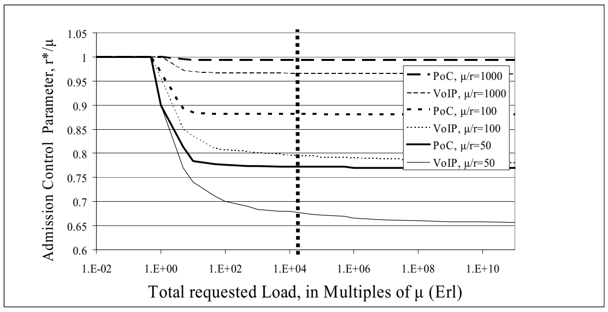

Fig. 5. Admission control threshold r^* for fixed $p_{\ell,p} = 10^{-3}$, as it depends on the total requested load, given in multiples of the requested traffic as in Fig. 4. Comparison between PoC (bold lines) and VoIP (thin lines) at different levels of aggregation $\tau = \mu/r$.

We therefore compare r^*/μ for both traffic models at the same level of τ/p, i.e. the aggregation divided by the throughput utilization given in Table I. The result depicted in Fig. 6 shows, that even now there is a big difference between VoIP and PoC, but in opposite to the results shown in Fig. 5, with VoIP higher values of r^*/μ are obtained than with PoC. Therefore we conclude at this point, that the admission control threshold r^*/μ depends in a more complex way on the parameters given in Table I, but leave a detailed evaluation for further studies.

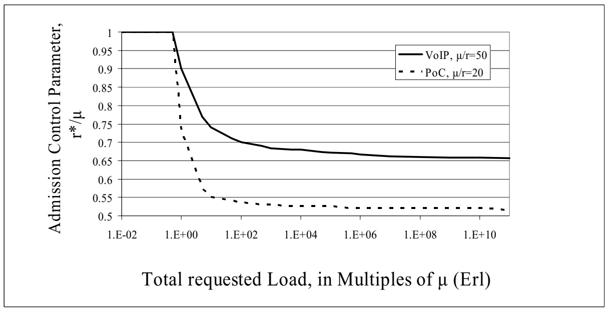

Fig. 6. Admission control threshold r^* for fixed $p_{\ell,p} = 10^{-3}$, depending on the total requested load (same as in Fig. 5). PoC and VoIP are compared at different levels of aggregation $\tau = \mu/r$, each. The comparison is done at equal aggregation divided by the throughput utilization given in Table I.

The effect of the bucket size b, as estimated within Section III/D, is depicted in Fig. 7 and shows a considerable reserve regarding packet loss.

The results for the admission control threshold are summed up within Fig. 8. As expected, PoC achieves higher values for r^*/μ than VoIP, because a PoC session transfers less data than VoIP, both within a certain time and in total.

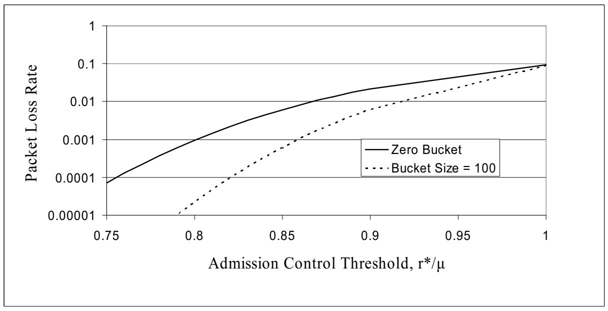

Fig. 7. The effect of bucket size b. The dependency of packet loss rate $p_{\ell,p}$ on r^* is shown for VoIP at fixed aggregation $\tau = \mu/r$.

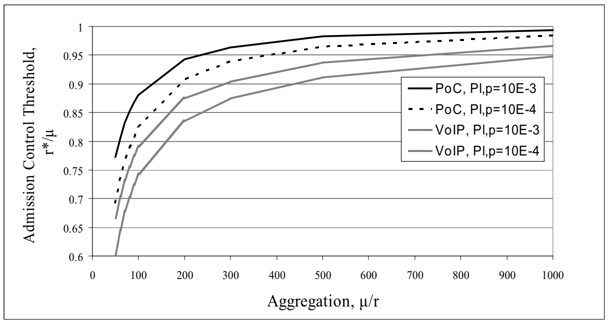

Fig. 8. Admission control threshold for VoIP and PoC as it depends on traffic aggregation $\tau = \mu/r$. The dependency is shown at different requirements for $p_{\ell,p}$: 10^{-3} and 10^{-4}, both given for a total requested traffic rate of 10^4 times τ, as shown in Fig. 5. Note, that $p_{\ell,p}$ is well below 10^{-6} if the requested traffic rates are within the regime of planned load.

We now consider the achieved mean rates ER and compare them to ER', cf. Section II. ER' is the accumulated mean rate related to the "static" approach, which only admits μ/r simultaneous sessions in maximum. ER' can be obtained from the total number of sessions μ/r and the requested session loss rate of 10^{-2} by the Erlang formula. The comparison is shown in Fig. 9 in the light of the example of VoIP. VoIP dynamic admission control pays at aggregation levels above 70. That is, if the dynamic resource under consideration is able to process the peak rates of 70 sessions in parallel, dynamic admission control performs better than the static approach. With increasing level of aggregation, ER tends to its maximum value which is given by the token rate μ, whereas the static approach is limited to $p\mu$, i.e. μ multiplied with the throughput utilization of $p = 0.625$ as given in Table I. Please note, that the packet loss rates listed in Fig. 9 only occur in very extreme cases, cf. Fig. 4. Under the conditions of planning load, packet loss is well below 10^{-6} in both cases. Fig. 8 compares the efficiency of the dynamic admission control for VoIP and PoC. Clearly, PoC is well above VoIP, as it is suggested by the theoretical maximum efficiency, which is shown within Table I as the reciprocal value of throughput utilization. Further it is evident, that VoIP traffic is not expected to be ruled efficiently by dynamic admission control. The dynamic approach however appears to be promising for services like PoC which show a low throughput utilization.

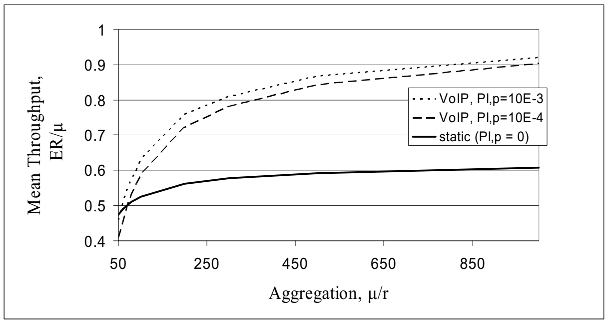

Fig. 9. Achieved mean throughput ER for VoIP as it depends on aggregation. The result is compared to what can be obtained by a static approach. The mean rates ER' for a static admission control is calculated according the Erlang formula.

Our results are summed up within Fig. 10. In contrast to VoIP, PoC shows an advantage of dynamic AC with regard to the static approach over the full range of aggregation which has been investigated. According to the throughput utilization of PoC, which is given within Table I, a static admission control can only use 25% of the given total throughput capacity μ. This is the main reason for the large enhancement in efficiency, which is obtained for PoC with dynamic admission control. As can be seen, the dynamic approach even pays considerably, if the admission control threshold r^* is set to a lower level by a conservative approach, which in order to be at the save side takes the worst case value (VoIP) of both kinds of traffic.

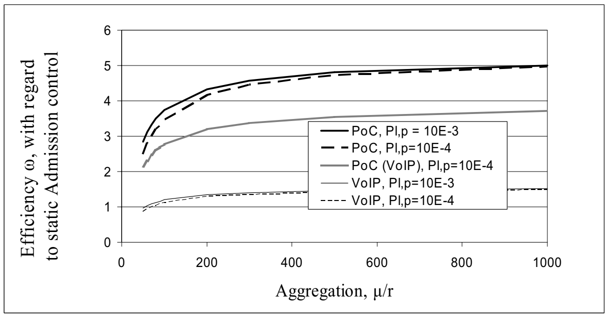

Fig. 10. Throughput efficiency $\omega := ER/ER'$ (i.e. mean throughput of dynamic admission control, compared with mean throughput of static admission control), as it depends on the aggregation level $\tau = \mu/r$. The figures are drawn for PoC and VoIP, based on the admission control thresholds shown within Fig. 8. With "PoC (VoIP)", a conservative engineering approach for PoC is depicted, which uses the values of r^*, which have been obtained for VoIP for a packet loss rate of 10^{-4}.

V. Conclusion

We have investigated a realization of the dynamic admission control for realtime traffic, which takes its decisions on base of the actual traffic rate. The details of the entire mechanism have been modeled analytically by a fluid flow model. The formalism, which calculates the probability for packet loss and admission denial for given traffic characteristics, provides good accuracy, even for aggregation levels of 10^4, and can therefore be applied to the entire mobile network from the radio access to the core. The findings of a first application to the traffic models of "VoIP" and "PoC" show the advantage of a dynamic admission control, which acts on actual throughput. It will pay, if idle periods take major parts of total RT session times. Even if setting the admission control threshold cautiously, the poor efficiency of static admission control will be enhanced considerably. Future studies have to be done on the detailed dependency of admission control performance on traffic model parameters, as well as the treatment of different packet or bit rates within the aggregated traffic stream.

References

[1] "Quality of service concept and architecture," *3GPP TS* 23.107.

[2] "End-to-end quality of service (QoS) concept and architecture," *3GPP TS* 23.207.

[3] "Technical specification; push-to-talk over cellular (PoC)," *PoC Release 2.0; Architecture V2.0.8 (2004–06).*

[4] I.J.B.F. Adan, E.A. van Doorn, J.A.C. Resing, and W.R.W. Scheinhardt, "Analysis of a single-server queue interacting with a fluid reservoir," *Queueing Syst.,* vol. 29, pp. 313–336, 1998.

[5] N. Akar and K. Sohraby, "Algorithmic solution of finite Markov fluid queues," in *Providing Quality of Service in Heterogeneous Environments, Proceedings of the 18th Int. Teletraffic Congress, Berlin, Germany, 2003,* J. Charzinski, R. Lehnert and P. Tran-Gia, Eds. Amsterdam: Elsevier Science, 2003, pp. 621–630.

[6] A. Brandt and M. Brandt, "On the distribution of the number of packets in the fluid flow approximation of packet arrival streams," *Queueing Syst.,* vol. 17, pp. 275–315, 1994.

[7] A. Brandt and M. Brandt. (2004, July) A birth-death approximation for a fluid source in a token bucket model. [Online]. Available: http://www.wiwi.hu-berlin.de/or/forschung/preprints.html

[8] S.K. Das, R. Jayaram, N.K. Kakani, and S.K. Sen, "A call admission and control scheme for quality-of-service (QoS) provisioning in next generation wireless networks," *Wireless Networks,* vol. 6, pp. 17–30, 2000.

[9] E.A. van Doorn and W.R.W. Scheinhardt, "Analysis of birth-death fluid queues," in *Proceedings of Applied Mathematics Workshop, Korea Advanced Institute of Science and Technology, Taejon, 1996,* B.D. Choi, Ed. pp. 13–29, 1996.

[10] A.I. Elwalid and D. Mitra, "Analysis and design of rate-based congestion control of high speed networks, I: Stochastic fluid models, access regulation," *Queueing Syst.,* vol. 9, pp. 29–63, 1991.

[11] R.B. Lenin and P.R. Parthasarathy, "A computational approach for fluid queues driven by truncated birth-death processes," *Methodol. Comput. Appl. Probab.,* vol. 2, pp. 373–392, 2000.

[12] C. Lindemann, M. Lohmann, and A. Thümmler, "A unified approach for improving QoS and provider revenue in 3G mobile networks," *Mobile Networks and Applications,* vol. 8, pp. 209–221, 2003.

[13] P.R. Parthasarathy, K.V. Vijayashree, and R.B. Lenin, "An M/M/1 driven fluid queue – continued fraction approach," *Queueing Syst.,* vol. 42, pp. 189–199, 2002.

[14] D.D. Trang, B. Sonkoly, and S. Molnár, "Fractal analysis and modelling of VoIP traffic," in *Proc. 11th International Telecommunications Network Strategy and Planning Symposium, July 2004.*