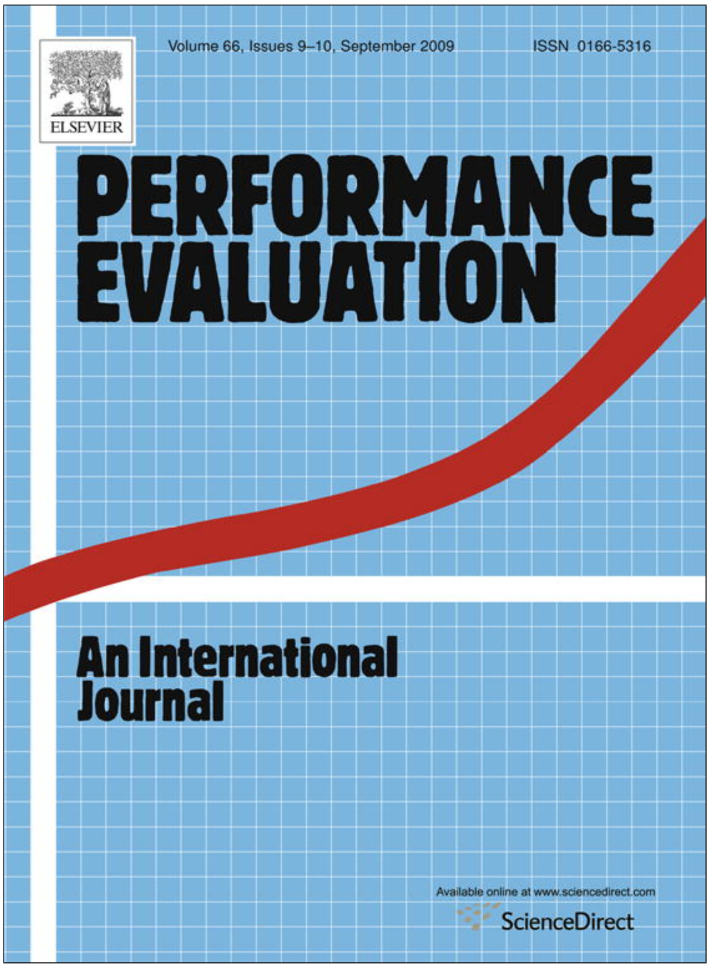
Volume 66, Issues 9–10, September 2009
ISSN 0166-5316
ELSEVIER
PERFORMANCE EVALUATION
An International Journal
Available online at www.sciencedirect.com
ScienceDirect

Performance Evaluation 66 (2009) 488–504

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Admission control for differentiated services in future generation CDMA networks

Hwee-Pink Tan [b,*], Rudesindo Núñez-Queija [c,1], Adriana F. Gabor [d], Onno J. Boxma [a,e]

[a] EURANDOM, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
[b] Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, Singapore 138632, Singapore
[c] CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
[d] Econometric Institute, Faculty of Economics, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
[e] Department of Mathematics and Computer Science, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands

ARTICLE INFO

Article history:
Received 3 April 2007
Received in revised form 28 June 2008
Accepted 15 March 2009
Available online 28 March 2009

Keywords:
Future generation wireless systems
CDMA
Differentiated admission control
Time-scale decomposition

ABSTRACT

Future Generation CDMA wireless systems, e.g., 3G, can simultaneously accommodate flow transmissions of users with widely heterogeneous applications. As radio resources are limited, we propose an admission control rule that protects users with stringent transmission bit-rate requirements ("streaming traffic") while offering sufficient capacity over longer time intervals to delay-tolerant users ("elastic traffic"). While our strategy may not satisfy classical notions of fairness, we aim to reduce congestion and increase overall throughput of elastic users. Using time-scale decomposition, we develop approximations to evaluate the performance of our differentiated admission control strategy to support integrated services with transmission bit-rate requirements in a realistic downlink transmission scenario for a single radio cell.

1. Introduction

Future Generation CDMA systems such as 3G are expected to support a large variety of applications, where the traffic they carry is commonly grouped into two broad categories. **Elastic traffic** corresponds to the transfer of digital documents (e.g., Web pages, emails and stored audio/videos) characterized by their size, i.e., the volume to be transferred. Applications carrying elastic traffic are flexible, or "elastic", towards transmission bit-rate fluctuations, the total transfer time being a typical performance measure. **Streaming traffic** corresponds to the real-time transfer of various signals (e.g., voice and streaming audio/video) characterized by their duration as well as their transmission bit-rate.

Stringent transmission bit-rate guarantees are necessary to ensure real-time communication to support applications carrying streaming traffic.[2] Consequently, the classical approach to resource sharing amongst *integrated* (elastic and streaming) traffic is to give head-of-line priority to packets of streaming traffic in order to offer packet delay and loss guarantees. Markovian models have been developed for the exact analysis of these systems [4,5]. However, they can be numerically cumbersome due to the inherently large dimensionality required to capture the diversity of user applications. Therefore, various approximations have been proposed [6,3], where closed-form limit results were obtained that can serve as performance bounds, and hence yield useful insight.

In this study, we consider downlink transmissions of integrated traffic in a single CDMA radio cell and propose an admission control strategy that allocates priority to streaming traffic through resource reservation while guaranteeing a

[*] Corresponding author.
E-mail addresses: hptan@i2r.a-star.edu.sg (H.-P. Tan), sindo@cwi.nl (R. Núñez-Queija), gabor@few.eur.nl (A.F. Gabor), boxma@win.tue.nl (O.J. Boxma).

[1] Present address: TNO Information and Communication Technology, The Netherlands.

[2] Streaming traffic with less stringent requirements, e.g., adaptive streaming traffic that is TCP-friendly and mimics elastic traffic, is considered in [1–3].

doi:10.1016/j.peva.2009.03.002

certain minimum transmission bit-rate requirements for all elastic users that share the remaining capacity equally. The location dependence of the wireless link capacity adds to the dimensionality problem already inherent in the performance analysis of corresponding *wireline* integrated services platforms.

We describe our system model in Section 2 and develop an approximation based on time-scale decomposition in Section 3 to evaluate the user-level performance. We define two base station models based on abstractions of the generic system model in Section 4 and present numerical results comparing both models in Section 5. Some concluding remarks are outlined in Section 6.

1.1. Related work

Various papers have been published recently that study communication links that carry integrated traffic:

- **Wired links**. In [6], an admission control policy is proposed which ensures *equal* blocking probabilities for streaming and elastic users. The thresholds used in the admission control are derived with the help of a fluid model. In [7], the impact on performance of streaming and elastic users is analyzed and the important issue of stability is raised. For the case of uniform stability (where the service rate for elastic users is higher than their arrival rate), by using time-scale decomposition, the authors propose bounds on the expected response time. Our analysis is largely motivated by Delcoigne et al. [7] and aims at incorporating more diversity of traffic classes, admission control rules and resource sharing strategies into the modeling framework.
- **Wireless links**. While a single class of elastic users is commonly assumed in wired links, the use of several classes of users seems more natural in *wireless* links, where geometry of the cell and interference play a major role. In [8], the integration of streaming and elastic traffic is analyzed for a time-slotted system with an admission control which ensures that the number of streaming users is not affected by the number of elastic users. For this model, good approximations based on time-scale decomposition are proposed. In [9,10], the complexity of the model is increased by taking into account the cell geometry and interference. The authors analyze several (fair) rate allocation schemes which lead to a feasible solution to the power control problem.

The sufficient conditions for decentralization proposed in [11,9] allow base stations to *independently* allocate transmission bit-rates among streaming and elastic users: If these conditions are satisfied (e.g., when all base stations transmit at a constant power), the use of a single-cell will be justified. Hence, our focus is on devising an allocation strategy that reserves capacity for streaming users while guaranteeing a certain minimum transmission bit-rate for all elastic users that share available capacity equally in a single CDMA cell. In a 3G radio system, this will lead to higher bit-rates for users near the base station. While our strategy may not satisfy common fairness criteria such as proportional-fairness and max–min fairness, intuitively, by analogy with opportunistic scheduling, it should result in reduced congestion (i.e., reduced blocking) and improved overall throughput for elastic users.

Our paper differs from [8] in that we account for interference and reserve a fixed capacity for streaming users. In our model, the number of streaming users is influenced by the number of elastic users present, which makes the analysis slightly more difficult. As compared to [6], we assume multiple classes of elastic users and account for interference between users. We approximate the model by using time-scale decompositions, in a similar way to [6–8].

2. System model

We consider a CDMA (e.g., UMTS/W-CDMA) radio cell with a single downlink channel whose transmission power at the base station (resource) is shared amongst users carrying streaming and elastic traffic. We assume that the base station transmits at full power, denoted by P, whenever there is at least one user in the cell. In addition, a part of the total power, $P_s \leq P$, is *statically* reserved for streaming traffic, where unclaimed power (subject to a maximum of $P_e = P - P_s$) is *equally* shared amongst all elastic users. Although in practice power may not be shared exactly equally, this assumption is reasonable when, for example, a Proportional Fair rate sharing mechanism is employed, cf. [12].

With W-CDMA technology, the base station can transmit to *multiple* users simultaneously using orthogonal code sequences. Let $P_u \leq P$ be the power transmitted to user u. The power received by user u is $P_u^r = P_u \Gamma_u$, where Γ_u denotes the attenuation due to path loss. For typical radio propagation models, Γ_u for user u at distance δ_u from its serving base station is proportional to $(\delta_u)^{-\gamma}$, where γ is a positive path-loss exponent.

As a measure of the quality of the received signal at user u, we consider *the energy-per-bit to noise-density ratio*, $\left(\frac{E_b}{N_0}\right)_u$, given by

$$\left(\frac{E_b}{N_0}\right)_u = \frac{W}{R_u} \frac{P_u^r}{\eta + I_u^a + I_u^r},$$

where W is the CDMA chip rate, R_u is the *instantaneous* data rate of user u, η is the background noise (assumed to be constant throughout the cell) and I_u^r is the inter-cell interference at user u caused by simultaneous *interfering* transmissions received at user u from base stations in *neighboring* cells. For linear and hexagonal networks, it can be shown [13] that I_u^r increases as δ_u increases. On the other hand, intra-cell interference, I_u^a, is due to simultaneous transmissions from the serving base station of user u using non-orthogonal codes (with total power P_u^a) to other users in the *same* cell received at user u. Quantitatively, we can write $I_u^a = \alpha P_u^a \Gamma_u$, where α is the code non-orthogonality factor.

To achieve a target error probability corresponding to a given Quality of Service (QoS), it is necessary that $\left(\frac{E_b}{N_0}\right)_u \geq \epsilon_u$, for some threshold ϵ_u. Equivalently, the data rate R_u of each admitted user u is upper-bounded as follows:

$$R_u \leq \frac{W P_u \Gamma_u}{\epsilon_u(\eta + \alpha P_u^a \Gamma_u + I_u^r)}. \tag{1}$$

Accordingly, for a given P_u, α and user type, the feasible transmission bit-rate of user u depends on its location (through Γ_u and I_u^r) and the intra-cell interference power, P_u^a.

2.1. Power control/allocation

According to Eq. (1), the transmission power, P_u, needed to support the transmission bit-rate requirement, r_u, of user u is given by:

$$P_u \geq \frac{r_u \epsilon_u [\alpha P_u^a \Gamma_u + \eta + I_u^r]}{W \Gamma_u} \equiv \tilde{P}_u. \tag{2}$$

Ideally, given perfect knowledge of the location of each user u at the base station, a maximum number of users can be admitted by allocating *exactly* $\tilde{P}_u$ to each user u. While this can be realised by users sending *power-up* or *power-down* signaling messages to the base station in response to *overly-strong* or *overly-weak* received signals, the actual power control is carried out in *discrete* steps, e.g., {0.5, 1, 1.5, 2} dB in UMTS [14].

For mathematical convenience, we manifest the discrete power control steps by dividing the cell into J disjoint segments, where J is chosen to adequately cover the dynamic range of the received power levels for a given step size. Hence, for a given dynamic range, a larger J corresponds to a smaller step size. The special cases of $J = 1$ ($J = \infty$) correspond to the scenario where power control is disabled or infeasible (perfect). We assume that the path loss, intra-cell and inter-cell interference are the same for any user in segment $j = 1, \dots, J$, denoted by (Γ_j, I_j^a, I_j^r), respectively.

Accordingly, we assume that elastic and streaming users arrive at segment j as independent Poisson processes at rates $\lambda_{j,e}$ and $\lambda_{j,s}$, with transmission bit-rate requirements of $r_{j,e} > 0$ and $r_{j,s} > 0$ respectively. Elastic users in segment j have a general file size (or service requirement) distribution with mean $f_{j,e}$ (bits) and, similarly, the holding times of streaming users may be taken to have mean $1/\mu_{j,s}$ (s). The total arrival rates of elastic and streaming users to the cell are denoted by $\lambda_e = \sum_{j=1}^J \lambda_{j,e}$ and $\lambda_s = \sum_{j=1}^J \lambda_{j,s}$. The minimum energy-to-noise ratio, ϵ_u, may depend on the user type and location [14], and will be denoted by $\epsilon_{j,e}$ and $\epsilon_{j,s}$ for elastic and streaming users in segment j, respectively.

2.2. Resource sharing

Given the transmission power, P_u, the mechanism via which the total power, P, is shared amongst all users (resource sharing) determines the total intra-cell interference power experienced at user u, P_u^a. When the base station transmits to all users in the cell simultaneously, each user u experiences the maximum intra-cell interference power, given by $P - P_u$; on the other hand, if time is slotted and the base station transmits only to one user in each time slot (*time sharing*), then there will be no interference power. Accordingly, we have the following expressions for P_u^a:

$$P_u^a \begin{cases} = P - P_u, & \text{simultaneous transmission to } \textit{all} \text{ users in the cell;} \\ < P - P_u, & \text{simultaneous transmission to } \textit{some} \text{ users in the cell;} \\ = 0, & \textit{no} \text{ simultaneous transmission (}\textit{time-sharing}\text{).} \end{cases}$$

2.3. Admission control

We propose an admission control strategy that ensures the required transmission bit-rate r_u of each admitted user u is satisfied. Let $N_{j,e}$ and $N_{j,s}$ denote the number of elastic and streaming users in segment j respectively, and define $N_j = N_{j,e} + N_{j,s}$. We further define the vectors $\mathbf{N}_e = (N_{1,e}, \dots, N_{J,e})$ and $\mathbf{N}_s = (N_{1,s}, \dots, N_{J,s})$ and let N_e and N_s be the total number of elastic and streaming users in the cell respectively. Let (β_j, γ_j) be the *minimum* transmission power required by an (elastic, streaming) user in segment j to sustain a transmission bit-rate requirement of $(r_{j,e}, r_{j,s})$, respectively. Depending on the resource sharing mechanism employed, (β_j, γ_j) can be evaluated using Eq. (2).

Provided there is sufficient capacity,[3] streaming users are always accommodated with exactly their required transmission bit-rate, consuming a total power of

$$P_s(\mathbf{N}_s) = \sum_{j=1}^J N_{j,s} \gamma_j.$$

[3] This, commonly referred to as the pole capacity of the cell, follows from the restrictions imposed in our admission control formulation.

The transmission bit-rate requirements of elastic users, on the other hand, must be achievable with power $P_e = P - P_s$. Since they receive an equal portion of the available power, we conclude that

$$N_e\beta_j \le P_e,$$

must hold for all j with $N_{j,e} > 0$, or equivalently,

$$N_e\beta_j \mathbf{1}_{(N_{j,e}>0)} \le P_e, \quad \forall j. \tag{3}$$

The indicator function $\mathbf{1}_E$ equals 1 if expression E holds and is 0 otherwise. Note that the J conditions in (3) only limit the *total* number of elastic users N_e, but that the maximum number of users does depend on the entire vector $\mathbf{N}_e$. Similarly, the fact that elastic users share power equally, together with the minimum power restrictions of both elastic and streaming users, implies that

$$N_e\beta_j \mathbf{1}_{(N_{j,e}>0)} + P_s(\mathbf{N}_s) \le P, \quad \forall j. \tag{4}$$

Conditions (3) and (4)[4] completely determine the admission policy: a newly-arrived user will be accepted only if the resulting system state, $(\mathbf{N}_e, \mathbf{N}_s)$, satisfies all $2J$ conditions.

Alternatively, these conditions may be formulated in terms of the *required power* for each user type. Similar to $P_s(\mathbf{N}_s)$, we determine the transmission power required by elastic users:

$$P_e(\mathbf{N}_e, \mathbf{N}_s) \equiv N_e \times \max_{j:N_{j,e}>0} \left\{\beta_j\right\}.$$

Note that this expression depends on the system state, $(\mathbf{N}_e, \mathbf{N}_s)$.

Our admission control policy for streaming users can now be formulated as follows: a newly-arrived streaming user in segment i will be admitted if

$$P_e(\mathbf{N}_e, \mathbf{N}_s + \mathbf{e}_i) + P_s(\mathbf{N}_s + \mathbf{e}_i) \le P,$$

where the vector $\mathbf{e}_i$ has its ith component equal to 1 and all other components are 0.

For elastic users, we must incorporate the power reservation restrictions as well. If we define

$$\overline{P}_s(\mathbf{N}_s) \equiv \max\left\{P_s, P_s(\mathbf{N}_s)\right\},$$

then a newly-arrived elastic user in segment i will be admitted if

$$P_e(\mathbf{N}_e + \mathbf{e}_i, \mathbf{N}_s) + \overline{P}_s(\mathbf{N}_s) \le P.$$

While the admission control proposed in [6] is similar, it results in equal blocking probabilities for both types of traffic. Due to resource reservation in our case, the blocking probabilities will depend on both the type and location of users.

2.4. Rate allocation

While streaming users are accommodated with exactly their required transmission bit-rate, i.e., $r_{j,s}$ in segment j, the transmission bit-rates allocated to elastic users depend on the number, type and location of other users. The available transmission power for elastic users is $P - P_s(\mathbf{N}_s)$, of which all active elastic users receive an equal portion regardless of their location. Using Eq. (1), an elastic user in segment j attains a transmission bit-rate

$$r_{j,e}(N_e, \mathbf{N}_s) = \frac{W\frac{P-P_s(\mathbf{N}_s)}{N_e}}{\epsilon_e[\alpha P^a_{j,e} + \frac{\eta+I^r_j}{\Gamma_j}]}, \tag{5}$$

where $P^a_{j,e}$ is the total intra-cell interference experienced by that user, which depends on the resource sharing mechanism. Accordingly, the *departure rate* of elastic users in segment j is given by:

$$\mu_{j,e}(\mathbf{N}_e, \mathbf{N}_s) = \frac{N_{j,e}r_{j,e}(N_e, \mathbf{N}_s)}{f_{j,e}}. \tag{6}$$

3. Analysis

Since exact analysis of our model is non-tractable in general and computationally involved when assuming exponentially distributed holding times and file sizes [4,5], we develop an approximation based on time-scale decomposition to evaluate the cell performance and to assess the accuracy through comparison with simulation. Our work is largely motivated by [7], where time-scale separation techniques were introduced for the analysis of integration of streaming and elastic traffic. The main goal in this section is to illustrate how the basic framework of [7] can be extended to cover various resource sharing strategies, admission control policies and a larger variety of user classes so as to capture the user heterogeneity exemplified in 3G wireless systems. In our discussion we explore the limits to such extensions if we wish to retain the desired tractability of their analysis.

[4] While this condition is pessimistic and may result in unnecessarily high blocking probability for elastic users, an admission policy that accounts for the location of elastic users would render the processor sharing model for elastic users intractable (the assumptions in [15] no longer hold). On the other hand, an overestimate of the power required for our admission control policy implies a better bit-rate, thus a better throughput for admitted elastic users.

3.1. Quasi-stationary approximation

We develop a quasi-stationary approximation for elastic users, to be denoted $\mathbf{A}(\mathbf{Q}, \mathbf{J})$, where we assume that the dynamics of streaming users take place on a much slower time scale than those of elastic users. More specifically, we assume that elastic traffic practically reaches statistical equilibrium while the number of active streaming calls remains unchanged, i.e., we assume that all $\mu_{j,s}$ and $\lambda_{j,s}$ are much smaller than any of the quantities $1/f_{j,e}$ and $\lambda_{j,e}$. This assumption is reasonable when we consider a combination of voice calls (streaming) and web-browsing or email (elastic) applications. Under this assumption, the dynamics of elastic users can be studied by fixing the number of streaming users in each segment, i.e., we fix the vector $\mathbf{N}_s \equiv \mathbf{n}_s$.

3.1.1. Conditional distribution for elastic traffic

We construct an approximation assuming that the number of active elastic users *instantaneously* reaches a new statistical equilibrium whenever $\mathbf{N}_s$ changes. For fixed $\mathbf{N}_s \equiv \mathbf{n}_s$, the elastic traffic behaves like a J-class $M/G/1$ processor sharing (PS) queue with admission control dictated by both (3) and (4). To avoid any confusion, we will append a superscript ^{Q} to all quantities (such as queue lengths and performance measures) resulting from this approximation.

For general service requirement distributions of elastic users and an admission region of the type $\sum_j N_{j,e}^Q \leq M$, the steady-state distribution of the number of jobs in each segment was shown to be a multivariate geometric distribution [15]. This can be shown to imply the same stationary distribution (up to a multiplicative constant) for the elastic users under the quasi-stationary assumption. For phase-type distributions, this can be proved formally by taking M large enough so that the set of allowable states (3) and (4) can be included. The joint process of queue lengths and service phases is reversible, so that state-space truncation does not destroy detailed balance and one can obtain the stationary distribution of the restricted process by renormalization of the steady-state measure:

$$\begin{aligned}\mathbb{P}^Q(\mathbf{n}_e|\mathbf{n}_s) &\equiv \mathbb{P}(\mathbf{N}_e^Q = \mathbf{n}_e \mid \mathbf{N}_s^Q = \mathbf{n}_s)\\ &= c_e^Q(\mathbf{n}_s) n_e! \prod_{j=1}^{J} \frac{\rho_{j,e}(\mathbf{n}_s)^{n_{j,e}}}{n_{j,e}!}, \end{aligned} \tag{7}$$

where we have defined $\rho_{j,e}(\mathbf{n}_s) = \frac{\lambda_{j,e}}{\mu_{j,e}(\mathbf{n}_s)}$ and the normalization constant $c_e^Q(\mathbf{n}_s)$ is such that summing (7) over all $\mathbf{n}_e$ that satisfy (3) and (4) gives a total of 1, for each fixed $\mathbf{n}_s$. We finally recall that $n_e = \sum_{j=1}^J n_{j,e}$.

The conditional acceptance probability of newly-arrived elastic users in segment i is

$$A_{i,e}^Q(\mathbf{n}_s) \equiv \mathbb{P}(P_e(\mathbf{N}_e^Q + \mathbf{e}_i, \mathbf{n}_s) \leq P - \overline{P}_s(\mathbf{n}_s) \mid \mathbf{N}_s^Q = \mathbf{n}_s).$$

From (7), we can also obtain the distribution of n_e by summing over all admitted combinations of $n_{j,e}$ such that $\sum_j n_{j,e} = n_e$.

3.1.2. Unconditional marginal distributions

Next, we consider the dynamics of streaming users. When $\mathbf{N}_s^Q = \mathbf{n}_s$, streaming users depart at a rate $\sum_j n_{j,s}\mu_{j,s}$. When a new streaming user arrives in segment i, due to admission control, it is either accepted or blocked. Under our approximation assumptions, the probability of acceptance in segment i, $A_{i,s}^Q(\mathbf{n}_s)$, is given by:

$$\mathbb{P}\left(P_e(\mathbf{N}_e^Q, \mathbf{n}_s + \mathbf{e}_i) \leq P - P_s(\mathbf{n}_s + \mathbf{e}_i) \mid \mathbf{N}_s^Q = \mathbf{n}_s\right).$$

Hence, the effective arrival rate of streaming users in segment i, $\Lambda_{i,s}^Q(\mathbf{n}_s)$, is given as follows:

$$\Lambda_{i,s}^Q(\mathbf{n}_s) = \lambda_{i,s} A_{i,s}^Q(\mathbf{n}_s).$$

As a side remark, note that $A_{i,s}^Q(\mathbf{n}_s) = 1$ if $P_s(\mathbf{n}_s + \mathbf{e}_i) \leq P_s$, since the admission control on elastic users ensures that $N_e^Q \beta_j \mathbf{1}_{(N_{j,e}>0)} \leq P - P_s$ for all j.

3.1.3. Evaluation of performance measures

We can now calculate several relevant performance measures by unconditioning on $\mathbf{N}_s^Q$. The unconditional distribution for the number of elastic users is

$$\mathbb{P}(\mathbf{N}_e^Q = \mathbf{n}_e) = \sum_{\mathbf{n}_s} \mathbb{P}^Q(\mathbf{n}_e \mid \mathbf{n}_s)\mathbb{P}(\mathbf{N}_s^Q = \mathbf{n}_s).$$

The unconditional blocking probabilities in segment i are

$$p_{i,s}^Q = \sum_{\mathbf{n}_s} (1 - A_{i,s}^Q(\mathbf{n}_s))\mathbb{P}(\mathbf{N}_s^Q = \mathbf{n}_s),$$

for streaming users; similarly, for elastic users, we have:

$$p_{i,e}^{Q} = \sum_{\mathbf{n}_s} (1 - A_{i,e}^{Q}(\mathbf{n}_s))\mathbb{P}(\mathbf{N}_s^{Q} = \mathbf{n}_s).$$

While the numerical evaluation of $\mathbb{P}^{Q}(\mathbf{n}_e \mid \mathbf{n}_s)$ and $\mathbb{P}(\mathbf{N}_s^{Q} = \mathbf{n}_s)$ is infeasible or cumbersome in general, we consider the following special cases where closed-form expressions exist:

- **Uniform admission control on elastic users**. For the special case where $\beta_i \equiv \beta$ for all i – we call this *uniform admission control*[5] – the distribution of n_e reduces to a simple truncated geometric distribution:

  $$\mathbb{P}(N_e^{Q} = n_e \mid \mathbf{N}_s^{Q} = \mathbf{n}_s) = \frac{\rho_e(n_s)^{n_e}(1 - \rho_e(\mathbf{n}_s))}{1 - \rho_e(\mathbf{n}_s)^{n_e^{Q,\max}(\mathbf{n}_s)}}, \tag{8}$$

  where $n_e^{Q,\max}(\mathbf{n}_s) = \lfloor (P - \overline{P}_s(\mathbf{n}_s))/\beta \rfloor$ and $\rho_e(\mathbf{n}_s) = \frac{\lambda_e}{\mu_e(\mathbf{n}_s)}$ is the total departure rate of elastic users from the cell.
- **Uniform admission control on streaming users**. Although we must assume exponential or phase-type holding time distributions and resort to standard methods to (numerically) solve the equilibrium distribution of $\mathbf{N}_s^{Q}$, the dimension of the finite-state Markov process $\mathbf{N}_s^{Q}$ is *much smaller* than that of the original process $(\mathbf{N}_e, \mathbf{N}_s)$: the component $\mathbf{N}_e$ is "eliminated" in the approximation.

  However, if we apply *uniform* admission control for streaming traffic by taking $\gamma_j \equiv \gamma$ independent of j (as above), then $A_{i,s}^{Q}(\mathbf{n}_s) \equiv A_s^{Q}(n_s)$ is independent of i and depends on $\mathbf{n}_s$ only through the total number of streaming users. $\mathbf{N}_s^{Q}$ can then be shown to be *balanced* [16] and can be reduced to the framework of [15]. It follows that, for *arbitrary* holding time distributions of streaming users, and $0 \le n_s \le n_s^{\max} = \lfloor \frac{P}{\gamma} \rfloor$:

  $$\mathbb{P}(\mathbf{N}_s^{Q} = \mathbf{n}_s) = c_s^{Q} \prod_{k=0}^{n_s-1} A_s^{Q}(k) \prod_{j=1}^{J} \frac{(\rho_{j,s})^{n_{j,s}}}{n_{j,s}!}, \tag{9}$$

  with $\rho_{j,s} = \lambda_{j,s}/\mu_{j,s}$ and $c_s^{Q} = P(N_s^{Q} = 0)$ can be determined by normalizing (9) to a probability distribution. Letting $\rho_s = \sum_j \rho_{j,s}$, we further obtain the distribution of the total number of active streaming users:

  $$\mathbb{P}(N_s^{Q} = n_s) = c_s^{Q} \frac{(\rho_s)^{n_s}}{n_s!} \prod_{k=0}^{n_s-1} A_s^{Q}(k),$$

  which in this case results again in a simple expression for the normalizing constant:

  $$c_s^{Q} = \left(\sum_{n_s=0}^{n_s^{\max}} \frac{(\rho_s)^{n_s}}{n_s!} \prod_{k=0}^{n_s-1} A_s^{Q}(k) \right)^{-1}.$$

We emphasize that, assuming quasi-stationarity, (7) and (8) are valid for general distributions of elastic users [15]. Note that these expressions are insensitive to the file size distributions, other than through their means. As a further remark, we observe that stability is of no concern in our model, since $\mathbf{N}_e^{Q}$ is bounded due to the assumption that $r_{j,e} > 0$. Often, when applying time-scale decomposition, the issue of stability is of considerable importance, giving rise to an additional assumption commonly referred to as *uniform stability* [7].

Remark 1. According to Eq. (6), the departure rate of elastic users depends on the system state, $(\mathbf{n}_e, \mathbf{n}_s)$. However, to apply Eqs. (7) and (8), the departure rate can depend on the system state through $\mathbf{n}_s$ only. We illustrate how this can be achieved with various resource sharing mechanisms in Section 4.

3.2. Fluid approximation

The fluid approximation (from the perspective of elastic users), denoted by $\mathbf{A}(\mathbf{F}, \mathbf{J})$, complements the quasi-stationary approximation: We now assume that the dynamics of elastic users are much slower than those of streaming users, i.e., the $\lambda_{j,s}$ and $\mu_{j,s}$ are much larger than the $\lambda_{j,e}$ and $1/f_{j,e}$. This assumption is valid when we consider the combination of voice calls (streaming) and large file transfer (elastic) applications. The dynamics of streaming users can then be studied by fixing the number of elastic users in each segment. This approximation will be reflected in the notations by adding a superscript ^{F}. Similar to $\mathbf{A}(\mathbf{Q}, \mathbf{J})$, we will construct an approximating $2J$-dimensional process under the assumption that $\mathbf{N}_s^{F}$ immediately reaches steady state, whenever $\mathbf{N}_e^{F}$ changes.

[5] With uniform admission control, the minimum required power is the same for all users, irrespective of their locations. As a consequence, the minimum rates are determined by the locations: users further away from the base station or with larger inter-cell interference must compromise for a lower rate. Thus, although the admission policy is the same, users in different segments are distinguished by the achievable rates (as well as their own traffic distributions).

3.2.1. Conditional distribution of streaming traffic

We fix the number of elastic users in each segment: $\mathbf{N}_e^F = \mathbf{n}_e$. Under the "fluid" approximation assumption, we can model the streaming users as a J-class Erlang-loss queue with finite capacity:

$$\begin{aligned}\mathbb{P}^F(\mathbf{n}_s \mid \mathbf{n}_e) &\equiv \mathbb{P}(\mathbf{N}_s^F = \mathbf{n}_s \mid \mathbf{N}_e^F = \mathbf{n}_e) \\ &= c_s^F(\mathbf{n}_e) \prod_{j=1}^{J} \frac{\rho_{j,s}^{n_{j,s}}}{n_{j,s}!},\end{aligned} \tag{10}$$

where $\rho_{j,s} = \frac{\lambda_{j,s}}{\mu_{j,s}}$. As before, we emphasize that the above expression depends on the holding time distribution only through its mean. The constant $c_s^F(\mathbf{n}_e)$ can again be determined by requiring that (10) adds up to 1 when summing (for fixed $\mathbf{n}_e$) over all $\mathbf{n}_s$ such that $P_e(\mathbf{n}_e, \mathbf{n}_s) + P_s(\mathbf{n}_s) \le P$.

3.2.2. Unconditional marginal distributions

Next, we consider the dynamics of elastic users. When $\mathbf{N}_e^F = \mathbf{n}_e > 0$, elastic users in segment j (if any) experience an average transmission bit-rate (recall that n_e is the sum over all components of the vector $\mathbf{n}_e$):

$$\begin{aligned}\bar{r}_{j,e}(\mathbf{n}_e) &\equiv \mathbb{E}[r_{j,e}(n_e, \mathbf{N}_s^F) \mid \mathbf{N}_e^F = \mathbf{n}_e] \\ &= \sum_{\mathbf{n}_s} r_{j,e}(n_e, \mathbf{n}_s)\, \mathbb{P}^F(\mathbf{n}_s \mid \mathbf{n}_e),\end{aligned} \tag{11}$$

where the summation is taken over all $\mathbf{n}_s$ for which $P_e(\mathbf{n}_e, \mathbf{n}_s) + P_s(\mathbf{n}_s) \le P$. The (state-dependent) departure rate of elastic users from segment j is

$$n_{j,e}\bar{r}_{j,e}(\mathbf{n}_e)/f_{j,e}.$$

In order to fully describe the dynamics of the elastic users, we now determine the arrival rate, which also depends on the state $\mathbf{n}_e$ because of the employed admission control. Under our approximation assumptions, the probability of acceptance in segment i is given by:

$$A_{i,e}^F(\mathbf{n}_e) \equiv \mathbb{P}(\bar{P}_s(\mathbf{N}_s^F) + \bar{P}_e(\mathbf{n}_e + \mathbf{e}_i, \mathbf{N}_s^F) \le P \mid \mathbf{N}_e^F = \mathbf{n}_e),$$

and, consequently, the effective arrival rate of elastic users in segment i is

$$\Lambda_{i,e}^F(\mathbf{n}_e) \equiv \lambda_{i,e} A_{i,e}^F(\mathbf{n}_e).$$

3.2.3. Evaluation of performance measures

We can now calculate the following unconditional performance measures:

$$\mathbb{P}(\mathbf{N}_s^F = \mathbf{n}_s) = \sum_{\mathbf{n}_e} \mathbb{P}^F(\mathbf{n}_s \mid \mathbf{n}_e)\mathbb{P}(\mathbf{N}_e^F = \mathbf{n}_e).$$

The unconditional blocking probabilities in segment i are

$$p_{i,e}^F = \sum_{\mathbf{n}_e} (1 - A_{i,e}^F(\mathbf{n}_e))\mathbb{P}(\mathbf{N}_e^F = \mathbf{n}_e),$$

and

$$p_{i,s}^F = \sum_{\mathbf{n}_e} (1 - A_{i,s}^F(\mathbf{n}_e))\mathbb{P}(\mathbf{N}_e^F = \mathbf{n}_e).$$

As for the quasi-stationary approximation, while the numerical evaluation of $\mathbb{P}^F(\mathbf{n}_s \mid \mathbf{n}_e)$ and $\mathbb{P}(\mathbf{N}_e^F = \mathbf{n}_e)$ is infeasible or cumbersome in general, we consider the following special cases where closed-form expressions exist:

- **Uniform admission control on streaming users**. For uniform admission control, i.e., $\gamma_i \equiv \gamma$ independent of i, we can obtain the following elegant form of the distribution for the *total* number of streaming users (a truncated Poisson distribution), as well as for the normalization constant:

  $$\mathbb{P}(N_s^F = n_s \mid \mathbf{N}_e^F = \mathbf{n}_e) = c_s^F(\mathbf{n}_e) \frac{(\rho_s)^{n_s}}{n_s!},$$

  and

  $$c_s^F(\mathbf{n}_e) = \left(\sum_{k=0}^{n_s^{F,\max}(\mathbf{n}_e)} \frac{(\rho_s)^k}{k!} \right)^{-1},$$

  where $n_s^{F,\max}(\mathbf{n}_e)$ is the maximum number of streaming users for which $P_e(\mathbf{n}_e, \mathbf{n}_s) + P_s(\mathbf{n}_s) \le P$.

- **Uniform admission control on elastic users and perfectly-orthogonal codes.** If we assume perfectly-orthogonal codes ($\alpha = 0$) **and** apply *uniform* admission control for elastic traffic by taking $\beta_j \equiv \beta$ independent of j, $\mathbf{N}_e^F$ is *balanced* [16]. In this case, the dynamics of $\mathbf{N}_s^F$ depends on $\mathbf{N}_e^F$ only through the total number of elastic users N_e, so if we define

  $$h(\mathbf{n}_s) = P - P_s(\mathbf{n}_s),$$

  then, we can write

  $$\mathbb{E}[h(\mathbf{N}_s^F) \mid \mathbf{N}_e^F = \mathbf{n}_e] = \mathbb{E}[h(\mathbf{N}_s^F) \mid N_e^F = n_e] \equiv g(n_e).$$

  If we further define $\nu_j = \frac{W\Gamma_j}{\epsilon_{j,e}[\eta + I_j^r]}$, then, from Eqs. (5) and (11), we obtain

  $$\bar{r}_{j,e}(\mathbf{n}_e) \equiv \bar{r}_{j,e}(n_e) = \frac{\nu_j\, g(n_e)}{n_e}.$$

  Furthermore, $A_{i,e}^F(\mathbf{n}_e)$ is independent of i and depends on $\mathbf{n}_e$ only through the total number of elastic users, i.e., $A_{i,e}^F(\mathbf{n}_e) \equiv A_e^F(n_e)$.

  It follows that, for *arbitrary* file size distributions, and $0 \le n_e \le n_e^{\max} = \lfloor \frac{P_e}{\beta} \rfloor$:

  $$\mathbb{P}(\mathbf{N}_e^F = \mathbf{n}_e) = c_e^F \prod_{k=1}^{n_e} \frac{k A_e^F(k-1)}{g(k)} \prod_{j=1}^{J} \left(\frac{\rho_{j,e}}{\nu_j}\right)^{n_{j,e}},$$

  with $\rho_{j,e} = \lambda_{j,e} f_{j,e}$ and $c_e^F = P(N_e^F = 0)$ can be determined after normalization. We further obtain the distribution of the *total* number of file transmissions:

  $$\mathbb{P}(N_e^F = n_e) = c_e^F \left(\sum_j \frac{\rho_{j,e}}{\nu_j}\right)^{n_e} \prod_{k=1}^{n_e} \frac{k A_e^F(k-1)}{g(k)},$$

  leading to a simple expression for the normalizing constant as before:

  $$c_e^F = \left(\sum_{n_e=0}^{n_e^{\max}} \left(\sum_j \frac{\rho_{j,e}}{\nu_j}\right)^{n_e} \prod_{k=1}^{n_e} \frac{k A_e^F(k-1)}{g(k)}\right)^{-1}.$$

  Remark 2. If the codes are not perfectly orthogonal ($\alpha > 0$), we can still apply the above analysis in case the background noise and inter-cell interference are negligible ($\eta_j + I_j^r \ll \alpha P_{j,e}^a \Gamma_j$) by choosing $\nu_j = \frac{W}{\alpha \epsilon_{j,e} P_{j,e}^a}$.

4. CDMA model abstractions

In order to appreciate how our model maps to actual CDMA systems, we define the following variants based on abstractions in terms of (a) power control granularity and (b) time sharing capability.

4.1. Fixed-Power, All-Users (FPAU) model

While power control is an essential element of the CDMA technology in terrestrial cellular systems such as UMTS, it may be impractical or undesirable in emerging wireless networks where CDMA has been identified as a promising candidate technology. For example, closed-loop feedback for power control may be impractical in underwater acoustic sensor networks [17] due to the extremely high propagation delay, unreliable links, limited bandwidth and half-duplex mode of operation of existing off-the-shelf underwater acoustic modems [18]. In addition, time slotting may be inefficient since large guard bands may be required to account for the large and highly-varying propagation delay.

Hence, we define a Fixed-Power, All-Users (FPAU) model, where power control and time sharing are disabled at the base station. Since $J = 1$, each user u is distinguished only in terms of its type (i.e., streaming ($u \equiv s$) vs elastic ($u \equiv e$)), the system state $(\mathbf{N}_e, \mathbf{N}_s)$ reduces to (N_e, N_s) and we can drop the subscript j from the notations. In addition, since the base station transmits simultaneously to *all* users, $P_u^a = P - P_u$, and Eq. (1) can be written as follows:

$$R_u \le \frac{WP_u}{\epsilon_u\left[\alpha(P - P_u) + \frac{\eta + I_u^r}{\Gamma_u}\right]},$$

which can be rewritten as follows:

$$P_u \ge \frac{R_u \epsilon\left(\frac{\eta + I_{\max}^r}{\Gamma_{\min}} + \alpha P\right)}{W + \alpha \epsilon R_u}.$$

For linear and hexagonal networks and typical propagation models, $\Gamma_u = \Gamma_{\min}$ and $I_u^r = I_{\max}^r$ when user u is located at the *edge* of the cell.

Accordingly, we can define the minimum power required by an (elastic, streaming) user to sustain transmission bit-rate requirements of (r_e, r_s) as follows:

$$\beta = \frac{r_e \epsilon \left(\frac{\eta + I^r_{\max}}{\Gamma_{\min}} + \alpha P \right)}{W + \alpha \epsilon r_e},$$
$$\gamma = \frac{r_s \epsilon \left(\frac{\eta + I^r_{\max}}{\Gamma_{\min}} + \alpha P \right)}{W + \alpha \epsilon r_s}. \tag{12}$$

Conditions (3) and (4) can be written as follows:

$$N_e \beta \leq P_e,$$
$$N_e \beta + N_s \gamma \leq P. \tag{13}$$

By substituting Eq. (12) into Condition (13) and defining $r = \max(r_e, r_s)$, we obtain the following:

$$N_e r_e + N_s r_s \leq \frac{P(W + \alpha \epsilon r)}{\epsilon \left(\frac{\eta + I^r_{\max}}{\Gamma_{\min}} + \alpha P \right)}. \tag{14}$$

4.1.1. Equivalent wired link analysis

According to Condition (14), if we define $c \equiv \frac{P(W+\alpha\epsilon r)}{\epsilon(\frac{\eta+I^r_{\max}}{\Gamma_{\min}}+\alpha P)}$, then the downlink transmission scenario in the FPAU model can be approximated by a wired link with capacity c shared amongst streaming and elastic users, where $c_s = \frac{P_s}{P} c$ is reserved for streaming users. Details of the analysis of this model based on the quasi-stationary and fluid approximations (denoted by $\mathbf{A}(\mathbf{Q})$ and $\mathbf{A}(\mathbf{F})$ respectively) can be found in [19].

4.1.2. General analysis

Referring to Remark 1, to apply the quasi-stationary approximation, the departure rate of elastic users from the cell should only depend on the system state $(\mathbf{N}_s, \mathbf{N}_e)$ through N_s. From Eq. (6), we have the following expression:

$$\mu_e(N_e, N_s) = \frac{W[P - P_s(N_s)]}{\epsilon_e \left[\alpha (P - \frac{P - P_s(N_s)}{N_e}) + \frac{\eta + I^r_{\max}}{\Gamma_{\min}} \right]}.$$

It is not straightforward to obtain an approximation, $\mu_e(N_s)$, for $\mu_e(N_e, N_s)$. On the other hand, the fluid approximation developed in Section 3.2 can be applied for this model.

4.2. Power Control, Time Sharing (PCTS) model

Based on our definition in Section 1, each streaming (elastic) user u has a fixed (minimum) transmission bit-rate requirement, denoted by r_u. According to our resource reservation policy, while each streaming user transmits at *fixed* bit-rate r_u, the transmission bit-rate of an elastic user u, R_u ($\geq r_u$), depends on the resource unclaimed by streaming traffic, given by $P - P_s(\mathbf{N}_s)$. From Eq. (1), R_u can be maximized by minimizing P^a_u. One approach to do so is to apply time sharing amongst elastic users.

If we aggregate all elastic users, the resource sharing mechanism is such that the base station transmits using (almost-) orthogonal codes to all users, where the aggregate elastic user may be assigned several codes. Within the aggregate user, elastic users sharing the same code are served in a time-slotted fashion so that they do not interfere with one another, but only with elastic users using different codes and streaming traffic.

Hence, we define a Power Control, Time Sharing (PCTS) model, where the base station can perform discrete power control (at different steps of power) and also supports time sharing resource sharing amongst elastic users as described above. This resource sharing mode is similar to UMTS/HSDPA, where up to $N_c = 4$ codes can be shared amongst data (elastic) users. However, a "true" HSDPA system relies on channel-awareness, link adaptation and turbo codes, which offer a gain factor of 3 in terms of mean throughput as demonstrated in [10]. Although these enhancements are not considered in our PCTS model, the resulting gain in performance may be manifested by a gain function G(N_e) [12], without introducing additional modeling complexity that may render the model non-tractable. We assume that $N_c = 1$ in our study.

4.2.1. Impact on admission control

According to the above resource sharing policy, the received signal at each streaming user u in segment j is interfered by simultaneous transmissions to all other users, i.e., P^a_u = $P - P_u$ and from (2) we obtain

$$\gamma_j = \frac{r_{j,s} \epsilon_{j,s} [\alpha P \Gamma_j + \eta + I^r_j]}{(W + \alpha r_{j,s} \epsilon_{j,s}) \Gamma_j}.$$

Table 1
UMTS cell and traffic parameters for performance evaluation.

UMTS and traffic parameters	
$P(W)$	(20, 0.2)
$P_s(W)$	10
$\eta(W)$	6.09×10^{-14}
W (chips/s)	3.84×10^{6}
ε (dB)	2
α	0.5
Propagation model	Okumura–Hata model [21]
Inter-cell interference model	Hexagonal network with maximum tx. power [13]
Link budget	Table 8.3 [14]
r_e (kbps)	128
r_s (kbps)	128

For an elastic user u in segment j, we have $P_u^a = P_s(\mathbf{N}_s)$ since its received signal is only interfered by streaming users. Hence, the power required by an elastic user in segment j to sustain its transmission bit-rate requirement, $r_{j,e}$, depends on the number and location of streaming users as follows:

$$\beta_j(\mathbf{N}_s) = \frac{r_{j,e}\epsilon_{j,e}[\alpha P_s(\mathbf{N}_s)\Gamma_j + \eta + I_j^r]}{W\Gamma_j}.$$

The admission control scheme is such that a newly-arrived user is blocked only if accepting it would violate either the static reservation policy or the minimum power requirement of any user. At any time, streaming traffic can claim a portion P_s of the total power P. Therefore, the power required by an elastic user in segment j is given by:

$$\beta_j \equiv \beta_j(\mathbf{N}_s) = \frac{r_{j,e}\epsilon_{j,e}[\alpha P_s(\mathbf{N}_s)\Gamma_j + \eta + I_j^r]}{W\Gamma_j}.$$

4.2.2. Impact on rate allocation

Using Eqs. (5) and (6), with time sharing amongst elastic users, the departure rate of elastic users in segment j is given by:

$$\mu_{j,e}(\mathbf{N}_e, \mathbf{N}_s) = \frac{N_{j,e}W[P - P_s(\mathbf{N}_s)]}{f_{j,e}N_e\epsilon_e[\alpha P_s(\mathbf{N}_s) + \frac{\eta + I_j^r}{\Gamma_j}]}. \quad (15)$$

Since $\frac{N_{j,e}}{N_e} \leq 1$, we have the following:

$$\mu_{j,e}(\mathbf{N}_e, \mathbf{N}_s) \leq \frac{W[P - P_s(\mathbf{N}_s)]}{f_{j,e}\epsilon_e[\alpha P_s(\mathbf{N}_s) + \frac{\eta + I_j^r}{\Gamma_j}]} \equiv \mu_{j,e}(\mathbf{N}_s).$$

Referring to Remark 1, to apply the quasi-stationary approximation, it is necessary to remove the dependence of $\mu_{j,e}(\mathbf{N}_e, \mathbf{N}_s)$ on $\mathbf{N}_e$ in Eq. (15). This can be achieved by approximating $\mu_{j,e}(\mathbf{N}_e, \mathbf{N}_s)$ with $\mu_{j,e}(\mathbf{N}_s)$; this approximation is exact when power control is disabled (i.e., $J = 1$). As with the FPAU model, the fluid approximation can be applied for this model.

Further details on the derivation of the quasi-stationary and fluid approximations (denoted by $\mathbf{A}(\mathbf{Q}, \mathbf{J})$ and $\mathbf{A}(\mathbf{F}, \mathbf{J})$ respectively) for this model can be found in [20].

5. Performance evaluation

In this section, our objective is to evaluate whether the performance gain achieved with time sharing and power control justifies the added processing complexity and signaling overhead at the base station in a UMTS downlink scenario.

We consider a single UMTS cell whose radius, δ_J, is computed using the reference link budget given in Table 8.3 of [14] and the Okumura–Hata propagation model [21] for an urban macro cell. The inter-cell interference at each location within the cell is computed based on the conservative approximation for a hexagonal network [13].

Elastic (streaming) users arrive at the cell according to a Poisson process at rates λ_e (λ_s), transmission bit-rate requirement r_e (r_s), target energy-to-noise ratio ϵ_e (ϵ_s), mean file size f_e (holding time $\frac{1}{\mu_s}$) and are assumed to be uniformly distributed over the cell. In addition to the mean number of users, (E[N_e], E[N_s]), and blocking probabilities, (p_e, p_s), for each class of traffic, we define the *stretch*, S_e, for each admitted elastic user by normalizing the expected residence time, $E[T_e]$, by the mean file size, f_e, i.e., $S_e = \frac{E[T_e]}{f_e} = \frac{E[N_e]}{\lambda_e(1-p_e)}$ (cf. Little's Theorem). A summary of the cell and traffic parameters is given in Table 1.

In [19] and [20], through simulations, we have demonstrated that the user performance obtained with the FPAU and PCTS model (as defined in Section 4) is almost insensitive to the actual distribution of the traffic parameters. This justifies the application of the approximation techniques we develop, which depend on the traffic parameter distribution only through

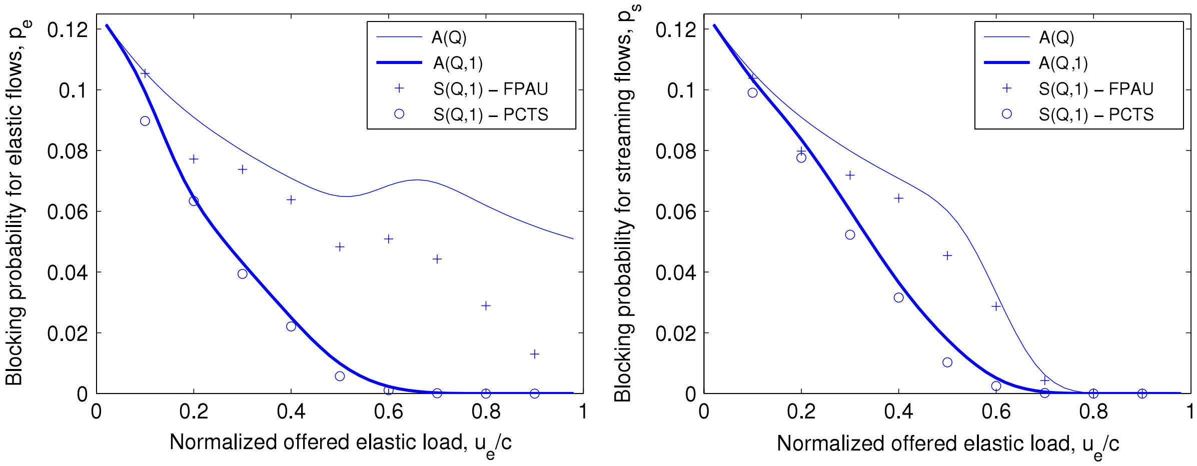

Fig. 1. Blocking probability for elastic (left) and streaming users (right) vs normalized offered elastic load obtained for quasi-stationary regime ($J = 1$).

the mean values. In addition, we also demonstrated the accuracy of the approximations, particularly for the extreme (quasi-stationary and fluid) traffic regimes.

Here, we focus on the comparison of the FPAU model and the PCTS model for the base station based on simulation as well as the approximations. Unless otherwise stated, we assume that (d_s, s_e) are exponentially distributed with mean $\frac{1}{\mu_s}$ and f_e respectively.

5.1. *Simulation procedure*

We develop a simulation program for our model by considering arrival/departure events of traffic users (elastic or streaming). Each simulation scenario is defined according to the following procedure:

1. Fix the granularity of power control, J:
   $J = 1$: no power control (FPAU or PCTS model);
   $J > 1$: discrete power control (PCTS model).
2. Fix the total offered traffic by choosing the *loading factor*, $l > 0$, where $u_e + u_s = lc$, $u_e = \lambda_e f_e$ and $u_s = \frac{\lambda_s r_s}{\mu_s}$;
3. For each l, fix the traffic *mix*, $\frac{u_e}{lc}$, by choosing u_e, $0 \le u_e \le lc$;
4. For each traffic mix, select (λ_e, λ_s) to fit one of the following traffic regimes:
   a. Quasi-stationary regime ($\mathbf{S}(\mathbf{Q}, \mathbf{J})$, cf. Section 3.1);
   b. Fluid regime ($\mathbf{S}(\mathbf{F}, \mathbf{J})$, cf. Section 3.2);
   c. Neutral regime ($\mathbf{S}(\mathbf{N}, \mathbf{J})$, fits neither a nor b)

We generate 5 sets of simulation results for each scenario, for which the sample mean for each performance metric is computed and used for performance comparison.

5.2. *Impact of time sharing (FPAU vs PCTS ($J = 1$))*

We begin by investigating the performance gain achieved with time sharing by comparing the performance obtained for the FPAU and PCTS model ($J = 1$) for various traffic regimes.

5.2.1. *Quasi-stationary regime*

We plot (p_e, p_s) and $(\mathrm{E}[N_e], S_e)$ as a function of the traffic mix, $\frac{u_e}{c}$, $0 \le u_e \le c$ in Figs. 1 and 2 respectively. We note that, since it is not straightforward to apply the quasi-stationary approximation to the FPAU model (cf. Section 4.1.2), we utilize the equivalent wired link analysis to obtain the corresponding quasi-stationary approximation, $\mathbf{A}(\mathbf{Q})$.

Based on the simulation results, we observe a performance gain achieved as a result of time sharing in terms of reduced blocking probabilities, queue length and sojourn time. This gain is expected since, for a given number of streaming users, time sharing amongst elastic users reduces the intra-cell interference power experienced by each elastic user, thereby increasing the transmission bit-rate per elastic user. This gain is marginal when elastic load is low, since the additional interference experienced by an elastic user due to other elastic users (without time sharing) is insignificant.

In terms of the accuracy of approximations, we observe that the performance obtained with the PCTS ($J = 1$) model is well tracked by the corresponding approximation; on the other hand, the equivalent wired link analysis results in overly conservative estimates of the performance for the FPAU model.

5.2.2. *Fluid regime*

We plot (p_e, p_s) and $(\mathrm{E}[N_e], S_e)$ as a function of the traffic mix, $\frac{u_e}{c}$, $0 \le u_e \le c$ in Figs. 3 and 4 respectively.

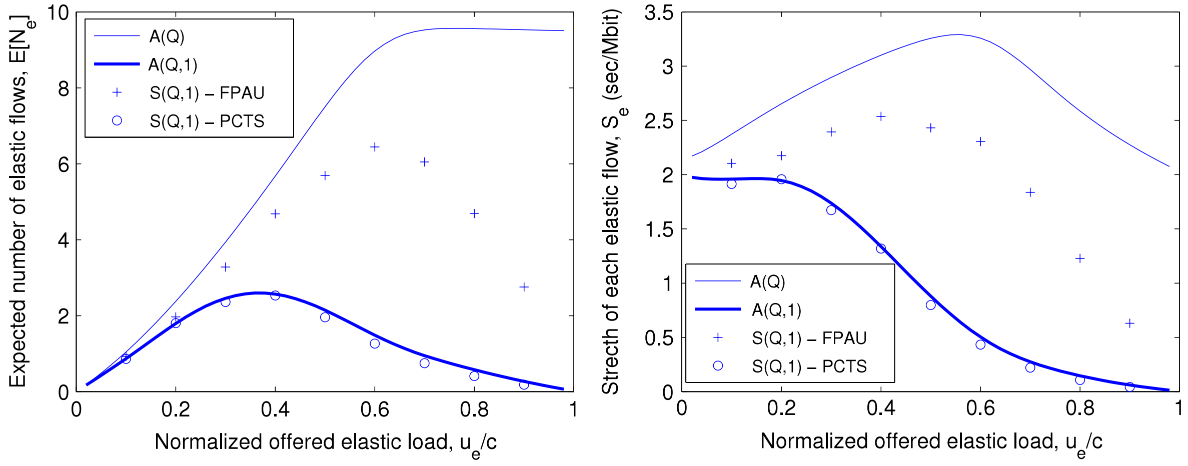

Fig. 2. Number of active elastic users (left) and stretch of each admitted elastic user vs normalized offered elastic load obtained for quasi-stationary regime ($J = 1$).

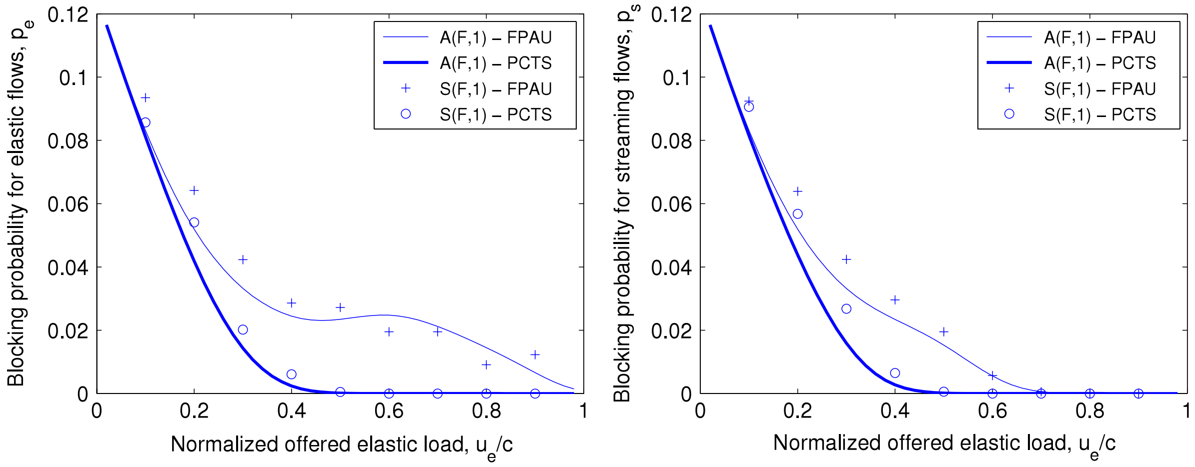

Fig. 3. Blocking probability for elastic (left) and streaming users (right) vs normalized offered elastic load obtained for fluid regime ($J = 1$).

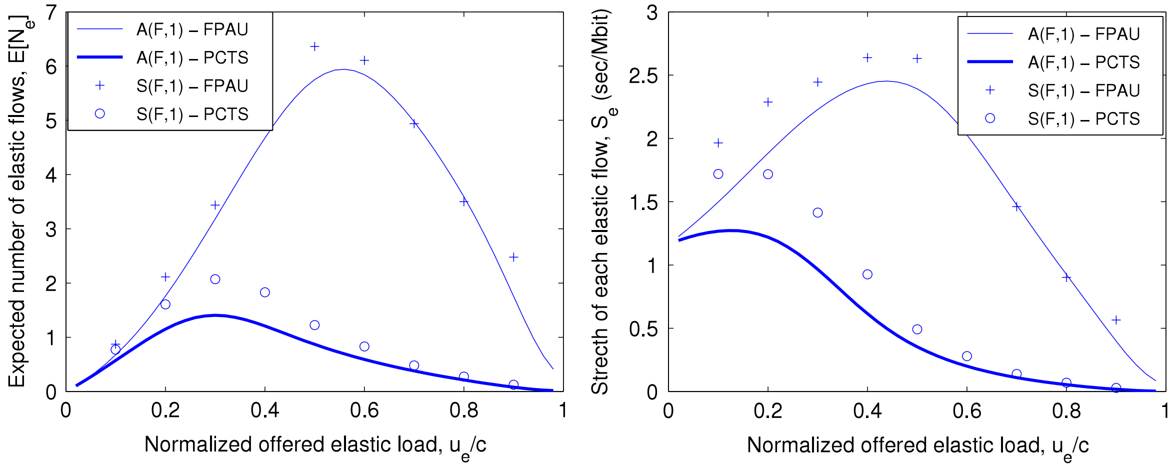

Fig. 4. Number of active elastic users (left) and stretch of each admitted elastic user vs normalized offered elastic load obtained for fluid regime ($J = 1$).

As with the quasi-stationary regime, we observe a performance gain achieved as a result of time sharing in terms of reduced blocking probabilities, queue length and sojourn time. In terms of the accuracy of approximations, the blocking performance obtained with both models is well tracked by the corresponding approximations. However, the approximations achieved more optimistic estimates of the queue length and sojourn time of elastic users.

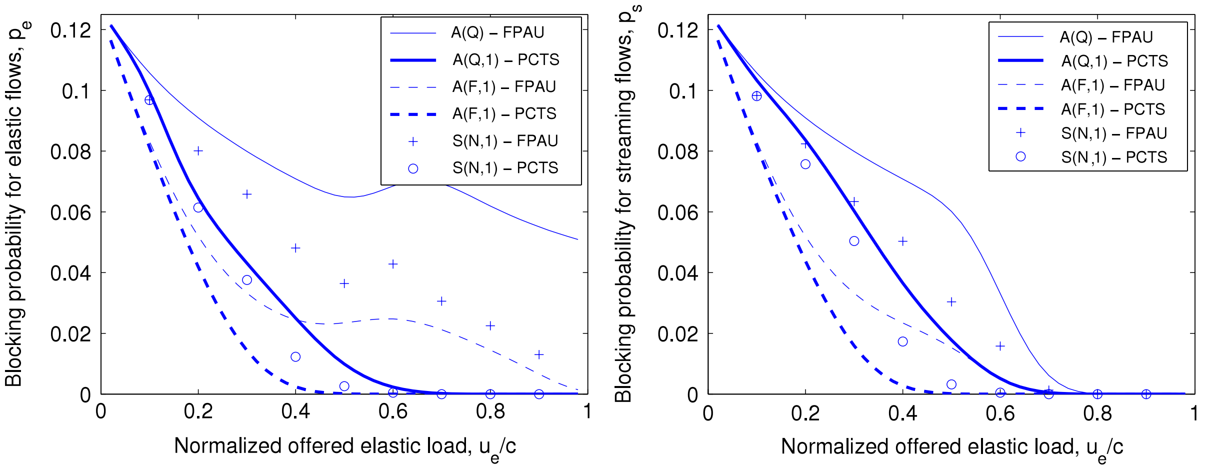

Fig. 5. Blocking probability for elastic (left) and streaming users (right) vs normalized offered elastic load obtained for neutral regime ($J = 1$).

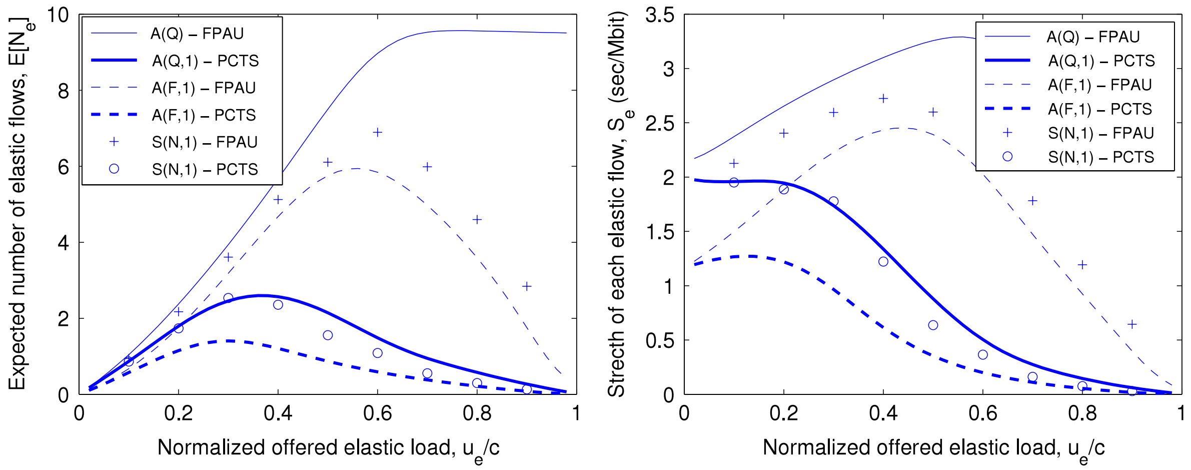

Fig. 6. Number of active elastic users (left) and stretch of each admitted elastic user vs normalized offered elastic load obtained for neutral regime ($J = 1$).

5.2.3. *Neutral regime*

We plot (p_e, p_s) and $(\mathrm{E}[N_e], S_e)$ as a function of the traffic mix, $\frac{u_e}{c}$, $0 \leq u_e \leq c$ in Figs. 5 and 6 respectively.

As with the "extreme" traffic regimes, we observe a performance gain achieved as a result of time sharing in terms of reduced blocking probabilities, queue length and sojourn time. For each performance metric, we note that the quasi-stationary (fluid) approximation upper (lower) bounds the performance obtained in the neutral traffic regime, where a tighter bound is obtained with the PCTS ($J = 1$) model.

5.3. *Impact of power control (PCTS model)*

Next, we investigate the performance gain achieved with various levels of power control granularity, J, for the PCTS model. We define each segment j as the annulus between concentric rings of radius δ_{j-1} and δ_j such that $\delta_j = \frac{j}{J}\delta_J$, $1 \leq j \leq J$. Since user arrivals are uniformly distributed over the cell, their arrival rate in each ring j is $\lambda_j = \frac{\delta_j^2 - \delta_{j-1}^2}{\delta_J^2}\lambda$, where δ_0 = 0.

5.3.1. $P = 20$ W

We plot (p_e, p_s) and $(\mathrm{E}[N_e], S_e)$ as a function of the traffic mix, $\frac{u_e}{c}$, $0 \leq u_e \leq c$, for $\mathbf{S(N, J)}$ in Figs. 7 and 8 respectively for $J = 1, 2$ and ∞ (corresponding to the case of perfect power control) for a neutral traffic regime. We observe that the cell performance obtained with simulation is lower bounded (well approximated) by $\mathbf{A(F, J = 1)}$ ($\mathbf{A(Q, J = 1)}$), and that $\mathbf{S(N, J)}$ is almost invariant with the value of J. Hence, no significant performance gain is achieved through finer power control in this case, and therefore, the performance can be approximated with the PCTS ($J = 1$) model.

5.3.2. $P = 0.2$ W

In order to demonstrate the performance gain with finer power control, we repeat the simulations for the case of $P = 0.2$ W, and plot (p_e, p_s) and $(\mathrm{E}[N_e], S_e)$ as a function of the traffic mix, $\frac{u_e}{c}$, $0 \leq u_e \leq c$, for $\mathbf{S(F, J)}$ in Figs. 9 and 10

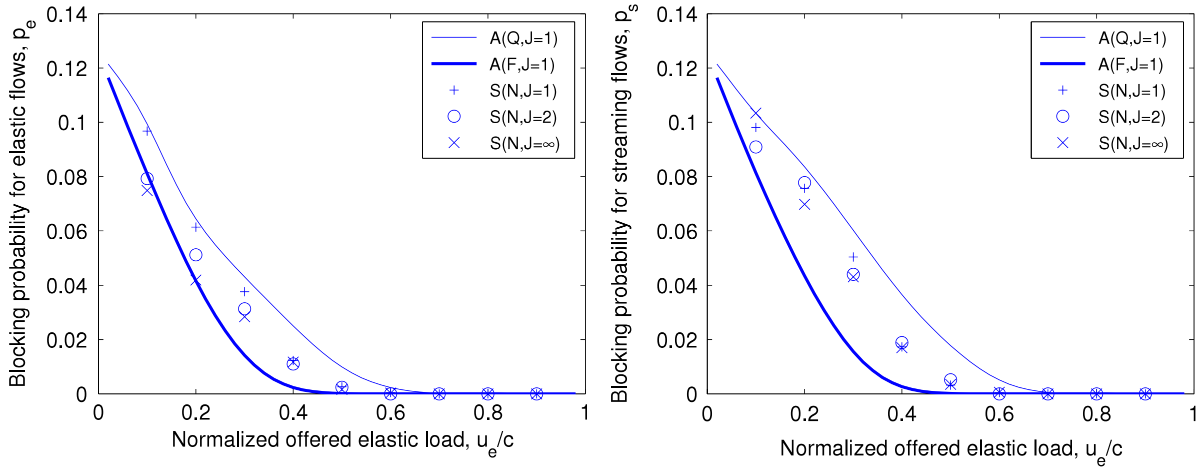

Fig. 7. Blocking probability for elastic (left) and streaming users (right) vs normalized offered elastic load obtained with approximation and simulation for PCTS model ($J = 1, 2, \infty$, $P = 20$ W, neutral regime).

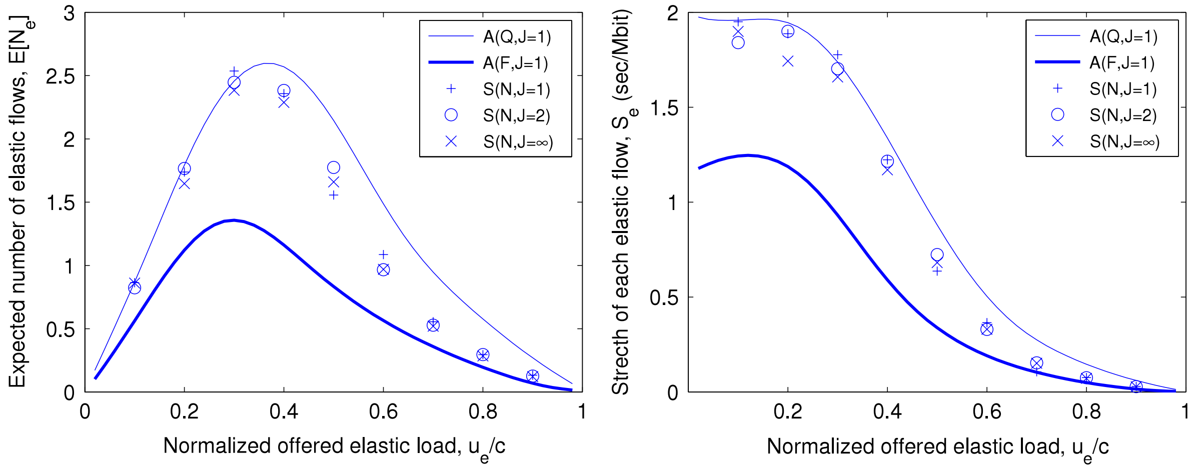

Fig. 8. Number of active elastic users (left) and stretch of each admitted elastic user vs normalized offered elastic load obtained with approximation and simulation for PCTS model ($J = 1, 2, \infty$, $P = 20$ W, neutral regime).

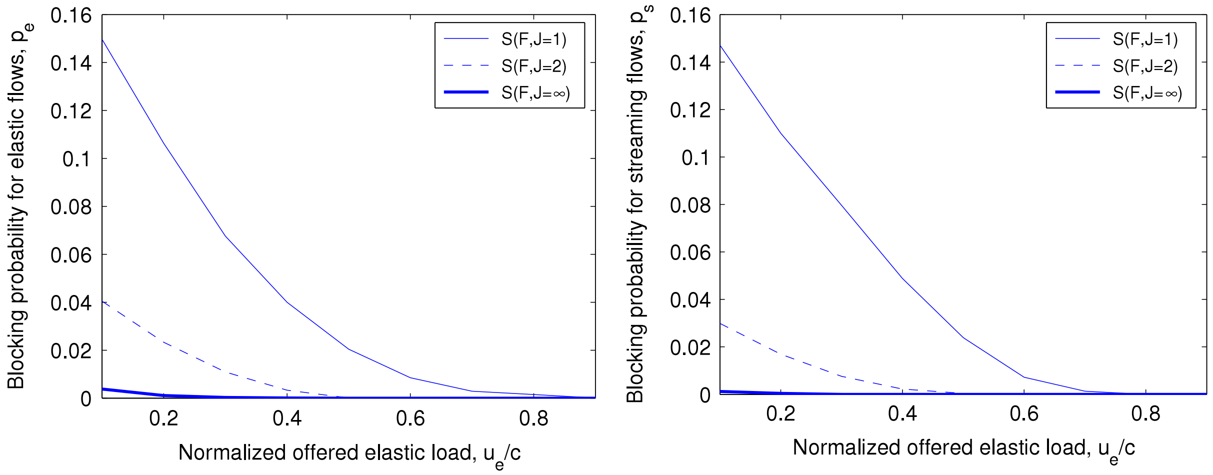

Fig. 9. Blocking probability for elastic (left) and streaming users (right) vs normalized offered elastic load obtained with simulation for PCTS model ($J = 1, 2, \infty$, $P = 0.2$ W, fluid regime).

respectively. In this case, we note that as power control becomes finer (increasing J), the performance obtained with **S(F, J)** is improved significantly (e.g., reduced blocking and sojourn time).

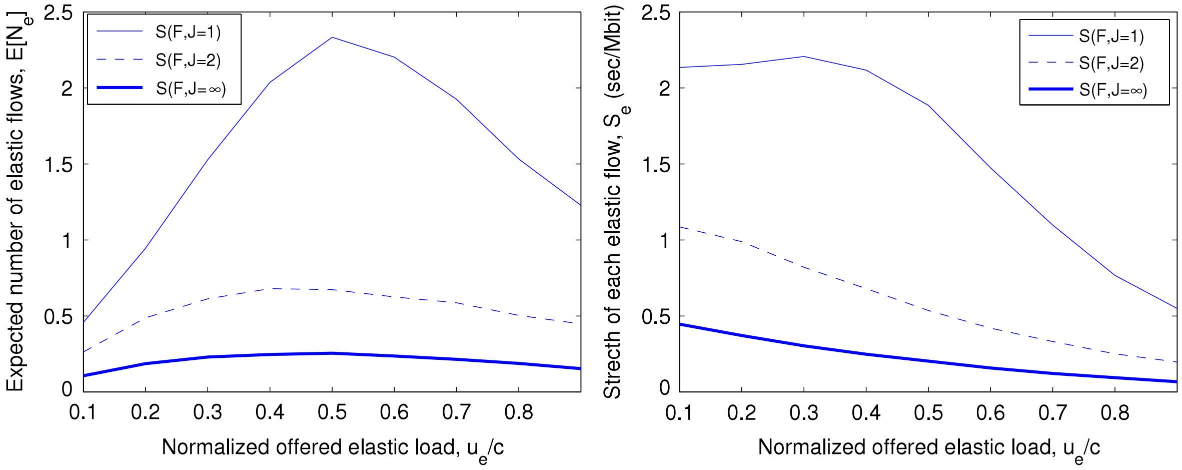

Fig. 10. Number of active elastic users (left) and stretch of each admitted elastic user vs normalized offered elastic load obtained with simulation for PCTS model ($J = 1, 2, \infty, P = 0.2$ W, fluid regime).

5.4. Comparison with other fair resource sharing policies

In this paper, we assume that all elastic users receive an equal portion of the available power, so that we can apply well-known results from Generalized Multi-Class Processor Sharing [15] to model their behavior in the quasi-stationary approximation. With such an allocation policy, the transmission bit-rate of an elastic user will decrease with its distance from base station. This follows the same pattern as an optimal bit-rate allocation, but it is not necessarily optimal. It is also not max–min fair, since it is possible to increase the bit-rate in a segment by decreasing the bit-rate in a segment closer to the base station while leaving the bit-rates in the other segments unchanged.

5.5. Extension to multi-cell scenario

In this paper, we considered the analysis of a single CDMA cell in order to simplify the way in which interference is taken into account. However, our one cell model can be easily embedded in a multi-cell scenario where each base station adjusts its power according to the level of the interference encountered or when the base stations transmit at a fixed power. The assumption of orthogonal codes or negligible interference was used only for obtaining a closed-form solution in the fluid approximation for the distribution of the number of elastic users.

6. Conclusions

Future Generation CDMA wireless systems can simultaneously accommodate users carrying widely heterogeneous applications. Since resources are limited, particularly in the air interface, admission control is necessary to ensure that all active users are accommodated with sufficient bandwidth to meet their specific Quality of Service requirements.

We propose a general traffic management framework that supports differentiated admission control, resource sharing and rate allocation strategies, such that users with stringent transmission bit-rate requirements ("streaming traffic") are protected while sufficient capacity over longer time intervals to delay-tolerant users ("elastic traffic") is offered. This framework permits discrete power control by distinguishing users according to their distance from the base station through cell partitioning, and also supports a time sharing resource sharing mode to improve rate allocation to elastic traffic while guaranteeing the transmission bit-rate requirements of all users. While our admission control strategy may not satisfy classical notions of fairness, we aim to reduce congestion and increase overall throughput of elastic users.

Since the exact analysis to evaluate the performance of such an integrated services system is non-tractable in general, we define extreme traffic regimes (quasi-stationary and fluid) for which time-scale decomposition can be applied to isolate the traffic streams, from which known results from fluid queueing models are used to approximate the performance for each user type. For the extreme traffic regimes, simulation results suggest that the performance is almost insensitive to traffic parameter distributions, and is well approximated by our proposed approximations. In addition, we also demonstrate the performance gain achieved with finer power control, as well as applying time sharing amongst elastic users to improve their rate allocation.

Acknowledgments

The support of Vodafone is gratefully acknowledged. This research is partially supported by the Dutch Bsik/BRICKS project and is performed within the framework of the European Network of Excellence Euro-NGI. Much of the research was performed while Adriana F. Gabor and Hwee-Pink Tan were affiliated with EURANDOM and Rudesindo Núñez-Queija was affiliated with the Eindhoven University of Technology. We are also grateful to the reviewers for their careful reading of the manuscript and their useful comments, which have led to several improvements.

References

[1] P. Key, L. Massoulié, A. Bain, F. Kelly, Fair internet traffic integration: Network flow models and analysis, Annales des Telecommunications 59 (2004) 1338–1352.
[2] T. Bonald, A. Proutière, On performance bounds for the integration of elastic and adaptive streaming flows, in: Proceedings of the ACM SIGMETRICS/Performance, 2004, pp. 235–245.
[3] P. Key, L. Massoulié, Fluid limits and diffusion approximations for integrated traffic models, Technical Report MSR-TR-2005-83, Microsoft Research, June 2005.
[4] R. Núñez-Queija, J.L. van den Berg, M.R.H. Mandjes, Performance evaluation of strategies for integration of elastic and stream traffic, in: D. Smith, P. Key. (Eds.), Proc. ITC 16, Elsevier, Amsterdam, 1999, pp. 1039–1050.
[5] R. Núñez-Queija, Processor-sharing models for integrated-services networks, Ph.D. Thesis, Department of Mathematics and Computer Science, Eindhoven University of Technology, 2000.
[6] N. Benameur, S.B. Fredj, F. Delcoigne, S. Oueslati-Boulahia, J.W. Roberts, Integrated admission control for streaming and elastic traffic, Lecture Notes in Computer Science 2156 (2001) 69–81.
[7] F. Delcoigne, A. Proutière, G. Regnie, Modeling integration of streaming and data traffic, Performance Evaluation 55 (2004) 185–209.
[8] S. Borst, N. Hegde, Integration of streaming and elastic traffic in wireless networks, Proceedings of the IEEE Infocom (2007) 1884–1892.
[9] F. Baccelli, B. Blaszczyszyn, M.K. Karray, Up- and downlink admission/congestion control and maximal load in large homogeneous CDMA networks, Mobile Networks and Applications (2004) 605–617.
[10] B. Blaszczyszyn, M.K. Karray, Performance evaluation of scalable congestion control schemes for elastic traffic in cellular networks with power control, Proceedings of the IEEE Infocom (2007) 170–178.
[11] F. Baccelli, B. Blaszczyszyn, F. Tournois, Downlink admission/congestion control and maximal load in CDMA networks, Proceedings of the IEEE Infocom (2003) 723–733.
[12] S. Borst, User-level performance of channel-aware scheduling algorithms in wireless data networks, IEEE/ACM Transactions on Networking (2005) 636–647.
[13] T. Bonald, A. Proutière, Wireless downlink data channels: User performance and cell dimensioning, in: Proc. of the ACM MOBICOM, 2003, pp. 339–352.
[14] H. Holma, A. Toskala, WCDMA for UMTS, Radio Access for Third Generation Mobile Communications, John Wiley and Sons, 2001.
[15] J.W. Cohen, The multiple phase service network with generalized processor sharing, Acta Informatica 12 (1979) 245–284.
[16] T. Bonald, A. Proutière, Insensitive bandwidth sharing in data networks, Queueing Systems 44 (2003) 69–100.
[17] I.F. Akyildiz, D. Pompili, T. Melodia, State of the art in protocol research for underwater acoustic sensor networks, ACM Mobile Computing and Communications Review (2007) 11–22 (invited paper).
[18] J. Partan, J. Kurose, B.N. Levine, A survey of practical issues in underwater networks, in: Proc. of the 1st ACM International Workshop on Underwater Networks, WUWNet, 2006, pp. 17–24.
[19] O.J. Boxma, A.F. Gabor, R. Núñez-Queija, H.P. Tan, Performance analysis of admission control for integrated services with minimum rate guarantees, in: Proc. of 2nd NGI, 2006, pp. 41–47.
[20] R. Núñez-Queija, H.P. Tan, Location-based admission control for differentiated services in 3G cellular networks, in: Proc. of the 9th ACM-IEEE MSWiM, 2006, pp. 322–329.
[21] Y. Wang, T. Ottosson, Cell search in W-CDMA, IEEE Journal on Selected Areas in Communications 18 (2000) 1470–1482.

Hwee-Pink Tan is a Senior Research Fellow with the Networking Protocols Department, Institute for Infocomm Research (I^{2}R), Singapore. He received his Ph.D. in September 2004 from the Technion, Israel Institute of Technology. Before joining I^{2}R, he was a Post-doctoral Researcher at EURANDOM, The Netherlands from December 2004 to June 2006, and a Research Fellow with CTVR, Trinity College Dublin, Ireland from July 2006 to March 2008. His research has mainly focused on the performance analysis of wireless networks, and his current research interests are in underwater networks, cognitive radio networks and wireless sensor networks powered by ambient energy harvesting.

Rudesindo Núñez Queija is associate professor of Operations Research and Management (ORM) at the University of Amsterdam (UvA) and is part-time affiliated with the Center for Mathematics and Computer Science (CWI) in Amsterdam. Since 2000 he has been a staff member at CWI (currently in the group of Stochastic Networks and Probability). He held part-time positions as assistant professor in Stochastic Operations Research at TU/e (2000–2006) and as a staff member at TNO Information and Communication Technology in Delft (2006–2008). In 2008 he joined the research group ORM of the Department of Quantitative Economics (Faculty of Economics and Business) at the UvA. His research has mainly focused on Queueing Theory and in particular its application to data transmission in bandwidth sharing networks and random file sharing networks.

Adriana F. Gabor received her Ph.D. in 2002 from University of Twente, The Netherlands. Currently she is assistant professor in the Econometrics department of the Erasmus School of Economics, Rotterdam, The Netherlands. Her main research interests are in combinatorial optimization problems with stochastic data and their applications in logistics and telecommunications.

Onno J. Boxma (1952; Ph.D. Utrecht, 1977) holds the chair of Stochastic Operations Research in Eindhoven University of Technology, and is scientific director of the European research institute EURANDOM. Onno Boxma is a co-author/co-editor of five books on queueing theory and performance evaluation. His main research interests are in queueing theory and its applications to the performance analysis of computer-communication and production systems. He has published over 150 refereed papers on these subjects. He serves on the editorial board of several journals, presently being editor-in-chief of Queueing Systems. Onno Boxma is member of IFIP WG7.3, and honorary professor in Heriot-Watt University, Edinburgh. In June 2009 he will receive a honorary doctorate from the University of Haifa.