Telecommunication Systems (2021) 77:241–253
https://doi.org/10.1007/s11235-021-00752-7

Performance analysis of cellular networks with delay tolerant users

Feng Yan[1] · Patrick Maillé[2] · Xavier Lagrange[2]

Accepted: 4 January 2021 / Published online: 28 January 2021

Abstract

In this paper, we analyze the impact of delaying delay-tolerant calls under certain conditions in cellular networks. We propose to queue the call if the user agrees when the terminal has bad radio conditions and the system is loaded. The call is served as soon as radio conditions become good or the current load goes below a given threshold. We model the system as a continuous-time Markov chain, which allows us to compute the blocking probability, the mean waiting time and the mean service time. Numerical results show that when the proportion of users with delay tolerance is 20%, the system can bear 16% more calls with the same blocking probability, and 113% more calls if 80% of users are delay tolerant.

Keywords Smart phone · Delay tolerant users · Cellular networks

1 Introduction

With the explosive growth of smart phones and tablets, mobile data traffic has been approximately doubling each year in the early 2010s. According to the last Ericsson report [1], the mobile data traffic grew around 54% between the first quarter of 2017 and the first quarter of 2018. Such a traffic growth raises big challenges to cellular networks. In dense areas, it is quite necessary to deploy heterogeneous networks [2–4] or to combine cellular and WiFi technologies [7,8] to cope with traffic growth.

A lot of applications can tolerate a delay (e.g., non-urgent file download, pull services). The delay tolerance feature has already been exploited in [5–10]. In [5], the trade-off between delaying a service and minimizing the energy consumption is studied. In [6], the authors propose to combine the use of neighbor terminals (crowd computing approach) and mobility prediction to limit the delay before getting the service. In [7], a model is proposed to predict WiFi connectivity and offload cellular networks by steering delay-tolerant data onto WiFi. In [8], the authors propose an integrated architecture to migrate data traffic from cellular networks to WiFi networks and quantify the number of WiFi access points (APs) required for a city-wide WiFi offloading. The performance of offloading in [7,8] is clearly closely related to the WiFi availability, which is better in urban than in rural environments. In [9], the authors study how to select a number of key locations in cellular networks to upgrade capacity, and shift delay-tolerant traffic to them. In [10], the authors investigate solutions for network-controlled WiFi offloading in Long Term Evolution (LTE) cellular networks when performance needs exceed the capacity of the LTE.

Both the capital and the operational expenses that are required for the deployment of WiFi or LTE small cells are significant. In rural environments, such additional expenses can be prohibitive, especially in developing countries where the monthly subscription fee should be kept as low as possible. Also in some cases, deploying micro-cells is technically difficult because of the lack of energy sources.

Data traffic generally exhibits some degrees of heterogeneity in both the time and space domains. It is well known that the dimensioning of network resource is done to cope with the traffic conditions for peak hours in the day or even for peak periods in one given day of the week. The average usage of network resource is thus generally very low.

Our objective is to analyze the capacity increase of cellular networks by exploiting user delay tolerance without deploying new base stations or access points and without adding any resource. The capacity is defined in this paper as the maximum traffic arrival rate for which the blocking probability of the system is below a target [11]. In [12], we proposed an architecture based on a specific server in the network

✉ Feng Yan
feng.yan@seu.edu.cn

1 National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China

2 IMT Atlantique/IRISA, 2 Rue de la Chataigneraie, 35510 Cesson Sevigne, France

and a mobile application that queues non-urgent download requests and determines the best time to trigger a download that is queued. We also presented a proof of concept. In [13], we analyze the impact of user delay tolerance on cellular networks but from the energy efficiency perspective. In this paper, we focus on the capacity increase that is possible by delaying the service of some users in the network.

In [14], the authors consider different priority schemes and the model they propose could be used for delay-tolerant systems. However, they assume the service rate is the same for all users in a cell. In contrast, our model considers two zones (inner and outer) with different rates, through a simple but appropriate model to take into account the effect of the radio conditions. In [15], the authors propose a call admission scheme that uses mobility prediction. The proposal shows good performance but it applies to voice calls as it considers a given number of channels and a loss system (a call is rejected if all channels are busy). Here, we consider data traffic with different scheduling policies and a queuing system. We propose an analytical model, that is of course a simplified view of reality, but is based on the widely accepted assumption of a regular hexagonal network, for which we build a Markov chain. Unlike simulation models, the way to compute all variables is explained, which ensures reproducibility.

We consider that the load of cells is mainly due to the data traffic, and focus on interactive services (for example, web browsing). Each user alternatively downloads some content and reads this content. We refer to the download request as a *packet call*. We assume that a call admission control is activated by the operator to guarantee a minimum bit rate. Hence, a packet call can be blocked in case of overload. Of course, enough frequency bandwidth should be allocated to the cell to ensure a low blocking probability (typically 0.1% in normal conditions).

In this paper, we refer to a user with delay-tolerant data traffic as a *Delay Tolerant User* (DTU). Users whose calls cannot tolerate delays are referred to as non-DTUs. Note that a user can be a DTU for a service and a non-DTU for another service. We propose to queue the call of a DTU, upon its arrival, if the user has bad radio conditions and the current load of the system is above a given threshold. The call gets served when either the radio conditions improve or the load goes below the threshold. We analyze the impact of DTUs on the capacity of cellular networks. Of course, delaying the service is not possible in all environments and all circumstances. However, due to its simplicity, it is worth studying such a solution in constrained networks (e.g., cost constraints, lack of energy, etc.).

The remainder of the paper is organized as follows. Section 2 presents the system model considered in this paper, including the mobility model, the traffic model, the scheduling strategies and the admission control policy. In Sect. 3, a Markov chain is defined for that system, and the blocking probability, mean service time and mean waiting time of DTUs are derived. Numerical results are given in Sect. 4. Finally, Sect. 5 concludes the paper.

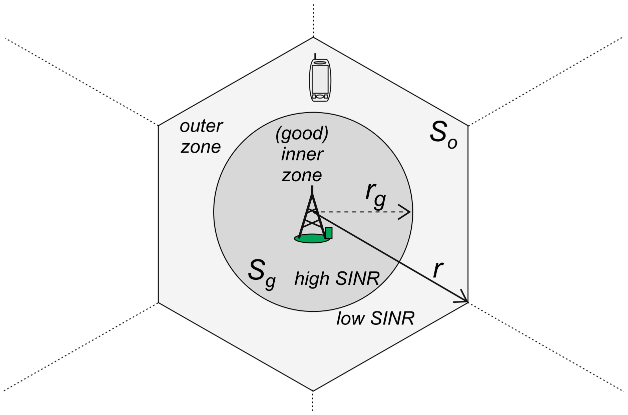


Fig. 1 Inner and outer zones of a cell

2 System model

We consider a regular hexagonal cellular network with the Okumura–Hata propagation model [16]: the received power at the terminal is proportional to $1/x^\eta$ where x is the distance to the base station and η is an environment-dependent parameter (typically $\eta = 3.3$). The same frequency carrier is assumed to be used in all cells (the reuse cluster size is 1). The SINR (Signal-to-Interference-and-Noise Ratio) thus depends only on x (see Sect. 2.2.1).

Each cell is divided into two zones: the inner zone in which users have a SINR higher than a threshold and can have a high transmission rate, the outer zone in which the SINR is low and hence the rate is reduced. In a simple hexagonal cellular network (see Fig. 1), the inner zone is a disk of radius r_g (we use subscript g for "good" instead of i for inner to avoid any confusion with index i, which will be used in the Markov chain) and the outer zone is the complementary of the disk in the hexagon of radius r. In order to have a system simple enough to make the analysis, we consider that each mobile in the inner (resp. outer) zone gets the same rate as the one at distance r_g (resp. r).

The inner zone is a disk whose area is

$$S_g = \pi r_g^2. \tag{1}$$

The outer zone is defined by the part of the hexagonal cell that is not in the inner zone. Let S_o be the area of the outer zone. Thus,

$$S_o = \frac{3\sqrt{3}}{2} r^2 - \pi r_g^2. \tag{2}$$

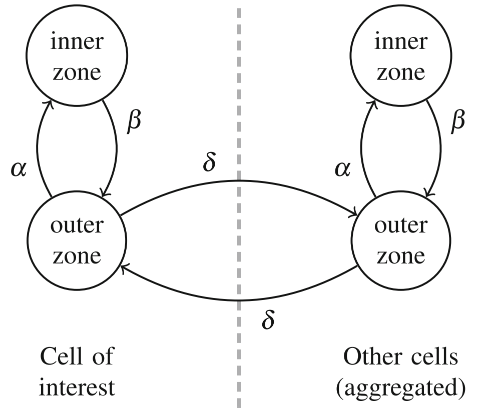


Fig. 2 Continuous-time Markov chain for the user mobility model

2.1 Mobility model

We assume the residence time of a user in any type of zone (inner or outer) is exponentially distributed and that mobility is a memory-less process [17]. When a user is in the outer zone, he/she can go either in the inner zone or in a neighbor cell. In the latter case, he/she is necessarily in the outer zone of the neighbour cell. The user can go an unlimited number of times in the inner zone and then go back to the outer zone. The network is assumed to be regular and we thus consider only two cells; the current cell and a neighbor cell (which becomes the current cell as soon as the user enters it, from the user point of view). The transition rate from the outer zone to the inner zone (resp. to the neighbor cell) is denoted by α (resp. δ). The transition rate from the inner zone to the outer zone is denoted by β. The mobility of each user is modeled by a continuous-time Markov chain as shown in Fig. 2.

Let q_g and q_o be the steady state probability to be in the inner and the outer zone, respectively. We have

$$\begin{cases} q_g = \frac{\alpha}{\alpha+\beta} \\ q_o = \frac{\beta}{\alpha+\beta} \end{cases} \tag{3}$$

We consider an underlying random walk mobility model, where the distribution of the speed is uniform in all the directions. In that case [19,20], it is possible to compute the average number of terminals $\frac{dN}{dT}$ that cross (outwards) the perimeter L of an area S per time unit:

$$\frac{dN}{dT} = \frac{v\rho_T L}{\pi} \tag{4}$$

where ρ_T is the density of terminals and v is the average speed. But with a Markovian model, that average number of outgoing crossings should equal the individual outgoing rate ω of the area multiplied by the number $\rho_T S$ of terminals in the area: $\frac{dN}{dT} = \omega\rho_T S$. Combining with (4) and denoting by T the average dwelling time in the area, we get:

$$\omega = \frac{1}{T} = \frac{vL}{\pi S}. \tag{5}$$

Note that (5) is very general and can be applied to any type of shape. In Sect. 2.1.1, we use it both for the inner zone (circle) and the global cell (hexagon).

Finally, since all users have the same mobility pattern, q_g is also the proportion of users in the inner zone. If the repartition of users over space is uniform, the steady state probability is clearly proportional to the area of each zone. This implies that

$$\frac{\alpha}{\beta} = \frac{S_g}{S_o}. \tag{6}$$

2.1.1 Computation of the transition rates of the mobility Markov chain

The inner zone is a circle and thus, $L = 2\pi r_g$. By using (5) for $S = S_g$ and $\beta = \omega$, we get:

$$\beta = \frac{2}{\pi}\frac{v}{r_g}. \tag{7}$$

Consider a terminal that enters a cell. It is necessarily in the outer zone, where it stays on average $1/(\alpha+\delta)$ seconds. Then it leaves the cell with probability $\delta/(\alpha+\delta)$ or goes in the inner zone with probability $\alpha/(\alpha+\delta)$. In the latter case it can go again in the inner zone then in the outer zone several times before leaving the cell. The dwell time in the inner zone is $1/\beta$. Let T_c be the average cell dwell time. Using the renewal theory we can write

$$T_c = \frac{1}{\alpha+\delta} + \frac{\alpha}{\alpha+\delta}\left(\frac{1}{\beta} + T_c\right), \tag{8}$$

yielding

$$T_c = \frac{1}{\delta}\left(\frac{\alpha}{\beta} + 1\right). \tag{9}$$

The cell crossing rate is given by $1/T_c$ and can be computed with (5). As for an hexagon $S = 3\sqrt{3}r^2/2$ and $L = 6r$, we have:

$$\frac{1}{T_c} = \frac{4}{\pi\sqrt{3}}\frac{v}{r}. \tag{10}$$

Given r_g, r, v, it is possible to compute α, β and δ very simply by solving a system of equations:

$$\begin{cases} \beta = \frac{2}{\pi} \, \frac{v}{r_g} \\ \frac{\alpha}{\beta} = \frac{S_g}{S_o} \\ \delta \frac{1}{1+\frac{\alpha}{\beta}} = \frac{4}{\pi\sqrt{3}} \frac{v}{r}. \end{cases} \tag{11}$$

After a few elementary steps, we get

$$\begin{cases} \alpha = \dfrac{\frac{2}{\pi} \, \frac{v}{r_g}}{\frac{3\sqrt{3}}{2\pi} \, \frac{r^2}{r_g^2} - 1} \\ \beta = \frac{2}{\pi} \, \frac{v}{r_g} \\ \delta = \dfrac{\frac{6}{\pi^2} \, \frac{v\,r}{r_g^2}}{\frac{3\sqrt{3}}{2\pi} \left(\frac{r}{r_g}\right)^2 - 1}. \end{cases} \tag{12}$$

2.2 Traffic model

2.2.1 Rate for each zone

In this subsection we compute the minimum SINR for each zone and then deduce the achievable rate. We assume that noise is negligible compared to interference; the SINR is thus equal to the Signal to Interference Ratio (SIR). In Kelif et al. [18] propose to compute the interference by considering a continuum of interference base stations. With this approach, they got a closed formula of the SIR $f(x)$ as a function of the distance x between the terminal and the base station in an hexagonal network. The formula can be written as:

$$f(x) = 4\pi \frac{\sqrt{3}}{3} \frac{1}{\eta - 2} \frac{\left(\frac{x}{\sqrt{3}r}\right)^{\eta}}{\left(1 - \frac{x}{\sqrt{3}r}\right)^{\eta-2}} \tag{13}$$

where η is the propagation exponent and r is the cell radius.

The rate $g(x)$ achievable at distance x is computed according to the Shannon formula:

$$g(x) = B \log_2 (1 + f(x)) \tag{14}$$

where B is the bandwidth used for the transmission. In the inner zone, the SIR is equal to or higher than $f(r_g)$. We consider that the rate in the whole zone is given by $g(r_g)$. The maximum rate R is given when there is only one terminal in the inner zone (and none in the outer zone). In that case, B is the system bandwidth. We thus have:

$$R = g\left(r_g\right) = B \log_2 \left(1 + \frac{4\pi\sqrt{3}}{3(\eta - 2)} \frac{\left(\frac{r_g}{\sqrt{3}r}\right)^{\eta}}{\left(1 - \frac{r_g}{\sqrt{3}r}\right)^{\eta-2}}\right). \tag{15}$$

Similarly, the rate for one terminal in the outer zone (and no terminal in the inner zone) is given by $g(r)$. Let χ be the ratio between the rates in the outer zone and in the inner zone (we have $0 < \chi < 1$):

$$\chi = \frac{g(r)}{g\left(r_g\right)} = \frac{\log_2 \left(1 + \frac{4\pi\sqrt{3}}{3(\eta-2)} \, \frac{\left(\frac{1}{\sqrt{3}}\right)^{\eta}}{\left(1-\frac{1}{\sqrt{3}}\right)^{\eta-2}}\right)}{\log_2 \left(1 + \frac{4\pi\sqrt{3}}{3(\eta-2)} \, \frac{\left(\frac{r_g}{\sqrt{3}r}\right)^{\eta}}{\left(1-\frac{r_g}{\sqrt{3}r}\right)^{\eta-2}}\right)}. \tag{16}$$

2.2.2 Practical considerations

In the model, the position of each mobile is known and the rate is a decreasing function of the distance between the terminal and the base station. The two bit-rate values are linked to geographical zones. In a real network, due to propagation phenomena (shadowing, fading), two terminals at the same distance can have different bit rates. Hence, the criterion to decide whether a DTU should be queued or not would no longer be based on the distance but on an SINR threshold. Considering these propagation phenomena is out of the scope of the paper.

2.2.3 Source model

As explained in Sect. 1, data services are modeled with transmission phases triggered by *packet calls* and reading periods [21]. Call arrivals are assumed to follow a Poisson process with rate λ. Each data transmission phase is equivalent to the transfer of an average amount of data (i.e., a file) equal to F bits. That file can be a video, a web page, etc. The size of the file is modeled as an exponential random variable.

2.3 Scheduling strategies

The transmission rate needs to be shared among all users. We adopt two schedulers as proposed in [22]: equal throughput (ET) scheduler and round robin (RR) scheduler. In this paper, we always use subscript i and j to denote the number of users being served in the inner zone and in the outer zone, respectively. We denote by $R_{g,i,j}$ and $R_{o,i,j}$ the average rate of a user in the inner zone and in the outer zone, respectively, when i users in the inner zone and j in the outer zone are served. Note that all users in the same zone have the same bit rate.

2.3.1 Equal throughput scheduler (ET)

An ET scheduler allocates the same rate to all users. It can be seen as a round robin scheduler where one bit is transmitted to each user. The time to transmit one bit to a user in the inner zone and in the outer zone is $1/R$ and $1/(\chi R)$, respectively. The total time for a round (one bit to each user) is thus $i/R + j/(\chi R)$. The rate, which is the same for all users, is therefore:

$$R_{g,i,j} = R_{o,i,j} = \frac{R}{i + \frac{j}{\chi}}. \tag{17}$$

Let $\mu_{g,i,j}$ and $\mu_{o,i,j}$ be the service rate of a user in the inner zone and in the outer zone, respectively. With ET, we have

$$\mu_{g,i,j} = \mu_{o,i,j} = \frac{R}{F(i + \frac{j}{\chi})}. \tag{18}$$

2.3.2 Round Robin scheduler (RR)

An RR scheduler allocates the same airtime to each user (but not the same rate). Thus,

$$\begin{cases} R_{g,i,j} = \frac{R}{i+j} \\ R_{o,i,j} = \frac{\chi R}{i+j}, \end{cases} \tag{19}$$

giving the service rates

$$\begin{cases} \mu_{g,i,j} = \frac{R}{F(i+j)} \\ \mu_{o,i,j} = \frac{\chi R}{F(i+j)}. \end{cases} \tag{20}$$

2.3.3 Load indicator

The resource management strategy is based on the global load of the cell. In this report, as already proposed in [11], we use the harmonic mean $R_{h,i,j}$ of the user rates as an indicator of the load. With ET, all users have the same rate and the harmonic mean is thus $\frac{R}{i+j/\chi}$. With RR, i terminals get $\frac{R}{i+j}$ and j terminals get $\frac{\chi R}{i+j}$. By definition of the harmonic mean, we have:

$$R_{h,i,j} = \frac{i + j}{\frac{i}{R_{g,i,j}} + \frac{j}{R_{o,i,j}}} \tag{21}$$

By combining (19) and (21), we easily get

$$R_{h,i,j} = \frac{R}{i + \frac{j}{\chi}}. \tag{22}$$

Note that the harmonic rate is given by (22) for both ET and RR, i.e., the load indicator does not depend on the scheduling policy.

2.4 Admission control

We propose and analyze the following admission control policy:

- in low load conditions, all packet calls are accepted and served immediately;
- in medium load conditions, packet calls made by DTUs in the outer zone are queued (within the queue length limit) and other packet calls are served immediately;
- in high load conditions, all calls are blocked.

More precisely, we define a maximum length K_b^T of the queue, and two thresholds R_q^T and R_b^T (with $R_q^T > R_b^T$) on the harmonic rate on which the admission decision will be based. Denoting by k the size of the queue at a given time, we propose that

- when $R_{h,i,j} > R_q^T$, all packet calls are served immediately;
- when $R_b^T < R_{h,i,j} \leq R_q^T$, a call made by a DTU in the outer zone is queued if $k < K_b^T$ and is blocked otherwise, while other calls are served (including calls by DTUs in the inner zone);
- when $R_{h,i,j} \leq R_b^T$, all calls are blocked.

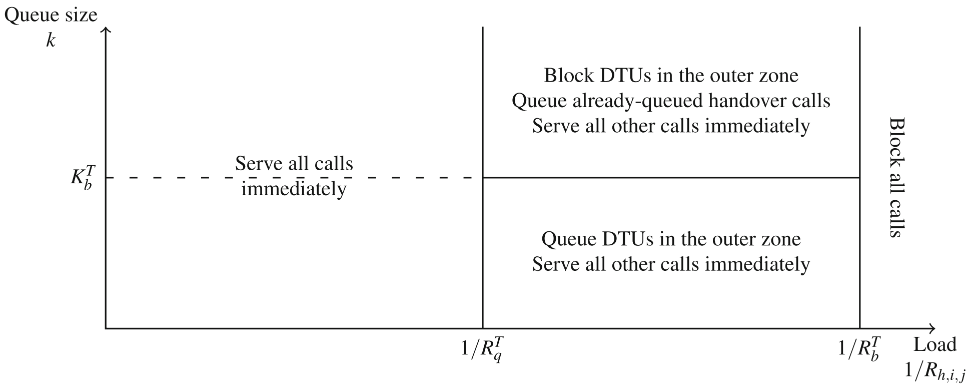
That policy is illustrated in Fig. 3, where we use $1/R_{h,i,j}$ on the horizontal axis to have an indicator that increases with the load. Note that the comparison with a threshold is a linear function of i and j. In other words, $R_{h,i,j} > R_q^T$ is equivalent to $i + j/\chi < R/R_q^T$. Similar expressions can be found for all tests.

Mobility events can happen at any time. The mobility events consist of:

- movement between the inner zone and the outer zone,
- handover of a queued call, indicating that a user with a call in the queue of a neighbor cell is entering the current cell,
- handover of an active call, indicating that a user in data transmission phase in a neighbor cell is entering the current cell.

In order to provide a good quality of service, the data transmission phase should not be interrupted. Hence, in the latter case, the call is accepted regardless of the load. Furthermore, a user who is served and who is switching from the inner to the outer zone (or vice-versa) is always kept in the system. If

Fig. 3 Illustration of the studied admission control policy


Queue size
k
K_b^T
Serve all calls immediately
Block DTUs in the outer zone
Queue already-queued handover calls
Serve all other calls immediately
Queue DTUs in the outer zone
Serve all other calls immediately
Block all calls
$1/R_q^T$
$1/R_b^T$
Load
$1/R_{h,i,j}$

the current load is below the threshold, a queued call from a neighbor cell is immediately served. In other cases, it is put in the queue of the current cell, regardless the current length of the queue. When a queued user enters the inner zone, he/she is immediately served.

For convenience, we summarize the main notations in Table 1.

3 Analytical method

3.1 Markov chain

The system is modeled by a continuous-time Markov chain with states (i, j, k), representing the respective number of calls that are served in the inner zone, served in the outer zone, and queued. This chain has an infinite number of states. In order to allow numerical processing, we keep the Markov chain with a finite number of states. We define $I_{\max}$, $J_{\max}$ and $K_{\max}$ as the upper bounds for i, j and k, respectively. We denote by Ω the set of possible values for (i, j, k) ($\Omega \subset [0, I_{\max}] \times [0, J_{\max}] \times [0, K_{\max}]$). Considering a chain with a finite number of states slightly modifies the behavior of the system. For example, according to the admission control policy, handovers are always accepted and a data transmission phase is never interrupted, while for our truncated Markov chain, a handover can be blocked when the system is in state $(i, J_{\max}, k)$. However, $I_{\max}$, $J_{\max}$ and $K_{\max}$ are chosen large enough to make such blocking probabilities negligible (typically 10^{-15}).

Due to the complexity of the Markov chain, it is not possible to represent it clearly in a figure. However, the transitions from state (i, j, k) to other states are shown in Table 2. For the sake of clarity, we consider a typical state (i, j, k), where $0 < i < I_{\max}, 0 < j < J_{\max}, 0 < k < K_{\max}$. For the extreme states, such as $i = 0$ or $j = 0$ or $k = 0$, the transition rates can also be easily derived. Table 2 is valid for ET and RR: variables $\mu_{g,i,j}$ and $\mu_{o,i,j}$ are chosen according to the scheduler selected by using either (18) or (20).

The infinitesimal generator $\mathbf{Q}$ of the Markov Chain is easily obtained from the transition rates. The stationary probability vector ` $= (\theta_{i,j,k})$ is computed from `$\mathbf{Q} = \mathbf{0}$ and $\sum_{(i,j,k)} \theta_{i,j,k} = 1$, which can be solved by a classical method.

3.2 Arrival rates of handovers of active calls and queued calls

Let λ_H be the arrival rate of handovers of active calls and λ_W the arrival rate of handovers of queued calls from DTUs. In homogeneous networks (all cells are equivalent), the departure rate of users is equal to the arrival rate. Therefore, we have

$$\lambda_H = \sum_{(i,j,k)\in\Omega} j\delta\theta_{i,j,k} \tag{23}$$

$$\lambda_W = \sum_{(i,j,k)\in\Omega} k\delta\theta_{i,j,k} \tag{24}$$

Solving the Markov chain involves solving some fixed-point system (the steady-state probabilities depend on λ_H and λ_W, that depend on the steady-state probabilities). We solve that problem numerically, using Algorithm 1, i.e., iterating those dependencies until variations are below a threshold. We know from simulation process that it usually requires from 4 to 9 iterations to solve the Markov chain.

Algorithm 1 Iteration algorithm for solving the Markov chain

```
Initialization $\lambda_H^0 = \frac{\lambda}{T_c R/F}$, $\lambda_W^0 = 0$
$\lambda_H^{new} \leftarrow \lambda_H^0$, $\lambda_W^{new} \leftarrow \lambda_W^0$
repeat
  Solve the Markov chain to obtain `
  $\lambda_H^{old} \leftarrow \lambda_H^{new}$, $\lambda_W^{old} \leftarrow \lambda_W^{new}$
  Compute $\lambda_H^{new}$ by (23) and $\lambda_W^{new}$ by (24)
until $|\lambda_H^{new} - \lambda_H^{old}| < \epsilon$ and $|\lambda_W^{new} - \lambda_W^{old}| < \epsilon$
```

Table 1 Main notations

Variable	Meaning
B	Bandwidth of the system
F	Average size of files
i	Number of users being served in the inner zone
$I_{\max}$	Maximum number of users served in the inner zone
j	Number of Users being served in the outer zone
$J_{\max}$	Maximum number of users served in the outer zone
k	Number of Users in the queue (only DTUs can be queued)
$\overline{K}$	Mean number of users in the queue (only DTU)
K_b^T	Threshold length of queue to block new DTU calls in the outer zone
$K_{\max}$	Maximum length of queue for handovers of queued DTUs
$\overline{N_c}$	Mean number of users being served in a cell
P_b	Blocking probability of packet calls
$P_{b,1}$	Blocking probability of packet calls for inner and non-DTU outer users
$P_{b,2}$	Blocking probability of packet calls for DTU outer users
R	Maximum bit rate for a user in the inner zone
$R_{g,i,j}$	Inner user bit rate when i and j users are being served in the inner and outer zones, respectively
$R_{h,i,j}$	Harmonic mean rate when i and j users are being served in the inner and outer zones, respectively
$R_{o,i,j}$	Outer user bit rate when i and j users are being served in the inner and outer zones, respectively
R_b^T	Threshold rate for blocking
R_q^T	Threshold rate for queuing DTUs in the outer zone
r	Radius of each cell
r_g	Radius of the inner zone
$\overline{T_c}$	Mean service time in the same cell (waiting time not included)
$\overline{T_s}$	Mean service time in the system (waiting time not included)
v	Average speed of users
$\overline{W_c}$	Mean waiting time of queued DTUs in a given cell
$\overline{W_s}$	Mean waiting time of queued DTUs in the system
α	Average transition rate from the outer zone to the inner zone
β	Average transition rate from the inner zone to the outer zone
δ	Average transition rate from the outer zone to neighbor cells
η	Propagation exponent
$\theta_{i,j,k}$	Steady state probability to be in state (i, j, k)
λ	Arrival rate of new calls in the cell
λ_H	Arrival rate of handovers of active calls
λ_W	Arrival rate of handovers of queued calls
χ	The ratio between the rate in the outer zone and in the inner zone
ρ	Proportion of DTUs
ω	Individual outgoing rate due to mobility for any shape

In the remainder of this section, we explain how we use the obtained steady-state distribution to derive performance measures.

3.3 Blocking probability

A new call from the inner zone or from a non-DTU in the outer zone is blocked if $R_{h,i,j} \leq R_b^T$, so the blocking probability for such users is

$$P_{b,1} = \sum\nolimits_{(i,j,k)\in\Omega_1} \theta_{i,j,k} \tag{25}$$

where $\Omega_1 = \{(i, j, k)|i + j/\chi \geq R/R_b^T, 0 \leq k \leq K_{\max}\}$.

A new call from a DTU in the outer zone is blocked if $R_{h,i,j} \leq R_b^T$ or if $R_b^T < R_{h,i,j} \leq R_q^T$ and $k \geq K_b^T$. The blocking probability is given as

$$P_{b,2} = \sum\nolimits_{(i,j,k)\in\Omega_2} \theta_{i,j,k} \tag{26}$$

Table 2 Transitions from state (i, j, k)

Event	Condition	Final state	Transition rate	Comments
A new call is made in the inner zone	If $i + j/\chi < R/R_b^T$	$(i+1, j, k)$	$\frac{\lambda\alpha}{\alpha+\beta}$	Call served
A new call is made by a non-DTU in the outer zone	If $i + j/\chi < R/R_b^T$	$(i, j+1, k)$	$\frac{\lambda\beta}{\alpha+\beta}(1-\rho)$	Call served
A new call is made by a DTU in the outer zone	If $i + j/\chi < R/R_q^T$	$(i, j+1, k)$	$\frac{\lambda\beta}{\alpha+\beta}\rho$	Call served
	Else if $k < K_b^T$	$(i, j, k+1)$	$\frac{\lambda\beta}{\alpha+\beta}\rho$	Call queued
A call is finished in the inner zone	If $i - 1 + j/\chi < R/R_q^T$	$(i-1, j+1, k-1)$	$i\mu_{g,i,j}$	A queued call served
	Else	$(i-1, j, k)$	$i\mu_{g,i,j}$	–
A call is finished in the outer zone	If $i + (j-1)/\chi < R/R_q^T$	$(i, j, k-1)$	$j\mu_{o,i,j}$	A queued call served
	Else	$(i, j-1, k)$	$j\mu_{o,i,j}$	–
A user with an active call moves from the inner to the outer zone	If $j < J_{\max}$	$(i-1, j+1, k)$	$i\beta$	–
	Else	$(i-1, j, k)$	$i\beta$	Call interruption
A user with an active call moves from the outer zone to the inner zone	If $i + (j-1)/\chi < R/R_q^T$	$(i+1, j, k-1)$	$j\alpha$	A queued call served
	Else if $i < I_{\max}$	$(i+1, j-1, k)$	$j\alpha$	–
	Else	$(i, j-1, k)$	$j\alpha$	Call interruption[a]
A user with an active call moves from the outer zone to a neighbor cell	If $i + (j-1)/\chi < R/R_q^T$	$(i, j, k-1)$	$j\delta$	A queued call served
	Else	$(i, j-1, k)$	$j\delta$	–
A user with a queued call moves from the outer zone to the inner zone	If $i < I_{\max}$	$(i+1, j, k-1)$	$k\alpha$	A queued call served
	Else	$(i, j, k-1)$	$k\alpha$	Call interruption[a]
A user with a queued call moves from the outer zone to a neighbor cell		$(i, j, k-1)$	$k\delta$	–
A user with an active call is entering the cell	If $j < J_{\max}$	$(i, j+1, k)$	λ_H	Call served
A user with a queued call is entering the cell	If $i + j/\chi < R/R_q^T$	$(i, j+1, k)$	λ_W	Call served
	Else if $k < K_{\max}$	$(i, j, k+1)$	λ_W	Call queued

[a] $I_{\max}$, $J_{\max}$ or $K_{\max}$ are large enough to ensure that the probability of such an event is negligible for the considered loads

where $\Omega_2 = \{(i, j, k) | i + j/\chi \geq R/R_b^T, 0 \leq k \leq K_{\max}$ or $R_q^T \leq i + j/\chi < R/R_b^T, K_b^T \leq k \leq K_{\max}\}$.

As users are assumed to be uniformly distributed, the overall blocking probability is

$$P_b = (1 - \rho\frac{\beta}{\alpha+\beta})P_{b,1} + \rho\frac{\beta}{\alpha+\beta}P_{b,2}. \tag{27}$$

We also compute the probability P_s that a new call from a DTU in the outer zone is served immediately (i.e., not queued):

$$P_s = \sum\nolimits_{(i,j,k)\in\Omega_s} \theta_{i,j,k} \tag{28}$$

where $\Omega_s = \{(i, j, k) | i + j/\chi < R/R_q^T, 0 \leq k \leq K_{\max}\}$.

3.4 Mean service time of users in the system

Another performance metric we are interested in is the mean service time of users in the system, which includes only the transmission time when the users are served.

Let $\overline{N_c}$ be the mean number of users being served in one cell. We have

$$\overline{N_c} = \sum\nolimits_{(i,j,k)\in\Omega} (i + j)\,\theta_{i,j,k}. \tag{29}$$

According to Little's law, the mean service time $\overline{T_c}$ in a given cell is

$$\overline{T_c} = \overline{N_c}/\lambda_{ec} \tag{30}$$

where $\lambda_{ec} = \lambda(1 - P_b) + \lambda_H$ is the arrival rate of calls[1].

Several handovers can happen during the data transmission phase. The mean service time $\overline{T_s}$ in the system is thus different from the service time $\overline{T_c}$ in a given cell. We have

$$\overline{T_s} = \sum_{n=0}^{\infty}(n + 1)p_H^n(1 - p_H)\overline{T_c} = \frac{\overline{T_c}}{1 - p_H} \tag{31}$$

[1] The blocking probability of handover is not taken into account since it is negligible.

where $p_H = \lambda_H/(\lambda+\lambda_H)$ is the handover probability during the transmission phase.

3.5 Mean waiting time of queued DTUs in the system

We use a similar approach to compute the mean waiting time of queued DTUs in the system. We first compute the equivalent arrival rate λ_{eq} of calls for the queue in the outer zone as

$$\lambda_{eq} = \lambda\rho\frac{\beta}{\alpha+\beta}(1-P_s-P_{b,2}) + \lambda_W(1-P_s) \tag{32}$$

By using Little's law, we obtain the mean waiting time of queued DTUs in the considered cell as

$$\overline{W_c} = \frac{\overline{K}}{\lambda_{eq}} \tag{33}$$

where $\overline{K}$ denotes the mean number of users in the queue and is given by

$$\overline{K} = \sum_{(i,j,k)\in\Omega} k\,\theta_{i,j,k} \tag{34}$$

We thus deduce the mean waiting time $\overline{W_s}$ of queued DTUs in the system:

$$\overline{W_s} = \sum_{n=0}^{\infty}\Big\{(n+1)[p_W(1-P_s)]^n(1-p_W)\overline{W_c} + n[p_W(1-P_s)]^{n-1}p_W P_s\overline{W_c}\Big\} = \frac{\overline{W_c}}{1-p_W(1-P_s)} \tag{35}$$

where $p_W = \lambda_W/(\lambda_W+\lambda\rho\frac{\beta}{\alpha+\beta})$ is the handover probability of queued DTUs.

4 Numerical results

The parameter values are listed in Table 3. The cell radius is 0.5 km and the inner-zone radius is 0.3215 km. This gives a proportion of users in the inner zone $\alpha/(\alpha+\beta) = 0.5$. The propagation exponent is assumed to be 3.3 and the bandwidth is 10 MHz. With (15) and (16), this gives a maximum speed rate $R = 16$ Mbit/s and a ratio between the bit rate in the outer zone and the rate in the inner zone $\chi = 0.25$. In addition, λ is chosen from 0.5 to 1.6 with interval 0.1.

Table 3 Parameter values

Symbol	Value	Symbol	Value
B	10 MHz	η	3.3
v	1 m/s	ρ	0.2, 0.5, 0.8
r	0.5 km	$I_{\max}$	30
r_g	0.3215 km	$J_{\max}$	20
F	4 Mbit	$K_{\max}$	25
R_q^T	4 Mbit/s	K_b^T	10, 15, 20
R_b^T	0.75 Mbit/s	λ	0.5–1.6 call/s

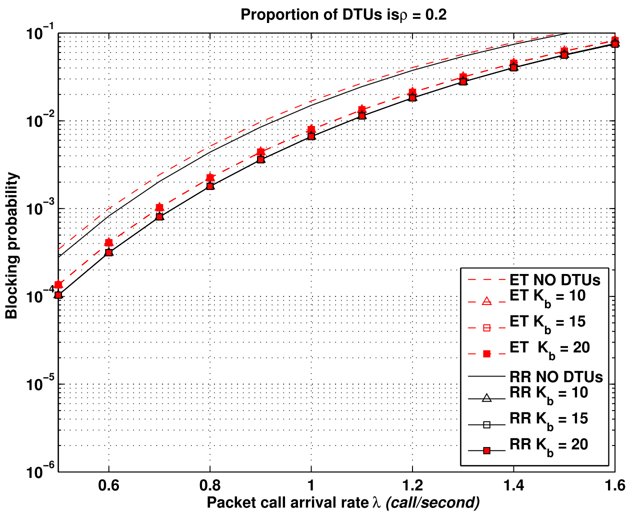


Fig. 4 Blocking probability of RR and ET schedulers under $\rho = 0.2$, $K_b^T = 10, 15, 20$

4.1 Blocking probability

4.1.1 Blocking probability against packet call arrival rate

The blocking probability against the packet call arrival rate is shown respectively in Figs. 4 and 5 for $\rho = 0.2$ and 0.5. From both figures, we can see that the blocking probability for the ET scheduler is very close to the one for the RR scheduler. In addition, when ρ is small, the blocking probability is insensitive to the increase of queue length K_b^T. Only when ρ is larger, the impact of K_b^T is obvious, see Fig. 6. Comparing Figs. 4, 5 and 6, we can find that the blocking probability is lower if K_b^T is larger but having a large queue provides a noticeable gain only if the proportion of DTUs is high and at medium load ($0.8 \le \lambda \le 1.4$), see Fig. 6.

4.1.2 Blocking probability against velocity of users

In Sect. 4.1.1, we fix the velocity of users to be 1 m/s and obtain the results of blocking probability against packet call arrival rate under different ρ. In this section, we aim to obtain the impact of users' velocity on blocking probability and set

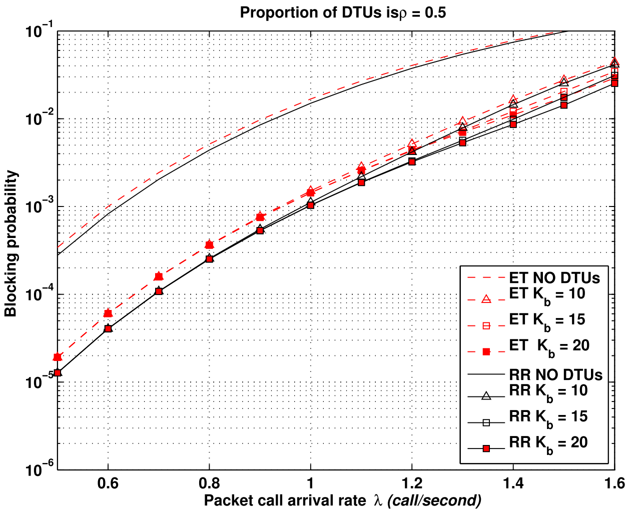


Fig. 5 Blocking probability of RR and ET schedulers under $\rho = 0.5$, $K_b^T = 10, 15, 20$

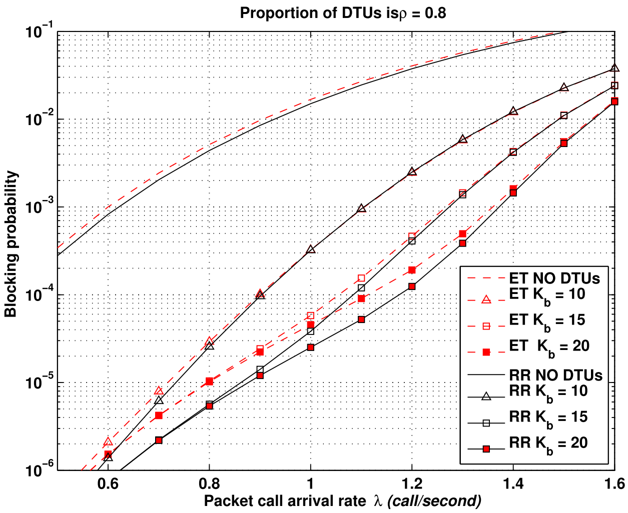


Fig. 6 Blocking probability of RR and ET schedulers under $\rho = 0.8$, $K_b^T = 10, 15, 20$

$\rho = 0.5$, $K_b^T = 20$, $v = 0.1, 1, 5, 10$ m/s respectively for both RR and ET schedulers.

The blocking probability under different velocities of users for the RR scheduler is shown in Fig. 7. We can see that the blocking probability decreases with the increase of users' velocity, but the gap is not very large. Similarly, the blocking probability for the ET scheduler is given in Fig. 8, from which it can be seen that the blocking probability decreases little with the increase of users' velocity. From both figures, we know that increasing users' velocity can relatively decrease the blocking probability, but the impact of users' velocity on blocking probability is very small.

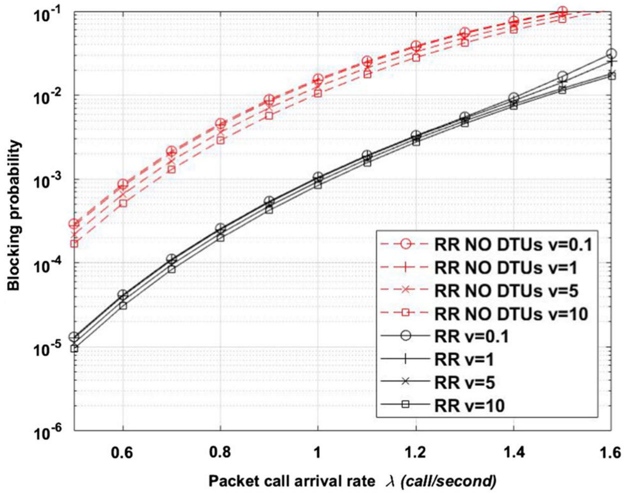


Fig. 7 Blocking probability of RR scheduler under different velocities of users

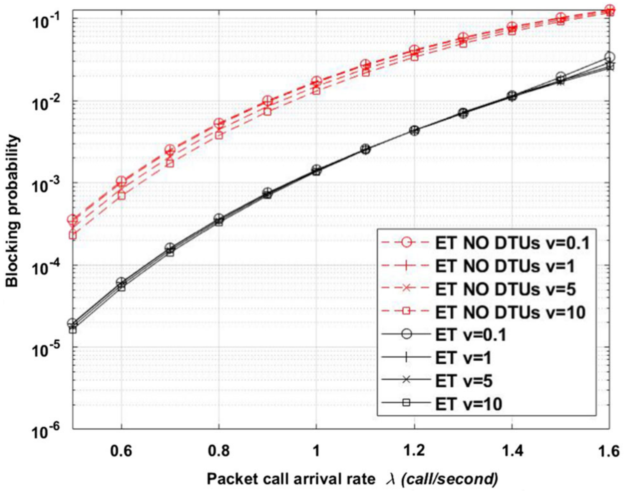


Fig. 8 Blocking probability of ET scheduler under different velocities of users

4.1.3 Capacity increase

The results of blocking probability for the RR scheduler under $K_b^T = 20$ and $\rho = 0.2, 0.5, 0.8$ are shown in Fig. 9. We can see that when K_b^T is fixed, the blocking probability decreases with the increase of ρ, as could be expected: having more delay-tolerant users allows more flexibility, and less blockings thanks to our queuing policy.

From Fig. 9, we compute the maximum arrival rates for different blocking probability targets, which represent medium to high load conditions (0.1%, 0.5%, 1%), and the relative capacity increase brought by having DTUs is shown in Table 4. The capacity increase is around 16% for $\rho = 0.2$, which is moderate, but the capacity can increase by 113% when $\rho = 0.8$ and $P_b = 0.1\%$.

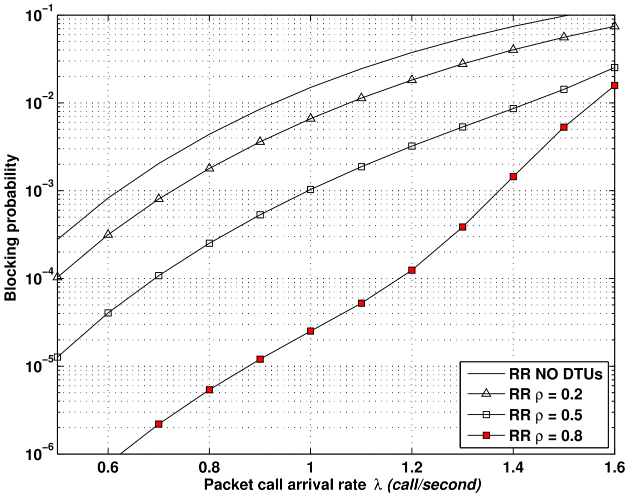


Fig. 9 Blocking probability of RR scheduler under $\rho = 0.2, 0.5, 0.8$, $K_b^T = 20$

Table 4 Capacity increase for different blocking probability targets

P_b (%)	$\rho = 0.2$ (%)	$\rho = 0.5$ (%)	$\rho = 0.8$ (%)
0.1	16.9	60.3	112.9
0.5	16.4	57.3	82.8
1	16.0	54.1	67.8

Table 5 Mean waiting time of queued DTUs for the RR scheduler under peak load in seconds

P_b (%)	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$
0.1	2.614	3.579	7.121
0.5	3.934	7.095	10.633
1	5.032	11.134	12.949

4.2 Mean waiting time of queued DTUs

The mean waiting time of queued DTUs under peak load for the RR scheduler is shown in Table 5 for different blocking probability targets. Here $v = 1$ m/s, $K_b^T = 20$. In all cases, the waiting time is less than 13 s, which is quite acceptable.

We are also interested in the impact of users' velocity on the mean waiting time of queued DTUs. We thus set $\rho = 0.5$, $K_b^T = 20$, and $v = 0.1, 1, 5, 10$ m/s to show the results for the RR and ET schedulers as in Tables 6 and 7 respectively. From both tables, we can see that the mean waiting time of queued DTUs decreases with the increase of users' velocity. This is because when the velocity of DTUs increases, the DTUs have high probability to go from the outer zone to the inner zone and the probability to be served immediately also increases, leading to the decrease of mean waiting time. In addition, comparing the third columns of Tables 5 and 6, we can see that they are consistent.

Table 6 Mean waiting time of queued DTUs for the RR scheduler with different users' velocities under peak load in seconds

P_b (%)	$v = 0.1$	$v = 1$	$v = 5$	$v = 10$
0.1	3.711	3.579	3.142	2.768
0.5	7.681	7.095	5.494	4.484
1	12.068	11.134	7.988	6.098

Table 7 Mean waiting time of queued DTUs for the ET scheduler with different users' velocities under peak load in seconds

P_b (%)	$v = 0.1$	$v = 1$	$v = 5$	$v = 10$
0.1	3.261	3.151	2.774	2.446
0.5	6.275	5.820	4.578	3.762
1	9.818	8.915	6.339	4.910

Table 8 Mean service time of users for the RR scheduler under peak load in seconds

P_b (%)	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	NO DTUs
0.1	1.017	1.048	1.023	1.011
0.5	1.234	1.272	1.081	1.224
1	1.371	1.402	1.099	1.353

Table 9 Mean service time of users for the RR scheduler with different users' velocities under peak load in seconds

P_b (%)	$v = 0.1$	$v = 1$	$v = 5$	$v = 10$
0.1	1.048	1.048	1.050	1.051
0.5	1.269	1.272	1.271	1.271
1	1.389	1.402	1.413	1.410

4.3 Mean service time of users

The mean service time of users for the RR scheduler under peak load is shown in Table 8. Here $v = 1$ m/s, $K_b^T = 20$. The maximum mean service time under these settings is about 1.4 s and is obtained for $\rho = 0.5$ and $P_b = 1\%$. As the mean size of files is 4 Mbit, the average bit rate is $4/1.4 = 2.86$ Mbit/s, which is acceptable. Though the mean waiting time is higher for $\rho = 0.8$ (compared with $\rho = 0.5$), the mean service time is relatively lower. This is a very interesting benefit of introducing DTUs: by limiting the number of users in the system who have a low rate, the average rate is higher, which is beneficial for both the system and users.

To show the impact of users' velocity on the mean service time of users, we set $\rho = 0.5$, $K_b^T = 20$, and $v = 0.1, 1, 5, 10$ m/s. The results for the RR and ET schedulers are shown in Tables 9 and 10 respectively. From both figures, we can see that the mean service time of users under different users' velocities changes slightly, which means that the impact of users' velocity on the mean service time is very

Table 10 Mean service time of users for the ET scheduler with different users' velocities under peak load in seconds

P_b (%)	$v = 0.1$	$v = 1$	$v = 5$	$v = 10$
0.1	1.194	1.192	1.185	1.178
0.5	1.532	1.527	1.506	1.488
1	1.737	1.738	1.709	1.680

small. In addition, by comparing the third column of Tables 8 and 9, we can see that they are consistent.

5 Conclusion

In this paper, we have proposed a model to analyze the benefit of having DTUs in cellular networks. We divide a cell into the inner zone where users have a higher SINR and the outer zone where users have a low SINR. When a packet call is from a DTU in the outer zone, it is queued if the current load is above a threshold. When the DTU moves to the inner zone or the load is below the threshold, its call is served. We then analyze the impact of such a policy on the system capacity. Numerical results show that when 20% of calls are DTUs, the capacity of cellular networks can increase about 16%, and when there is 80% DTUs, the capacity can even increase by 113% without affecting the blocking rate.

We have computed the mean waiting time, which is a first QoS indicator. However, users are generally sensitive to the occurrence of long delays. A next step of this work is thus to compute either the distribution of the waiting time or its standard deviation to estimate the last decile. Another extension is to consider an impatience threshold and to serve a user as soon as the experienced delay is higher than a threshold. Another possible extension of this work is to analyze the management of DTUs in a system that combines cellular networks and WiFi.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Jonsson, P., & Carson, S. (2018). Ericsson mobility report.
2. Yeh, S. P., Talwar, S., Wu, G., Himayat, N., & Johnsson, K. (2011). Capacity and coverage enhancement in heterogeneous networks. *IEEE Wireless Communications*, *18*(3), 32–38.
3. Falowo, O. E., & Chan, H. A. (2011). Effect of mobile terminal heterogeneity on call blocking/dropping probability in cooperative heterogeneous cellular networks. *Telecommunication Systems*, *47*(3–4), 337–349.
4. Ghosh, A., Mangalvedhe, N., Ratasuk, R., Mondal, B., Cudak, M., Visotsky, E., et al. (2012). Heterogeneous cellular networks: From theory to practice. *IEEE Communications Magazine*, *50*(6), 54–64.
5. Li, P., Gong, S., Gao, S., Hu, Y., Pan, Z., & You, X. (2019). Delay-constrained sleeping mechanism for energy saving in cache-aided ultra-dense network. *Science China Information Sciences*, *62*(8), 82301.
6. Pramanik, P. K. D., & Choudhury, P. (2020). Mobility-aware service provisioning for delay tolerant applications in a mobile crowd computing environment. *SN Applied Sciences*, *2*(3), 1–17.
7. Balasubramanian, A., Mahajan, R., & Venkataramani, A. (2010). Augmenting mobile 3G using WiFi. In *Proceedings of the ACM MobiSys* (pp. 209–222), San Francisco, California, USA, June 15–18.
8. Dimatteo, S., Hui, P., Han, B., & Li, V.O.K. (2011) Cellular traffic offloading through WiFi networks. In *Proceedings of the IEEE MASS* (pp. 192–201).
9. Trestian, I., Ranjan, S., Kuzmanovic, A., & Nucci, A. (2012). Taming the mobile data deluge with drop zones. *IEEE/ACM Transactions on Networking*, *20*(4), 1010–1023.
10. Hagos, D. H. (2016). The performance of network-controlled mobile data offloading from LTE to WiFi networks. *Telecommunication Systems*, *61*(4), 675–694.
11. Bonald, T., & Proutière, A. (2003). Wireless downlink data channels: User performance and cell dimensioning. In *Proceedings of the ACM MobiCom* (pp. 339–352).
12. Munoz, E. A. C., Le Denmat, F., Morin, A., & Lagrange, X. (2015). Multimedia content delivery trigger in a mobile network to reduce the peak load. *Annals of telecommunications-annales des télécommunications*, *70*(7–8), 321–330.
13. Gamboa, S., Pelov, A., Maillé, P., Lagrange, X., & Montavont, N. (2017). Reducing the energy footprint of cellular networks with delay-tolerant users. *IEEE Systems Journal*, *11*(2), 729–739.
14. Baloch, R. A., Awan, I., & Min, G. (2010). A mathematical model for wireless channel allocation and handoff schemes. *Telecommunication Systems*, *45*(4), 275–287.
15. Fazio, P., Tropea, M., Sottile, C., Marano, S., Voznak, M., & Strangis, F. (2014). Mobility prediction in wireless cellular networks for the optimization of call admission control schemes. In *Proceedings of 27th IEEE Canadian conference on electrical and computer engineering (CCECE)* (pp. 1–5). IEEE.
16. Rappaport, T. S. (2014). *Wireless communications principles and practice* (2nd ed.). Upper Saddle River: Prentice Hall.
17. Guérin, R. A. (1987). Channel occupancy time distribution in a cellular radio system. *IEEE Transactions on Vehicular Technology*, *36*(3), 89–99.
18. Kelif, J.-M., Coupechoux, M., & Godlewski, P. (2007). Spatial outage probability for cellular networks. In *Proceedings of IEEE GLOBECOM* vol. 522 (pp. 4445–4450).
19. Thomas, R., Gilbert, H., & Mazziotto, G. (1988). Influence of the moving of the mobile stations on the performance of a radio mobile cellular network. In *Proceedings of the Nordic seminar on digital land mobile radio communications*.
20. Munir, K., Lagrange, X., Bertin, P., Guillouard, K., & Ouzzif, M. (2015). Performance analysis of mobility management architectures in cellular networks. *Telecommunication Systems*, *59*(2), 211–227.
21. Mahmoud, H. K., Coupechoux, M., & Godlewski, P. (June 2009). Traffic studies for DSA policies in a simple cellular context with packet services. In *Proceedings of CROWNCOM*.
22. Berggren, F., & Jäntti, R. (2004). Asymptotically fair transmission scheduling over fading channels. *IEEE Transactions on Wireless Communications*, *3*(1), 326–336.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Feng Yan received the B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 2005, the M.S. degree from Southeast University, Nanjing, China, in 2008, and the Ph.D. degree from Telecom ParisTech, Paris, France, in 2013, all in electrical engineering. From November 2013 to April 2015, he was a post-doctoral researcher in Telecom Bretagne (now IMT Atlantique), Rennes, France. He is currently an Associate Professor in the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China. His current research interests are in the areas of wireless sensor networks, UAV networks and IoT networks.

Patrick Maillé graduated from Ecole polytechnique and Telecom Paris, France. He has been with IMT Atlantique (formerly Telecom Bretagne) since 2002, where he obtained the Ph.D. in applied mathematics in 2005 and is now a professor. His research interests are in all economic aspects of telecommunication networks, from pricing schemes at the user level, to auctions for spectrum and regulatory issues (net neutrality, search neutrality). He authored or co-authored more than 100 research contributions on those topics, including the book “Telecommunication Network Economics”, published by Cambridge University Press in 2014.

Xavier Lagrange received the master of engineering degree from Ecole Centrale Paris, France, in 1984 and the Ph.D. degree from Telecom Paris in 1998. He is professor in “Network Systems, Cybersecurity and Digital Law” department at IMT Atlantique (Rennes, France) and leads research group “Advanced Technologies for Operated Networks” at IRISA. He co-authored more than 180 publications. His domain of interest includes resource allocation, medium access control and performance analysis for 5th and 6th generation cellular networks.