Modelling Load Balancing and Carrier Aggregation in Mobile Networks

F. Bénézit, S. Elayoubi, R-M. Indre, A. Simonian,
Orange Labs, Issy-les-Moulineaux, France

***Abstract*—In this paper, we propose analytical models to derive the performance of dual carrier mobile HSDPA mobile networks. Specifically, we analyze the flow-level performance of two inter-carrier load balancing schemes and the gain engendered by Carrier Aggregation (CA). CA is one of the most important features of HSPA+ networks; it allows devices to be served simultaneously by several carriers. We propose Volume Balancing (VB) and Join the Fastest Queue (JFQ), a load-balancing mechanism that allows the traffic of non-CA users to be distributed over the aggregated carriers. We then evaluate the performance of both CA and non-CA users by means of analytical modeling. We show that the proposed schemes achieve efficient load balancing. We also investigate the impact of mixing traffic of CA and non-CA users in the same cell and show that performance is practically insensitive to the traffic mix.**

I. Introduction

One of the most important features of High Speed Packet Access (HSPA) and Advanced Long Term Evolution (LTE-A) is *Carrier Aggregation* (CA) which allows users to be served simultaneously by two or more carriers. The HSPA spectrum is divided into carriers of 5 MHZ each, while the LTE spectrum is divided into carriers with sizes ranging from 1.4 MHZ to 20 MHZ. The classical way of managing these carriers is to consider an independent scheduler per carrier. This type of resource management, however, may result into inefficiencies due to load discrepancies between carriers. To cope with this inefficiency, HSPA+ has defined carrier aggregation. Specifically, Release 8 introduces the Dual Carrier (DC) feature where two carriers are aggregated on the frequency band of 2.1 GHZ. Release 10 has extended this concept to the aggregation of a carrier on 2.1 GHZ with another carrier on the 900 MHZ band; this is the Dual Band (DB) HSPA feature in which the two carriers can have significantly different capacities due to the difference of propagation conditions between the 900 MHZ and the 2.1 GHZ bands. In LTE-A, it is possible to aggregate two or more carriers of different sizes, leading to large capacity difference between carriers.

It is widely agreed that carrier aggregation brings gain in the following two aspects. First, the user peak rates are substantially increased (e.g., doubled for DC devices). However, this is only true at low load where the user is almost always alone in the cell. In this paper, we focus mainly on DC and DB technologies and investigate the *actual* throughput gain of CA users when the carrier capacity is dynamically shared by several mobile users. The second advantage is joint scheduling or *pooling*. Indeed, since a single scheduler is used for two or more carriers, the scheduler can implement intelligent load balancing schemes so as to equalize the traffic over the carriers and thus increase the traffic capacity. In the present paper, we propose two load balancing schemes and evaluate their performance in terms of traffic capacity and mean flow throughput.

* This work has been partially financed by the ANR IDEFIX project.

A. Related work

Quantifying the gain of carrier aggregation in downlink HSPA (HSDPA) has been the object of several papers. Authors in [1] have shown that DC HSDPA Release 8 doubles the throughput of dual carrier users only at low network load. The authors also show that even at high network load, this feature still yields considerable gain when compared to single carrier HSPA Release 7. In [2], the gains expected from frequency diversity and higher multi-user diversity have been evaluated; these gains are used as inputs for the capacity model used in the present paper. Simulation results are provided in [3] and [4] for full-buffer traffic, i.e., the number of users in the system is fixed and these users have an infinite amount of data to transmit. However, the trunking gain cannot be observed for full buffer traffic, as explained in [2]. This gain has been assessed in [5]; results show that we have a 45% improvement in DC-HSDPA capacity. The latter could exceed 160% if Dual Carrier is combined to MIMO (DC HSDPA Release 9).

Another set of works has focused on carrier aggregation in the uplink. The feasibility of dual carrier in uplink HSPA (HSUPA) has been studied in [6]. It is shown that, in cells with a relatively large radius, cell edge users can barely benefit from their dual carrier capabilities. The pooling gain for DC HSUPA has been studied in [7], where it has been shown that, due to power limitations at the user device, the gain from carrier aggregation is lower in the uplink than in the downlink.

In LTE-A, carrier aggregation has been initially designed in order to allow the extension of the offered bandwidth by transmitting over multiple carriers, so that, from a user's point of view, the aggregated carriers are seen as a single, large carrier [8], [9]. System level simulations in [10] show that the carrier aggregation gain is large at low load and decreases when the number of scheduled users becomes large. The same trend is shown in [11], also using system level simulation; it is also observed that the full buffer traffic model yields an over-estimation of the carrier aggregation gain. Paper [12] confirms this trend in the framework of inter-band carrier aggregation considering carriers on 800 MHZ and 2.1 GHZ frequency bands.

978-3-901882-63-0/2014 - Copyright is with IFIP

B. Contribution

The contribution of this paper is twofold. First, we propose two inter-carrier load balancing schemes that allow to efficiently split incoming CA and non-CA traffic over the two carriers. We then analyse the performance of the proposed schemes by modelling the cell occupancy by means of a Markov process. Based on this analytical model, we show that the proposed schemes achieve ideal load balancing over the aggregated carriers.

As discussed in the previous section, most of the existing work on carrier aggregation uses system level simulation based on the full-buffer traffic model in order to assess the performance gains of carrier aggregation. While the full-buffer model allows to accurately model the lower layers and to estimate the impact of physical effects such as path loss, shadowing and fast fading, it does not represent a realistic traffic model since it does not account for the dynamic behaviour of users. There is thus a need for analytical models that capture the flow-level dynamics and help understand the system performance. This paper fills this gap by proposing analytical models that allow to evaluate the system performance while accounting for the following essential features:

- Aggregation of carriers of different capacities, which is the case of DB HSDPA
- A mix of CA capable devices and non-CA devices. This includes the case of legacy devices that do not support CA coexisting with CA devices and also the case of users that are covered by only one of the carriers (e.g., an indoor user that is covered by the 900 MHZ band only)
- The interaction of the load balancing feature with the CA. Indeed, legacy users do not select randomly the serving Carrier as it is supposed in [13], but take into account the rate that they will get on both carriers before making a decision.

II. System description

In this section, we present the considered network architecture and briefly describe the scheduling schemes proposed for *Single Carrier* (SC) and *Dual Carrier* (DC) users.

A. Radio conditions

We consider the downlink of a cell equipped with two carriers. The capacity of each carrier is time-shared between active users. Depending on their mobile subscription, there are two types of users in the cell: SC users, which are allowed to use only one of the two carriers and DC users, which are allowed to simultaneously use both carriers. The resources of the two carriers are shared using a single joint scheduler, as explained in the next section.

Let C_1 and C_2 denote the peak data rates at which a single user may transmit on Carrier 1 and Carrier 2 in the vicinity of the base station. This peak bit rate depends on the radio environment, the bandwidth and the coding efficiency. Typically, the peak data rate varies over time due to user mobility, shadowing and multipath reflections. As the aim of this paper is not to model these physical effects, we simply consider that there are J areas in the cell, with area A_j characterized by throughputs $C_{1,j}$ and $C_{2,j}$ corresponding to the maximum rate that can be attained on Carrier 1 and Carrier 2 in area A_j when a single user is present in the cell. Furthermore, we assume that the position of active users remains constant in time and that the probability of being in area j is equal to q_j. A classical assumption is to consider a cell divided into concentric rings and that users are uniformly distributed over the cell surface. Let R denote the total cell radius. The probability of being in area A_j is thus given by

$$q_j = \frac{|A_j|}{\pi R^2} \tag{1}$$

where $|A_j| = \pi(r_j^2 - r_{j-1}^2)$ is the area of ring j (r_j is the radius of ring j and by convention, $r_0 = 0$ and $r_J = R$).

B. Scheduling schemes

The scheduler must first decide how to assign SC and DC users to one or both carriers and then share the capacity of each carrier among the different users.

Upon the arrival of a new SC user, the scheduler decides which of the two carriers should serve this user for its entire session; and upon the arrival of an incoming DC flow, the scheduler must decide how much of the volume of that flow should be treated by each carrier, given that DC users are served by both carriers. We assume that this volume allocation can vary during the DC session.

Once these scheduling decisions are made, the scheduler divides the occupation time of each carrier fairly among the active users present on each carrier. Specifically, if n_1 users are present on Carrier 1, each user in area A_j is allocated a fraction $C_{1,j}/n_1$ of the total carrier rate $C_{1,j}$; this corresponds to a *Processor Sharing* (PS) service discipline in which the server capacity is equally shared among active users on that Carrier in area A_j. Similarly, if n_2 users are present on Carrier 2, each user in area A_j is allocated a fraction $C_{2,j}/n_2$.

We propose SC and DC scheduling schemes that maximize the immediate rate of each user in order to study a *best-case scenario*.

1) Scheduling SC users: We propose to schedule single carrier users via the *Join the Fastest Queue* (JFQ) scheduling policy, where the incoming SC flow is assigned to the carrier which would provide this flow with the smallest completion time, or equivalently with the largest throughput. Specifically, assume that Carriers 1 and 2 are currently serving n_1 and n_2 flows, respectively. Under the assumption of fair sharing, the carrier that provides the smallest completion time is the one that ensures the largest ratio among $C_{1,j}/(n_1 + 1)$ and $C_{2,j}/(n_2 + 1)$. Note that the completion times are computed only on the basis of the state of the system at the arrival time of the SC customer; system state changes (i.e., arrival or departure of other flows) that may occur during the transmission of the considered flow are thus not (and cannot be) taken into account in the scheduling scheme.

2) Scheduling DC users: Suppose that, at a given moment, the base station has an amount of data σ_0 that is relative to a given DC user in its buffer. The scheduler must decide how to split the volume σ_0 between the two carriers. Let σ_1 and σ_2 denote the volumes of that flow respectively transferred by Carrier 1 and Carrier 2 (with $\sigma_1 + \sigma_2 = \sigma_0$), and let T_1 and T_2 be the associated completion times. Ideally, volume σ_0 should be split such that the transfer of volumes σ_1 and σ_2 is completed simultaneously on both carriers. In fact, the actual completion time of the DC flow is determined by the slowest of the two carriers, i.e., $T = \max(T_1, T_2)$. To minimize T, the total volume must be split such that the completion times on the two carriers are identical, i.e., $T_1 = T_2$. In the following, we will propose a scheduling policy that minimizes T and ensures simultaneous completion times. It will be referred to as the *Volume Balancing* (VB) and will be formally defined in Section III-A2.

In practice, perfect Volume Balancing may be difficult to implement due to rapidly changing rates caused by fast fading. We therefore propose the following simple implementation. The scheduler feeds carriers with one new frame as soon as either carrier finishes serving a frame. Then, both carriers will finish serving their last frame approximately at the same time, that is, when the scheduler's buffer turns empty. Indeed, using this *frame-per-frame scheduler*, the completion time of the two carriers can differ by at most one frame transmission time, which is negligible at the flow time-scale. In the following, we will study the latter frame-per-frame scheduler, assuming that the data is perfectly split among carriers.

III. System model

In this section, the network model is presented. We first specify the proposed scheduling policies, i.e. Join the Fastest Queue and Volume Balancing, in terms of system parameters. We then model the system by means of a Markov process and discuss system stability and throughput performance.

A. Scheduler modeling

Given the fair sharing assumption (cf. §II-B), the system can be modeled by two multiclass PS queues representing the two carriers, as depicted in Figure 1. At any time t, the state of the system is defined by the $3J$-dimensional vector

$$\mathbf{n}(t) = (n_{1,1}(t), n_{2,1}(t), m_1(t); \ldots; n_{1,J}(t), n_{2,J}(t), m_J(t))$$

where

- $n_{1,j}(t)$ is the number of SC users in area A_j currently transmitting on Carrier 1,
- $n_{2,j}(t)$ is the number of SC users in area A_j currently transmitting on Carrier 2,
- $m_j(t)$ is the number of DC users in area A_j currently transmitting on Carriers 1 and 2.

The fact that a single variable is sufficient to characterize DC flows in both queues will be justified at the end of this section. For $1 \le j \le J$, let $\mathbf{e}_{1,j}$, $\mathbf{e}_{2,j}$ and $\mathbf{e}_{3,j}$, denote the unit vectors such that $\mathbf{e}_{k,j}$ has the $(3j+k-1)$-th element equal to 1, while all other elements are 0. Finally, define the rate functions $D_{1,j}$ and $D_{2,j}$ by

$$\begin{cases} D_{1,j}(\mathbf{n}(t)) = \dfrac{C_{1,j}}{n_1(t) + m(t)}, \\[2ex] D_{2,j}(\mathbf{n}(t)) = \dfrac{C_{2,j}}{n_2(t) + m(t)} \end{cases} \tag{2}$$

where

$$n_1(t) = \sum_{j=1}^{J} n_{1,j}(t), \quad n_2(t) = \sum_{j=1}^{J} n_{2,j}(t), \quad m(t) = \sum_{j=1}^{J} m_j(t).$$

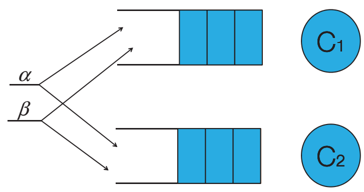

Fig. 1. *Queueing model for joint SC and DC flows.*

1) JFQ policy for SC users: As explained in Section II-B1, upon the arrival of a SC user at time t in area A_j, the user is directed towards either Carrier 1 queue or Carrier 2 queue according to the JFQ scheduling discipline. Let t_+ denote any instant close enough to time t so that no other event occurs in the system during interval $]t, t_+]$; the system state at time t_+ is then updated as follows:

i1) if $\dfrac{C_{1,j}}{n_1(t) + m(t) + 1} > \dfrac{C_{2,j}}{n_2(t) + m(t) + 1}$,

the flow is directed to Carrier 1 and $\mathbf{n}(t_+) = \mathbf{n}(t) + \mathbf{e}_{1,j}$;

i2) if $\dfrac{C_{2,j}}{n_2(t) + m(t) + 1} > \dfrac{C_{1,j}}{n_1(t) + m(t) + 1}$,

the flow is directed to Carrier 2 and $\mathbf{n}(t_+) = \mathbf{n}(t) + \mathbf{e}_{2,j}$;

i3) if $\dfrac{C_{1,j}}{n_1(t) + m(t) + 1} = \dfrac{C_{2,j}}{n_2(t) + m(t) + 1}$,

the flow is directed either to Carrier 1 or to Carrier 2 with probability $1/2$.

2) VB policy for DC users: DC flows are scheduled according to the VB policy which is defined by the following property:

for any given DC flow in any area A_j, and for any time interval $[t, t')$ where the system state $\mathbf{n}(\cdot)$ remains constant, Carrier 1 and Carrier 2 transfer volumes σ_1 and σ_2, respectively, such that

$$\begin{cases} \sigma_1 = D_{1,j}(\mathbf{n}(t))(t' - t), \\[1ex] \sigma_2 = D_{2,j}(\mathbf{n}(t))(t' - t). \end{cases} \tag{3}$$

In other words, this property ensures that any DC flow exploits the full potential of both its carrier ressources. The feasibility of Volume Balancing has been discussed in II-B2; besides, Volume Balancing satisfies the following interesting properties.

Remark 1. *The VB policy ensures that the transfer of any DC flow on Carriers 1 and 2 is completed at the same time on both carriers.*

In fact, as soon as a DC flow exists at time t, the VB policy ensures that *both* carriers are serving the flow. Therefore, they finish serving simultaneously.

Remark 2. *In the interval $[t, t')$ defined above, frames are served proportionally to rates, that is,*

$$t' - t = \frac{\sigma_1}{D_{1j}(\mathbf{n}(t))} = \frac{\sigma_2}{D_{2j}(\mathbf{n}(t))} = \frac{\sigma_1 + \sigma_2}{E_j(\mathbf{n}(t))} \tag{4}$$

where

$$E_j(\mathbf{n}(t)) = D_{1j}(\mathbf{n}(t)) + D_{2j}(\mathbf{n}(t)). \tag{5}$$

This readily follows from Eq. (3) and simple algebra. The last equality in Eq. (4) illustrates that the system behaves as a single carrier of rate $E_j(\mathbf{n}(t))$ for the entire DC volume $\sigma_1 + \sigma_2$.

Given that DC flows arrive and depart simultaneously in both queues, a single state variable can be therefore used to characterize the number of DC flows in the system, as claimed in the beginning of this section.

B. A Markov process

Assume that SC and DC flows arrive in the system according to Poisson processes with respective arrival rates α and β, and that the volume of data of any (SC or DC) user is exponentially distributed with mean σ. We denote by $\Lambda = \alpha + \beta$ the total arrival rate in the cell. We further assume that the traffic is uniformly distributed over the cell such that in area A_j, SC and DC flows arrive with intensities $\alpha_j = \alpha q_j$ and $\beta_j = \beta q_j$, with probability q_j introduced in (1).

Based on the assumptions of exponential flow size distribution and Poisson flow arrivals and given conditions ***i1***, ***i2*** and ***i3*** defined in Section III-A1, the system state described by vector $\mathbf{n}(t)$, $t \geq 0$, defines a Markov process which can move in an infinitesimal time interval from state $\mathbf{n}$ to state:

T1) $\mathbf{n} + \mathbf{e}_{1j}$ with transition rate $\alpha_j \mathbf{1}_{i1} + \frac{\alpha_j}{2}\mathbf{1}_{i3}$

T2) $\mathbf{n} + \mathbf{e}_{2j}$ with transition rate $\alpha_j \mathbf{1}_{i2} + \frac{\alpha_j}{2}\mathbf{1}_{i3}$

T3) $\mathbf{n} + \mathbf{e}_{3j}$ with transition rate β_j

T4) $\mathbf{n} - \mathbf{e}_{1j}$ with transition rate $n_{1j}D_{1j}(\mathbf{n})/\sigma$, since Carrier 1 behaves as a PS queue with departure rate proportional to the number n_{1j} of SC users in A_j

T5) $\mathbf{n} - \mathbf{e}_{2j}$ with transition rate $n_{2j}D_{2j}(\mathbf{n})/\sigma$, since Carrier 2 behaves as a PS queue with departure rate proportional to the number n_{2j} of SC users in A_j

T6) $\mathbf{n} - \mathbf{e}_{3j}$ with transition rate $r_j(\mathbf{n}) = m_j E_j(\mathbf{n})/\sigma$, where $E_j(\mathbf{n})$ is defined in Eq. (5).

To justify the value of $r_j(\mathbf{n})$, recall from Remark 2 that the completion of a DC flow ends simultaneously on both carriers, which behave as one single carrier of rate $E_j(\mathbf{n})$.

C. System stability

We now address the system stability defined as the ergodicity of the Markov process $(\mathbf{n}(t))_{t\geq 0}$. We start by considering the case in which radio conditions are uniform over the entire cell so that $J = 1$ and the capacity of each carrier is simply $C_{1,1} = C_1$ and $C_{2,1} = C_2$.

Proposition 1. *Defining the system load (with single area) by*

$$\rho = \frac{(\alpha + \beta)\sigma}{C_1 + C_2}, \tag{6}$$

the stability condition is $\rho < 1$ (for the above defined service disciplines).

Proof. A *necessary* condition for stability for a single area cell is that the total traffic intensity $(\alpha + \beta)\sigma$ must be less than the total cell capacity $C_1 + C_2$, that is, $\rho < 1$ with load ρ defined in (6).

To prove that $\rho < 1$ is also a *sufficient* stability condition, we can first invoke a fluid limit approach ([16], Corollary 9.8). Using the transition rates expressed in Section III-B for $J = 1$, the fluid limit associated with process $(\mathbf{n}(t))_{t\geq 0}$ is then defined by the differential equations

$$\begin{cases} \dfrac{dn_1}{dt} = \alpha \mathbf{1_{i1}} + \dfrac{\alpha}{2}\mathbf{1_{i3}} - \dfrac{C_1 n_1}{\sigma(n_1 + m)}\mathbf{1}_{n_1+m>0}, \\ \dfrac{dn_2}{dt} = \alpha \mathbf{1_{i2}} + \dfrac{\alpha}{2}\mathbf{1_{i3}} - \dfrac{C_2 n_2}{\sigma(n_2 + m)}\mathbf{1}_{n_2+m>0}, \\ \dfrac{dm}{dt} = \beta - \left(\dfrac{C_1 \mathbf{1}_{n_1+m>0}}{n_1 + m} + \dfrac{C_2 \mathbf{1}_{n_2+m>0}}{n_2 + m}\right)\dfrac{m}{\sigma}, \end{cases} \tag{7}$$

where conditions ***i1***, ***i2*** and ***i3*** are introduced in Section III-A1. The process $(\mathbf{n}(t))_{t\geq 0}$ is then ergodic if the associated fluid limit is stable in the sense that, starting from any initial state, the sum $n_1 + n_2 + m$ reaches state $(0, 0, 0)$ in a finite time. Assuming that $n_1 + m > 0$ and $n_2 + m > 0$ at any time, and summing all three equations (7) side by side, we obtain $\frac{d}{dt}(n_1 + n_2 + m) = \alpha + \beta - (C_1 + C_2)/\sigma$ which is a negative constant as long as $\rho < 1$. This ensures that $n_1 + n_2 + m$ is a decreasing function of time, and that the system empties in a finite time.

If either $n_1 + m = 0$ or $n_2 + m = 0$ at some time, however, the latter argument no longer applies. To take these boundaries into account, we can invoke results for the ergodicity of Markov processes based on Lyapounov functions ([16], Prop. 8.14). In fact, the fluid approach invoked above shows that $L_0 : \mathbf{n} \mapsto n_1 + n_2 + m$ is a natural Lyapounov function for system (7) in the region $n_1 + m > 0$, $n_2 + m > 0$; we therefore introduce the "weighted" Lyapounov function L defined by $L(\mathbf{n}) = v_1(\mathbf{n})n_1 + v_2(\mathbf{n})n_2 + w(\mathbf{n})m$ for all $\mathbf{n} = (n_1, n_2, m) \in \mathbb{N}^{3+}$, where

$$v_1(\mathbf{n}) = \mathbf{1}_{T_1 > T_2} + f\left(\frac{T_1}{T_1 + T_2}\right)^{\delta}\mathbf{1}_{T_1 \leq T_2},$$

$$v_2(\mathbf{n}) = v_1(n_2, n_1, m), \quad w(\mathbf{n}) = \min(v_1(\mathbf{n}), v_2(\mathbf{n})),$$

with $T_1 = \frac{n_1+m+1}{C_1}$, $T_2 = \frac{n_2+m+1}{C_2}$, $f(x) = 4x(1-x)$ and some constant $\delta > 0$. Let $\mathcal{Q}$ denote the transition operator of process $(\mathbf{n}(t))_{t\geq 0}$, defined by the transition rates expressed in

Section III-B for $J = 1$; simple calculations then show that for some small enough $\delta > 0$ and with $\rho < 1$, there exist constants $\gamma > 0$ and $K > 0$ such that $\mathcal{Q}(L)(\mathbf{n}) \leq -\gamma$ when $L(\mathbf{n}) > K$ and the set $\{\mathbf{n},\ L(\mathbf{n}) \leq K\}$ is finite. This ensures that inequality $\rho < 1$ is a sufficient condition for stability. □

Let us now analyze the system stability in the more general case with $J > 1$ different areas in the cell, each with its own radio conditions.

Proposition 2. *Define the system capacity $\overline{C}$ (with multiple areas) by the harmonic mean*

$$\frac{1}{\overline{C}} = \sum_{j=1}^{J} \frac{q_j}{C_{1,j} + C_{2,j}} \tag{8}$$

of total capacities $C_{1,j} + C_{2,j}$ in area A_j, weighted by probabilities q_j, $1 \leq j \leq J$, introduced in (1). Denoting the system load by

$$\rho = \frac{(\alpha + \beta)\sigma}{\overline{C}}, \tag{9}$$

then $\rho < 1$ is a necessary stability condition (for the above defined service disciplines).

Proof. Denote by ρ_j the load in area A_j associated with both Carriers 1 and 2; ρ_j is defined as the traffic intensity offered by users in area A_j divided by the total capacity $C_{1,j} + C_{2,j}$ in A_j, namely,

$$\rho_j = \frac{(\alpha_j + \beta_j)\sigma}{C_{1,j} + C_{2,j}}. \tag{10}$$

According to [13], a necessary stability condition for such a work-conserving multiclass systems is that the total offered load on both Carriers 1 and 2 must be less than 1, that is, $\sum_{j=1}^{J} \rho_j < 1$ which, in view of definitions (8), (9) and (10), is equivalent to inequality $\rho < 1$, as claimed. □

Neither a fluid or an "à la Lyapounov" approach used in Proposition 1 proves, however, easy to extend to the multi-area case to justify the sufficiency of the stability condition $\rho < 1$. The following argument can, nevertheless, be proposed in the favor of this *conjecture*. First define the simple *Bernoulli scheduling policy* as follows: a flow arriving in area A_j is scheduled on Carrier 1 with probability $C_{1,j}/(C_{1,j}+C_{2,j})$ and on Carrier 2 with probability $C_{2,j}/(C_{1,j}+C_{2,j})$ (here both SC and DC traffic is scheduled according to the Bernoulli policy). For the Bernoulli policy, both carriers behave independently as a multiclass PS queue; a necessary and sufficient stability condition for Carrier 1 under the Bernoulli policy is then ([14], Proposition 3.1)

$$\sum_{j=1}^{J} \rho_{1,j} < 1 \tag{11}$$

where $\rho_{1,j}$ is the load on Carrier 1 induced by users from area A_j, that is,

$$\rho_{1,j} = \left[(\alpha_j + \beta_j)\frac{C_{1,j}}{C_{1,j} + C_{2,j}}\right] \times \frac{\sigma}{C_{1,j}} = \rho_j \tag{12}$$

with ρ_j given in (10) (we similarly show that the load on Carrier 2 induced by users from area A_j is $\rho_{2,j} = \rho_j$). In view of (9) and (12), condition (11) equivalently reads $\rho < 1$. Considering the stability region for a given policy as that associated with a positive throughput, the above discussion shows that the throughput γ^B for the Bernoulli policy is strictly positive in the capacity region defined by $\rho < 1$. Given that JFQ and VB scheduling policies both take into account the state of the system, they are expected to perform better than the blind Bernoulli scheduling in terms of throughput, we have $\gamma \geq \gamma^B > 0$ as soon as $\rho < 1$. We could thus infer that the multiclass system implementing JFQ and VB scheduling is also stable in the region defined by $\rho < 1$.

In the following, we assume that $\rho < 1$ is a necessary and sufficient stability condition.

D. Throughput performance

If stability condition $\rho < 1$ is fulfilled, the system has a stationary distribution. Let then

$$\Pi(\mathbf{n}) = \lim_{t \uparrow +\infty} \mathbb{P}(\mathbf{n}(t) = \mathbf{n}), \quad \mathbf{n} \in \mathbb{N}^{3J},$$

define the stationary distribution of the Markov process $(\mathbf{n}(t))_{t \geq 0}$. That stationary distribution can be computed by writing the associated balance equations which are not given here for the sake of brevity. Solving these equations enables us to determine the stationary distribution Π which in turn allows us to derive various performance indicators. In particular, we are interested in deriving the mean flow throughput defined as the ratio of the mean flow size to the mean flow duration. Using Little's law, the mean flow throughput $\gamma_{SC,j}$ and $\gamma_{DC,j}$ for SC and DC flows arrived in area A_j can be expressed by

$$\gamma_{SC,j} = \frac{\alpha_j \sigma}{\mathbb{E}(n_{1,j} + n_{2,j})}, \qquad \gamma_{DC,j} = \frac{\beta_j \sigma}{\mathbb{E}(m_j)}. \tag{13}$$

Assume for instance that there are no SC users in the system. In view of (2), rates are given by $D_{1,j}(m) = C_{1,j}/m$ and $D_{2,j}(m) = C_{2,j}/m$ when the system contains m DC flows, and thus all DC clients in area A_j are served with a total rate $C_{1,j} + C_{2,j}$. The system then corresponds to a multiclass PS queue where each area A_j defines a different service class; the mean number of clients in each area A_j can thus be written as

$$\mathbb{E}(m_j) = \frac{\rho_j}{1 - \rho} \tag{14}$$

where $\rho_j = \beta_j \sigma/(C_{1,j} + C_{2,j})$ is the load due to DC users in area A_j and $\rho = \sum_{j=1}^{J} \rho_j$. From (13) and (14), we then obtain

$$\gamma_{DC,j} = (C_{1,j} + C_{2,j})(1 - \rho). \tag{15}$$

In view of (15), it is important to note that when only DC users are present in the system, the VB scheduling policy achieves ideal load balancing between the two carriers. Indeed, the throughput obtained by the DC flows under VB is equal to the throughput obtained for a single carrier of capacity $C_{1,j} + C_{2,j}$ shared according to the PS policy.

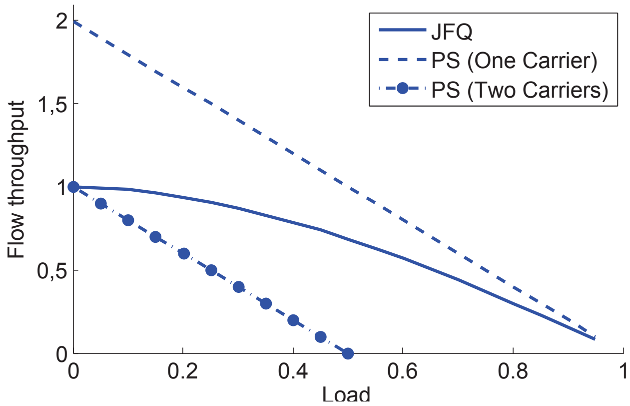

Fig. 2. Mean throughput of SC and DC users in terms of load: JFQ scheduling over 2 carriers of capacity $C = 1$ (SC users), PS scheduling over two carriers of capacity $C = 1$ (DC users) and PS scheduling over one carrier of capacity $C = 1$.

IV. Same capacity carriers

In this section, we analyze the case where both carriers have equal capacities. This corresponds to the Dual Carrier feature of HSDPA in Release 8 where two identical carriers of 5 MHZ, both in the 2.1 GHZ band, are aggregated. For the sake of simplicity, we consider the case of a single area, i.e. $J = 1$, and drop index j in the notation. The general case where $J > 1$ and $C_{1,j} \neq C_{2,j}$, $1 \leq j \leq J$, is considered in Section V. We thus set $C_1 = C_2 = C$. Let ϕ denote the proportion of SC traffic out of the total traffic, so that $\alpha = \phi\Lambda$ and $\beta = (1-\phi)\Lambda$.

The throughput performance of DC users when only DC devices are present in the cell is simply given by (15) for $C_1 = C_2 = C$, that is,

$$\gamma_{DC} = 2C(1-\rho). \tag{16}$$

A. SC users only

SC users are distributed among the two queues according to the JFQ discipline. When the two carriers have equal capacities, i.e., $C_1 = C_2$ for the JFQ policy coincides with to the well-known *Join the Shortest Queue* (JSQ) policy. Indeed, the shortest queue in terms of the number of active users is then also the queue which yields the smallest completion time, which coincides with JSQ when the two carriers have the same capacity. Note by symmetry that the average number of customers in both queues is equal, i.e., $\mathbb{E}(n_1) = \mathbb{E}(n_2) = n(\rho)$. An analytical expression for $n(\rho)$ for two parallel queues of equal capacity ruled by the JSQ policy is derived in [15]. Based on this result and using Little's law, the throughput of SC users under the JSQ policy can be written as

$$\gamma_{SC} = \frac{\alpha\sigma}{n(\rho)} = \frac{C\rho}{n(\rho)} \tag{17}$$

with ρ defined in (6) and $n(\rho)$ is given in [15].

Figure 2 gives the mean flow throughput expressed in (17) and compares it to the throughput obtained when a single carrier is available for SC users and to the throughput of DC users. We note that the throughput provided by JFQ is lower than that of DC users, which is expected since JFQ schedules users on only one carrier at a time. However, as shown in Figure 2, compared to the case when a single carrier is available, JFQ improves the throughput and doubles the capacity region.

Moreover, using results in [15], it can be shown from (17) that as the load increases, $\gamma_{SC} \sim 2C(1-\rho)$, i.e., γ_{SC} is asymptotic to γ_{DC}, the throughput of DC users given in (16). In other words, at high network load, JSQ allows SC users to efficiently utilise both carriers and to attain a throughput which is equivalent to that of DC users.

B. Joint performance of SC and DC users

We now consider the practically interesting case in which both SC and DC users are present in the cell. Figure 3 represents the flow throughput as a function of the network load for different values of ϕ. The results are obtained by numerically solving the balance equations of the Markov process $\mathbf{n}(t)$ presented in Section III-D. Note that truncation of the state space is necessary in order to numerically solve the balance equations. This engenders less accuracy at high network load, say $\rho > 0.9$.

We note that the performance of DC users is slightly improved when 50% of the traffic originates from SC users. This is because each SC user is constrained to using a single carrier which is favourable to DC users; conversely, the performance of SC users is slightly degraded by the presence of DC users. The results of Figure 3, however, indicate that mixing both SC and DC traffic does not significantly impact the throughput performance of neither SC nor DC users.

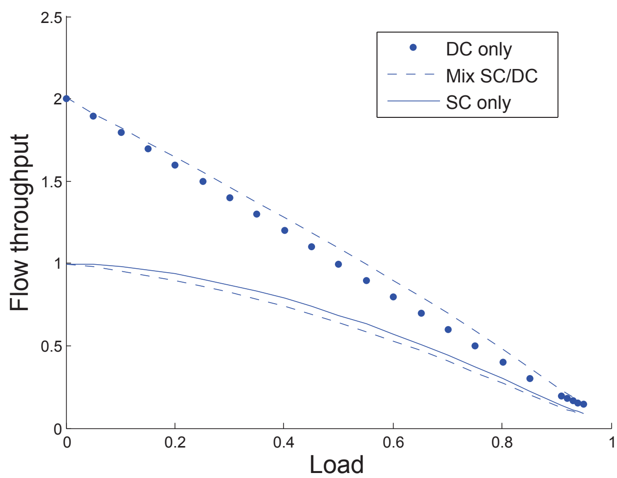

Fig. 3. *Mean throughput of SC and DC users in terms of network load for different values of ϕ: $\phi = 0$ (only DC traffic), $\phi = 0.5$ (mix of SC and DC traffic) and $\phi = 1$ (only SC traffic) and for $C = 1$.*

V. Different capacity carriers

We now consider the more general case of two carriers having different capacities, i.e. $C_1 \neq C_2$. This corresponds to Dual Band HSDPA and in LTE-A with carriers of different sizes. At the end of this section, we also consider the case of having two different areas in the cell. In the first part of the section, we normalize results with respect to C_1, i.e., $C_1 = 1$ and C_2 is either $C_2 = 1.3$ or $C_2 = 2$ in order to model DB HSDPA and LTE-A networks. The choice of these values is motivated as follows. DB HSDPA aggregates two carriers of 5 MHZ each, but the one in the 900 MHZ band has a slightly larger capacity due to favourable propagation conditions. On the other hand, LTE-A allows to aggregate carriers of different capacities; for instance, a carrier of 20 MHZ can be aggregated with a carrier of 10 MHZ within the same frequency band. A numerical application for HSDPA networks is presented in Section V-C.

The performance of DC users is given by (15) for $J = 1$. We subsequently consider the case in which there are only SC users in the system and the case of mixed SC and DC users.

A. SC users only

We compare the proposed JFQ discipline to the well-known JSQ discipline in the case in which only SC users are present in the cell, i.e., $\phi = 1$.

Figure 4 represents the mean throughput of SC flows under the two scheduling disciplines for $C_1 = 1$ and $C_2 = 2$. The JFQ discipline clearly outperforms JSQ. This is because JFQ takes scheduling decisions based not only on the queue size but also on the capacity of each queue. Based on the numerical evaluations, the throughput of JFQ can be approximated by

$$\gamma_{SC} \approx C_2(1 - \rho), \tag{18}$$

which corresponds to the throughput of a PS server of capacity C_2. As $C_2 > C_1$, the JFQ policy therefore yields a performance similar to that obtained when the users are always assigned to the carrier with the largest capacity.

B. Joint performance of SC and DC users

Assume now that both SC and DC users are present in the cell. Throughput performance is given in Figure 5 for different values of ϕ and for $C_1 = 1$, $C_2 = 1.3$. As in the case of equal capacity servers, we notice a slight improvement of DC performance when SC traffic is present in the system (equivalently, when SC traffic replaces DC traffic with identical total load). Once more, the impact of ϕ is not considerable. In other words, since performance is quasi-independent of the mix of SC and DC traffic, we can study the performance of each class independently by considering the system with only SC or only DC traffic.

Figure 6 shows the impact of the percentage of DC traffic on the *average throughput* of the cell for different values of the load. The average troughput in area A_j, $\overline{\gamma}_j$, is defined by

$$\overline{\gamma}_j = \phi \cdot \gamma_{SC,j} + (1 - \phi) \cdot \gamma_{DC,j}. \tag{19}$$

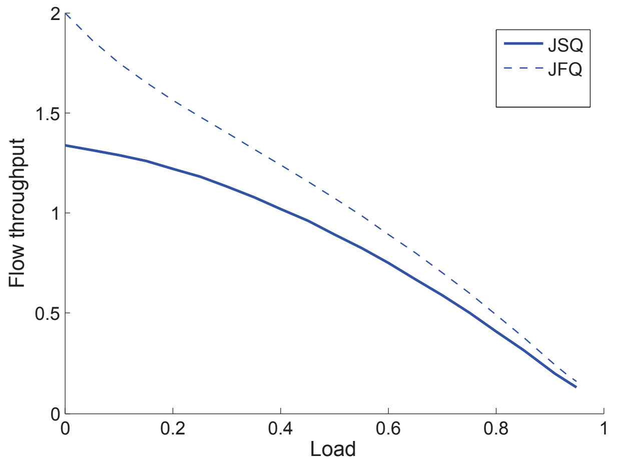

Fig. 4. *Mean throughput of SC users under JSQ and JFQ scheduling for* $C_1 = 1$ *and* $C_2 = 2$.

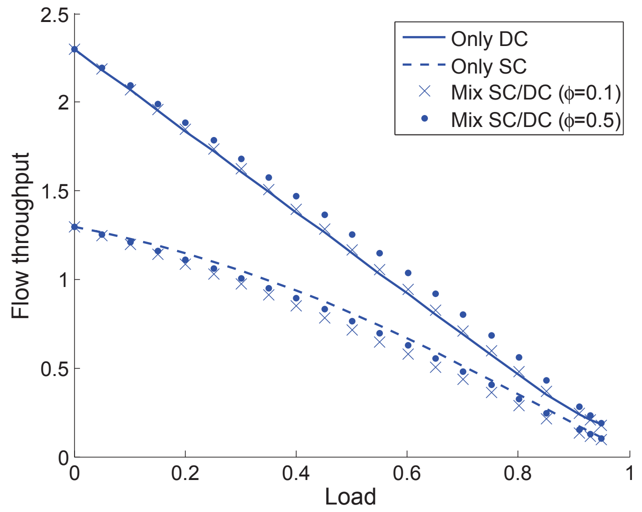

Fig. 5. *Mean throughput of SC and DC users in terms of network load for different values of* ϕ*:* $\phi = 0$ *(only DC traffic),* $\phi = 0.1$*,* $\phi = 0.5$ *(mix of SC and DC traffic) and* $\phi = 1$ *(only SC traffic) and for* $C = 1$.

Note that at any network load the total throughput increases as the percentage of DC traffic increases. The increase is more significant at lower network load. In fact, we have seen that γ_{SC} and γ_{DC} are quasi-insensitive to variations of ϕ; in view of (19), $\overline{\gamma}$ varies quasi-linearly with ϕ with slope $\gamma_{SC} - \gamma_{DC}$; and, as shown in Figure 5, this slope is maximal at low load.

C. Application to HSDPA

We are now interested in applying our model so as to estimate the traffic capacity of Dual Cell HSDPA and Dual Band HSDPA. We consider a cell having two distinct areas, one corresponding to transmission conditions in the cell center and the other corresponding to transmission conditions at the cell edge. Our aim is to determine the traffic that can be

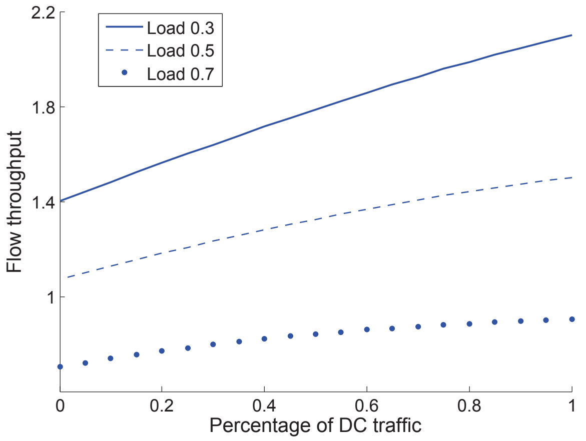

Fig. 6. *Mean throughput of SC and DC users for different values loads and for $C_1 = 1$, $C_2 = 2$.*

TABLE I

Traffic intensity for different percentages of SC traffic (ϕ) when the target average throughput at cell edge is $\overline{\gamma}_J$ =1 Mbit/s

	$\phi = 1$	$\phi = 0.8$	$\phi = 0.5$	$\phi = 0.2$
DB HSDPA (Mbit/s)	1.75	2.11	2.38	2.65
DC HSDPA (Mbit/s)	-	1.08	1.48	1.73

sustained by the cell such that SC users in the cell edge attain a certain target throughput.

We consider a cell with a total radius of $R = 600$ m and consider that users are uniformly distributed over the cell. In [17], the capacities of DC and DB HSDPA are obtained by means of a static system level simulator. According to [17], capacities of DC HSDPA correspond to $C_{1,1} = C_{2,1} = 10$ Mbit/s, while DB HSDPA has $C_{1,1} = 10$ Mbit/s and $C_{2,1} = 14$ Mbit/s, for carriers in the 2.1 GHZ and the 900 MHZ band, respectively.

Table I shows the maximum traffic intensity that can be sustained by a HSDPA cell such that the target average throughput at the cell edge is 1 Mbit/s and 10 Mbit/s, respectively. As for the total average throughput (cf. §V-B), the traffic intensity increases as the percentage of DC traffic increases, for both DB and DC HSDPA. As expected, the traffic capacity of DB HSDPA is superior to that of DC HSDPA.

VI. Conclusion

This paper addresses the performance of load balancing and carrier aggregation in HSPA+ networks.

In such networks, a joint scheduler is used to manage two carriers. Load balancing schemes are needed in order to evenly distribute the incoming traffic. We have proposed two such schemes: **1)** JFQ ("Join the Fastest Queue"), which allows to distribute the traffic of non-CA users, **2)** and VB ("Volume Balancing"), which balances the traffic of CA users. By means of mathematical modeling, we have shown that both schemes achieve efficient load balancing over the carriers. Indeed, when only DC users are present in the system, the throughput obtained by DC flows under VB is equal to the throughput obtained for a single carrier of capacity $C_{1,j} + C_{2,j}$ shared according to the PS policy; VB thus maximizes the utilization of the two carriers. SC users, on the other hand, are constrained to using only one of the two carriers. We have seen that JFQ favors the usage of the highest capacity carrier, thus maximizing carrier utilization.

The proposed model also allows us to gain insight into the performance of multi-carrier networks in which traffic is generated by both CA and non-CA users. We have shown that the throughput of both CA and non-CA users is practically insensitive to the percentage of CA users. Consequently, we can evaluate the performance of each class independently by considering the system with only SC or only DC traffic. In future work, we intend to extend the developed models to a larger number of aggregated carriers.

VII. Acknowledgment

The authors thank Bruno Kauffmann for his useful inputs and his contribution on the proof of the stability condition.

References

[1] H. Holma and A. Toskala, *WCDMA for UMTS: HSPA Evolution and LTE*, John Wiley and Sons, 2010.
[2] D. Morais, A. Klein, H. Holma, I Viering, G. Liebl, *Performance evaluation on dual-cell HSDPA operation*, IEEE VTC Fall 2009.
[3] D. Zhang, P. Kumar, B. Mohanty, J. Hou, *Performance analyis of dual-carrier HSDPA*, Proc. IEEE VTC Spring 2010.
[4] E. Seidel, J. Afzal, G. Liebl, *Dual cell HSDPA and its future evolution*, White paper, Nomor Research GmbH, January 2009.
[5] T. Bonald, S.E. Elayoubi, A. El Falou, J.B. Landre, *Radio capacity improvement with HSPA$^+$ dual-cell*, Proceedings of IEEE ICC 2011.
[6] I. Repo, K. Aho, S. Hakola, T. Chapman, and F. Laakso, *Enhancing HSUPA system level performance with dual carrier capability*, in Proc. Wireless Pervasive Computing (ISWPC), 2010 5th IEEE International Symposium on, 2010, pp. 579-583.
[7] A. Razzac,S.E. Elayoubi, A. El Falou, B. El Hassan, *Performance Evaluation of Dual Carrier Feature in the Uplink of HSPA+ Systems*. In Vehicular Technology Conference (VTC Fall), 2012 IEEE (pp. 1-5).
[8] R. Ratasuk, D. Tolli, A. Ghosh, *Carrier Aggregation in LTE-Advanced*, Vehicular Technology Conference - VTC Spring , pp. 1-5." 2010.
[9] G. Yuan, X. Zhang, W. Wang, Y. Yang., *Carrier Aggregation for LTE-Advanced*, Mobile Communication Systems, IEEE Communications Magazine, February 2010.
[10] K.I. Pedersen, F. Frederiksen, C. Rosa, H. Nguyen, L.G.U. Garcia, Y. Wang, *Carrier aggregation for LTE-advanced: functionality and performance aspects*, IEEE Communications Magazine, Volume:49 , Issue: 6.
[11] L. Zhang, et al., *QoS performance analysis on carrier aggregation based LTE-A systems*. Wireless Mobile and Computing (CCWMC 2009), IET International Communication Conference on. IET, 2009.
[12] H. Wang, C. Rosa, and K. Pedersen, *Performance analysis of downlink inter-band carrier aggregation in LTE-advanced*. Vehicular Technology Conference (VTC Fall), 2011 IEEE. IEEE, 2011.
[13] T. Bonald and A. Proutière, *Wireless downlink data channels: user performance and cell dimensioning*, Proc. of MobiCom, 2003.
[14] S. Borst, *User-level performance of channel-aware scheduling algorithms in wireless data networks*, Proc. of IEEE InfoCom, 2003.
[15] L. Flatto, H.P. McKean, *Two queues in parallel*, Communications in Pure and Applied Mathematics, vol. XXX, pp.255-263, 1977
[16] P. Robert, *Stochastic Networks and Queues*, ed. Springer 2003
[17] N. Khaji, S. E. Elayoubi, F. Marache, *HSDPA capacity gain in the 900 MHz band*, Vehicular Technology Conference, 2009. 69th IEEE VTC Spring 2009 (pp. 1-5).