Dynamic downlink aggregation carrier scheduling scheme for wireless networks

Kan Zheng¹, Fei Liu², Wei Xiang³, Xuemei Xin¹

¹Wireless Signal Processing and Network Lab, Key Laboratory of Universal Wireless Communication, Ministry of Education, Beijing University of Posts & Telecommunications, P.O. Box 93, No. 10, Xi Tu Cheng Road, Beijing 100088, People’s Republic of China
²Institute for Networked Systems, RWTH Aachen University, Kackertstrasse 9, D-52072 Aachen, Germany
³School of Mechanical and Electrical Engineering, University of Southern Queensland, Toowoomba, QLD 4350, Australia

E-mail: zkan@bupt.edu.cn; kzheng@ieee.org

Abstract: Carrier aggregation has been accepted as a means of bandwidth extension in the third generation long-term evolution-advanced (LTE-advanced) network, in an effort to support high data rate transmission with backwards compatibility. Since there are two or more component carriers (CCs) to be aggregated, it is crucial to design efficient carrier scheduling schemes. In this study, the authors propose a novel dynamic aggregation carrier (DAC) scheme for downlink transmission, which enables CCs to aggregate with each other in a dynamic manner. The dynamic nature of the new scheme allows the total capacity of all CCs to be fully utilised to serve flows, whereas the number of aggregated supplementary CCs is decreased so as to lower the computational complexity at user equipment (UE). Furthermore, the performances of the new scheme and two other carrier scheduling schemes are evaluated thoroughly through both analytical and simulation results. It is demonstrated that the DAC scheme offers good performances in terms of delay and throughput while reducing energy consumption and the signalling overhead at UEs.

1 Introduction

Wide bandwidth, that is, up to 100 MHz, is required for the third generation (3G) long-term evolution-advanced (LTE-advanced) network, whose target is to support a peak data rate of 100 Mbps under high mobility and 1 Gbps under low mobility [1]. Generally speaking, the most straightforward way to meet such requirements is to use large contiguous bandwidth. However, such a straightforward method proves to be quite challenging because of highly scarce radio frequency (RF) spectra available for mobile communications. On the other hand, spectral segments belonging to an operator, which may be of different sizes and not necessarily located contiguously in the same frequency band, should still be utilised. As a result, carrier aggregation (CA) is adopted by the 3G partnership project (3GPP) as a solution to bandwidth extension for LTE-advanced [2]. Through CA techniques, two or more component carriers (CCs) can be aggregated to support high data rate transmission with backwards compatibility.

At present, there exist three types of spectrum configurations for CA, that is, contiguous CC, non-contiguous CC in a single frequency band and non-contiguous CC in multiple frequency bands. When CA was first introduced in LTE-advanced, the radio access network has to be modified or complemented in an effort to achieve high data rates while reducing upgrade costs and implementation complexity. There are a number of technical challenges relating to CA-based LTE-advanced systems such as uplink CC selection, radio resource connection procedures, radio resource management of multiple CCs, among which the carrier scheduling technique is the most significant in affecting the performance of CA-based LTE-advanced systems. The cross-CC scheduling functionality in CA for improved control channel optimisation is discussed in [3]. The performance of CA with elastic traffic is analysed and compared with those of the independent carrier (IC) approach [4]. Then, a burst-level carrier scheduling scheme is proposed in [5], which achieves a good trade-off between performance and complexity. In [6], two novel methods of search space design for mapping control information on one CC are proposed, which can reduce the search space overlapping of different user equipment (UE). Moreover, CA technique also helps reducing interference among femtocells in LTE-advanced networks [7].

In CA, the base station, termed the eNodeB in an LTE-advanced network, manages all available CCs by using a carrier scheduling scheme. An UE may simultaneously receive or transmit on one or multiple CCs depending on its capabilities. The simplest approach of managing multiple CCs is to allocate each user onto a single CC. The UE can receive data only on one of the CCs at a time. This approach does not require any change to 3GPP Rel-8 specifications, and above all, the UE costs and complexity are greatly reduced by not having to support bandwidth wider than 20 MHz. This is referred to as ‘IC’ in the rest of this paper. However, the IC
scheme performs poorly in terms of spectrum utilisation. By contrast, another straightforward way is to jointly schedule the radio resources of all the CCs, which is termed ‘joint carrier’ (JC). It can attain the optimal performance, but is less practical because of overwhelming signal processing complexity at the UE. It has been shown that the mean throughput of the IC approach with a random routing dispatcher is only 1/L of that of its JC counterpart, when there are L CCs with the same bandwidth in the system, irrespective of the bandwidth of these CCs [4]. Furthermore, the performance of the IC scheme can be improved by using a smarter dispatcher [8]. As a result, advanced carrier scheduling schemes are expected to strike a good balance between practicability and efficiency. However, to the best of our knowledge, this problem has not yet been well researched to date.

In this paper, the performances of a variety of carrier scheduling schemes for CA-based LTE-advanced systems under elastic traffic are analysed. Owing to the limited capabilities of UEs and the switch signalling overhead, it is better to maintain the connection of an UE with only one CC as long as possible. In other words, infrequent switching between CCs is desirable. Motivated by the coupled processors in [9], we propose a novel ‘dynamic aggregation carrier (DAC)’ scheme which enables CCs to help one another in a dynamical manner. When no CC has an empty queue, each CC serves only its own queue. Whenever a queue of any CC becomes empty, this carrier will be able to help its counterparts. By this way, the total capacity of the carriers can be utilised to serve flows of the CCs with a non-empty queue. When the number of the CCs is greater than two, two methods, that is, ‘serving the longest queue (SLQ)’ and ‘round robin with priority (RRP)’, can be used for cooperation among the CCs. Simulation and numerical results are presented for evaluating the performances of the various schemes.

The remainder of this paper is organised as follows. The system model is introduced in Section 2. Several carrier scheduling schemes are described in Section 3. In Section 4, the performances of different carrier scheduling schemes are discussed. Section 5 compares all of the schemes in terms of simulation and analytical results. Section 6 concludes this paper.

Owing to a large number of abbreviations used in this paper, a list of acronyms is given in Table 1.

<table>
<thead>
<tr>
<th>Abbreviations</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>3G</td>
<td>third generation</td>
</tr>
<tr>
<td>CA</td>
<td>carrier aggregation</td>
</tr>
<tr>
<td>CC</td>
<td>component carrier</td>
</tr>
<tr>
<td>DAC</td>
<td>dynamic aggregation carrier</td>
</tr>
<tr>
<td>IBS</td>
<td>ideally balanced system</td>
</tr>
<tr>
<td>IC</td>
<td>independent carrier</td>
</tr>
<tr>
<td>JC</td>
<td>joint carrier</td>
</tr>
<tr>
<td>JSO</td>
<td>joint shortest queue</td>
</tr>
<tr>
<td>LTE</td>
<td>long-term evolution</td>
</tr>
<tr>
<td>PCC</td>
<td>primary component carrier</td>
</tr>
<tr>
<td>PS</td>
<td>processor-sharing</td>
</tr>
<tr>
<td>RA</td>
<td>random allocation</td>
</tr>
<tr>
<td>RR</td>
<td>round robin</td>
</tr>
<tr>
<td>RRP</td>
<td>round robin with priority</td>
</tr>
<tr>
<td>RS</td>
<td>resource scheduler</td>
</tr>
<tr>
<td>SCC</td>
<td>secondary component carrier</td>
</tr>
<tr>
<td>SLQ</td>
<td>serving the longest queue</td>
</tr>
</tbody>
</table>

### 2 System model

Let us consider the downlink of an OFDMA-based network, which consists of a single eNB communicating with multiple UEs in a time-slotted fashion. Although CA can support the aggregation of both non-contiguous and heterogenous CCs, we assume for simplicity that the eNB employs L adjacent CCs in the same frequency band, and that the bandwidth of each CC is the same. A dynamic flow model with elastic traffic is assumed, where a new flow arrives at the system with a finite-length file request, and leaves the system after the file is transmitted [10]. Without loss of generality, each UE is assumed to start a new transmission only after the old one is finished, and each new transmission by the same UE is treated as a new flow. The arriving flow to the network follows a Poisson process with an average arrival rate of λ. The sizes of the flows are exponentially distributed as shown below

\[ P\{X^{(n)} \leq a\} = 1 - e^{-a/F}, \quad a \in \{0, 1, 2, \ldots, \infty\} \quad (1) \]

where \(F^{(n)}\) is the file length of user n, and \(F = \mathbb{E}[F^{(n)}]\) is the average size of the flows.

Each CC is comprised of K resource blocks (RBs), also known as subchannels, which span multiple subcarriers in the frequency domain per time slot. One CC can serve multiple UEs through its resource scheduler (RS) at the eNB. Each UE is able to receive data from multiple CCs simultaneously, albeit with an increased computational complexity and power consumption at the UE. In each time slot, the RS allocates RBs to serve given UEs in accordance with a certain scheduling strategy. In this paper, an equal number of RBs are allocated to UEs in the queue circularly using the round robin (RR) strategy. The queuing model of the RR scheduler can be modelled as a processor-sharing (PS) queue [10], where the UEs share the wireless resources of one or multiple aggregated CCs allocated to them.

For simplicity, the transmit power is assumed to the same for all RBs, that is, \(P_{R,k} = P_{R}\), \(1 \leq l \leq L\), \(1 \leq k \leq K\). The received signal on RB k at CC l of UE n is given by

\[ y_{l,k}^{(n)} = \sqrt{P_{R}} \alpha_n h_{l,k}^{(n)} x_n + z_{l,k}^{(n)}, \quad 1 \leq l \leq L, \quad 1 \leq k \leq K \quad (2) \]

where \(x_n\) represents the transmitted signal from the nth UE, \(h_{l,k}^{(n)}\) denotes the independent complex fading channel gain on RB k at CC l of UE n, modelled as \(h_{l,k}^{(n)} \sim CN(0, \sigma^2)\) with \(\sigma^2 = \mathbb{E}\left|h_{l,k}^{(n)}|^2\right|\), \(\alpha_n\) represents the path loss attenuation factor of UE n and \(z_{l,k}^{(n)}\) is modelled as zero-mean additive white Gaussian noise with a noise power of \(\sigma^2\). Then, the instantaneous signal-to-noise ratio (SNR) on RB k at CC l of UE n can be calculated as

\[ y_{l,k}^{(n)} = \frac{\alpha_n |h_{l,k}^{(n)}|^2 P_{R}}{\sigma^2}, \quad 1 \leq l \leq L, \quad 1 \leq k \leq K \quad (3) \]

Correspondingly, the achievable data rate on RB k at CC l of UE n is given by

\[ C_{l,k}^{(n)} = W \log_2 \left(1 + \frac{\alpha_n |h_{l,k}^{(n)}|^2 P_{R}}{\sigma^2}\right), \quad 1 \leq l \leq L, \quad 1 \leq k \leq K \quad (4) \]

where W is the bandwidth of each subchannel.
3.1 Joint carrier

In JC, each UE has access to one or more CCs, and is served across all CCs. In other words, the data of any current flows may be transmitted over all the CCs. As illustrated in Fig. 1, there is only a single RS at the eNB, which is responsible for jointly assigning the RBs of all the CCs to the flows.

Such a joint processing mechanism can lead to maximum spectral efficiency and full resource utilisation in the CA-based LTE-advanced system. Therefore, the optimal system performance can be achieved by using the JC scheme. However, with JC, each UE has to maintain simultaneous connections with all the CCs. A user needs to use multiple fast Fourier transform (FFT) and RF units to receive data from more than one CC. Thus, the effectiveness of JC is largely constrained by the computational complexity and power consumption that can be afforded by the UE.

3.2 Independent carrier

In IC, each UE can only use a single CC to exchange data with the eNB. When a new flow arrives at the system, a flow dispatcher immediately assigns it to a CC, as shown in Fig. 1. Then, the data of this flow can only be transmitted over the assigned CC. Specifically, there are one independent RS and flow queue for each CC. The RS of the assigned CC allocates RBs to its incoming flows.

In an effort to balance the traffic load across the CCs, the dispatcher needs to sensibly select an allocation strategy. For example, the RR, random allocation (RA) or joint shortest queue (JSQ) method. RR is the simplest allocation method, where flows are assigned to the queues of all the CCs circularly. With the RA approach, the dispatcher assigns arriving flows to different CCs with given probabilities. As for the JSQ method, it takes into account the length of each queue, and allocates an arriving flow to the queue with the least number of flows.

3.3 Dynamic aggregation carrier

When the IC scheme is applied, a CC remains idle so long as all its assigned flows have been transmitted, although there may be flows assigned to other CCs that are still being served. On the other hand, the JC approach has a better
trunking efficiency, that is, less number of backlogged user packets, than its IC counterpart at the expense of computational complexity and energy consumption at the UE. We propose a ‘DAC’ scheme which is capable of striking an elegant balance between performance and complexity especially for elastic traffic.

In accordance with certain dispatching principles such as RR and JSQ, arriving flows are allocated by a dispatcher to the queue at each CC. An UE in the queue of a CC regards this CC as its primary CC (PCC) to transmit both signalling information and data. An UE has only a single PCC at a time, which is not supposed to be changed during transmission since switching the PCC is very time-consuming for 3GPP LTE-advanced systems [12].

With the proposed DAC method, each CC first serves the flows in its own queue when it is non-empty. Whenever a queue of any CC becomes empty, the CC with an empty queue, dubbed the supplementary CC (SCC), will assist other CCs. In this way, the total capacity of all the CCs can be fully utilised to serve flows of the CCs with non-empty queues. Furthermore, we propose two methods for cooperation between the CCs under the DAC scheme in the following.

3.3.1 Serving the longest queue: All the CCs with an empty queue, that is, the SCCs, are aggregated with the PCC having the longest queue. If more than one PCC has the same longest queue length, the SCCs are allocated to any of them randomly. For illustrative purposes, an example of the system with four CCs using the SLQ scheme is illustrated in Fig. 2a. Assume that the queues of CCs 2 and 3 are empty, that is, the queue length \( Q_2 = Q_3 = 0 \), whereas those of CCs 1 and 4 are non-empty. Then, CCs 2 and 3 may be used as the SCCs to help others. CC 1 will be aggregated with CCs 2 and 3 if the queue of CC 1 is longer than that of CC 4, that is, \( Q_1 > Q_4 \). Then, the RBs of CCs 2 and 3 are aggregated with those of CC 1, altogether serving the flows in the queue of CC 1.

3.3.2 Round robin with priority: The PCCs are first sorted according to their queue length before being coupled with the SCCs. The SCCs are circularly allocated to the ordered PCCs based on the RR principle. With this method, the PCC with a longer queue length has a higher priority to be coupled with one or more SCCs. Meanwhile, the PCC with a shorter queue length also has the opportunity to be aggregated with SCCs when there are more than one SCC. As shown in Fig. 2b, unlike the SLQ scheme, CC 2 is firstly allocated to CC 1 because CC 1 has a longer queue length. Next, CC 3 starts helping CC 4.

4 Performance analysis of carrier scheduling schemes

4.1 Queueing performance

Since each RS employs the RR algorithm to schedule RBs, the CCs can be modelled as an \( L \) PS queueing system for both the IC and JC schemes [10]. Each CC in the IC scheme or all the CCs working as one in the JC scheme, are regarded as a processor, the capabilities of which are shared by distributing RBs to the flows in their serving queue by the RS. In consideration of arriving and departing processes, each PS queueing system is an \( M/G/1 \) system [13]. The solution of the average number of backlogged users and response time is given in [14, 15], and proved independent of the distribution of the service time. Moreover, the steady-state results of the average queue length, which are the backlogged flows in an \( M/G/1 \) – PS system, and the average time consumption of each flow in the system are consistent with those of \( M/M/1 \) – PS [16]. Based on the above discussions, we present the average performance of the system under a variety of carrier scheduling schemes in the sequel.

1. JC: with JC, the system can be formulated as a PS server with an arrival rate of \( \lambda \). The average service rate of this PS server is the sum of all the CCs, that is, \( \mu = \sum_{l=1}^{L} \mu_l = L\mu_c \). Define the traffic load of the network as \( \rho = \lambda \mu_c \).

According to existing results on the PS model with a state-dependent service rate [4, 10], the average queue length, average sojourn time (AST) and average throughput
(ATP) can be calculated by

\[
\overline{Q}_{IC} = \frac{\rho}{1 - \rho},
\]

\[
\overline{D}_{IC} = \frac{\overline{Q}_{IC}}{\lambda} = \frac{1}{\mu - \lambda},
\]

\[
\overline{\xi}_{IC} = \frac{\mathbb{E}[F^{(t)}]}{\overline{D}_{IC}} = F(\mu - \lambda)
\]

respectively. The AST consists of two parts, namely, the waiting time and service time in a general queueing system. However, in a PS queueing system, there is no waiting time for each user. Every packet is transmitted immediately after its arrival.

2. IC: the system can be formulated as \( L \) parallel PS servers with a state-dependent service rate, corresponding to the \( L \) CCs. The serving rate of each PS server is \( \mu_j \). There is a dispatcher routing flows to servers such that the arrival rate of flows on server \( l \) is \( \lambda_l \). The dispatch algorithm will impact on the system performance as will be analysed in the following part.

- RA

When the dispatcher employs the RA algorithm, flows probabilistically arrive at server \( l \) independently with probability \( \beta_l \), where \( \sum_{l=1}^{L} \beta_l = 1 \). Then, the arriving process at each server is still Poisson and its rate is \( \lambda_l = \beta_l \lambda \). Define the traffic load of server \( l \) as \( \rho_l = \lambda_l / \mu_l \). Specially, when \( \beta_l = 1/L \), the average queue length, AST and ATP of the whole system can be derived as follows

\[
\overline{Q}_{IC-RA} = \sum_{l=1}^{L} \overline{Q}_l = \sum_{l=1}^{L} \frac{\rho_l}{1 - \rho_l} = \frac{L \rho}{1 - \rho}
\]

\[
\overline{D}_{IC-RA} = \frac{\overline{Q}_{IC-RA}}{\lambda} = \frac{L}{\mu - \lambda}
\]

\[
\overline{\xi}_{IC-RA} = \frac{\mathbb{E}[F^{(t)}]}{\overline{D}_{IC-RA}} = \frac{F(\mu - \lambda)}{L}
\]

- RR

Another simple dispatch algorithm is to allocate packets to each server using the RR approach. The interval of flow arrivals at each server is Erlang distributed [13]. Thus, each CC behaves like an \( E_1/G/1 – PS \) system with an average arrival rate of \( \lambda/L \) and average serving rate of \( \mu/L \).

The analytical performance of this queueing system is attainable only when the serving process satisfies certain conditions. We use the solution of the \( E_1/M/1 – PS \) system with the same arrival process and average serving rate as an approximation. As indicated in [13], its performance is better than that of the \( E_1/G/1 – PS \) system, when the serving time of process \( G \) has a larger variance than that of Poisson process \( M \).

The solution to the \( E_1/M/1 – PS \) queueing system can be obtained according to Ramaswami [17]. Thus, the average queue length, AST and ATP can be derived as follows

\[
\overline{Q}_{IC-RR} = L \frac{\rho}{1 - \eta}
\]

\[
\overline{D}_{IC-RR} = \frac{\overline{Q}_{IC-RR}}{\lambda} = \frac{L}{\mu} \frac{1}{1 - \eta}
\]

\[
\overline{\xi}_{IC-RR} = \frac{\mathbb{E}[F^{(t)}]}{\overline{D}_{IC-RR}} = \frac{\mu}{L} \frac{1 - \eta}{\mu - \lambda}
\]

where \( \eta \) denotes the unique root in the unit circle of (14) below. The pdf of the time interval between flows is given by

\[
p(t) = \frac{\lambda(\lambda)^{L-1}}{(L-1)!} e^{-\lambda t}, \quad t \geq 0
\]

The Laplace transformation of (12) is

\[
f(s) = \int_0^{+\infty} p(t) e^{-st} dt = \left( \frac{\lambda}{\lambda + s} \right)^L
\]

According to the method in [17] in solving the AST in the \( G/M/1 – PS \) system, (5) in [17] can be written as

\[
\eta = f(\mu_j - \mu_l \eta) = \left( \frac{\lambda}{\lambda + \mu_j - \mu_l \eta} \right)^L
\]

\[\Rightarrow \lambda^L = \eta(\lambda + \mu_j - \mu_l \eta)^L\]

The minimal non-negative root of (14) is the solution to \( \eta \) in (11).

- JSQ

The JSQ method allocates a new arriving flow to the serving queue with the least number of flows backlogged [18]. The arrival process of each processor is not independent, since the relationship of the queue lengths needs to be taken into account.

By using a special single queue approximation (SQA) method, the performance of the system based on the PS serving processors with the JSQ method can be analysed according to Gupta et al. [19]. The performance of the \( M/G/1/JSQ – PS \) system may be approximated by that of the \( M_\infty/M/1 – PS \) system. The \( M_\infty/M/1 – PS \) system obtained by SQA is insensitive to the serving process, similar to \( M/G/1 – PS \). Thus, the average queue length and sojourn time of \( M_\infty/M/1 – PS \) are the same as those of \( M_\infty/M/1 – PS \) as follows

\[
\overline{C}_{l,k} = \mathbb{W} \int_0^{\infty} \log_2 \left( 1 + \gamma_l^{(k)} \right) f(\gamma_l^{(k)}) d\gamma_l^{(k)}
\]

\[1 \leq l \leq L, \quad 1 \leq k \leq K\]

The average number of backlogged users for various \( \rho \) and \( L \) can be found in Table 2 [20].

DAC: the system employing the DAC scheme can be modelled as a special \( M/G/K – PS \) system. Different from the IC scheme, the number of service lines \( K \) and the capabilities of them vary depending on the CA state. According to the conservation law for time-shared systems in [13], the average queueing performance of the DAC scheme is the same as that of the IC scheme, because no
transmission bandwidth of the carriers is wasted when there are still flows in the queues waiting to be served. Thus, the DAC system ensures full utilisation of carrier resources, and the average queue length, AST and ATP can be derived as per (9). Moreover, the departing process of the DAC system is also a Poisson one because of the insensitivity of the PS scheduling system, and that the aggregation can be regarded as a combination of Poisson processes [14–16]. The proof of the insensitivity of the $M/G/1−\text{PS}$ queue on the distribution of the service time is detailed in the Appendix.

### 4.2 Number of SCCs

It is well known that, to achieve the same delay and throughput performance, the less number of SCCs used by each UE is, the more suitable the load balance scheme is for practical systems. Firstly, an UE served by only the PCC has lower power consumption in comparison with one served by both PCC and SCCs. The power consumption increases with the number of CCs an UE has to receive and process because of the more FFT and RF units required and more data streams combined by the multiplexer [3]. Secondly, a less control signalling overhead is needed when a less number of SCCs are aggregated with the PCC. For example, in the JC scheme, the channel state information of all the CCs need to be estimated and transmitted for each user. This will consume more bits than using fewer SCCs or only the PCC. Therefore the number of active SCCs plays a crucial role in evaluating the performance of the proposed carrier load balance schemes.

For ease of analysis delay because of CA scheduling is not considered in this paper. The CCs with an empty queue can be immediately allocated to users with flows by the proposed scheduling schemes. In addition, all the CCs can be used as the PCCs of different users if the number of the CCs is no less than that of existing users in the system. Thus, a minimum number of active SCCs are needed in such a system. We call it the ideally balanced system (IBS). However, when the number of active users is less than the total number of CCs, each UE can use only one CC as its PCC and thus does not need to aggregate SCCs in the IBS. However, when the number of active users is less than that of CCs, to guarantee the full utilisation of wireless resources, some UEs can use SCCs aggregated with their PCCs in accordance with a certain DAC scheme. Using a variety of DAC schemes may result in different SCC aggregation states. A four-CC system illustrated in Fig. 3 is used as an example to demonstrate this fact.

The steady-state probability of each state of this $M/M/1$ system can be easily solved as follows:

\[
\begin{align*}
\pi(0) &= 1 - \rho \\
\pi(n) &= (1 - \rho)\rho^n, \quad n = 1, 2, \ldots
\end{align*}
\]  

- SLQ method

There is no CC working when the system is in the idle state. When there is a single UE in the system, that is, in state 1, the

1. JC: for the JC scheme, the number of SCCs is always $L − 1$, indicating $P[N_{\text{SCC}} = L − 1] = 1$, since there is only a single PCC aggregated by all the other CCs. Hence, JC can only be used in the LTE-advanced system, in which UEs are able to transmit data through multiple CCs, and thus are not backward compatible with LTE UEs.

2. IC: for the IC scheme, each CC serves UEs independently without aggregating any other SCC, making the number of SCCs to be 0 at all times, that is, $P[N_{\text{SCC}} = 0] = 1$. However, radio resources may be wasted without an appropriate CA scheme.

3. DAC: as shown in Fig. 3, the arriving and departing flow processes in the DAC system can be modelled as a birth–death process. The birth–death process is a special case of the continuous-time Markov process, where each state represents the current number of active UEs and transitions occurring between neighbouring states. This is a typical $M/M/1$ process, where the arrival of users follows a Poisson distribution while the service time is exponentially distributed. When the number of active users is no less than the total number of CCs, each UE can use only one CC as its PCC and thus does not need to aggregate SCCs in the IBS. However, when the number of active users is less than that of CCs, to guarantee the full utilisation of wireless resources, some UEs can use SCCs aggregated with their PCCs in accordance with a certain DAC scheme. Using a variety of DAC schemes may result in different SCC aggregation states. A four-CC system illustrated in Fig. 3 is used as an example to demonstrate this fact.

The steady-state probability of each state of this $M/M/1$ system can be easily solved as follows:

\[
\begin{align*}
\pi(0) &= 1 - \rho \\
\pi(n) &= (1 - \rho)\rho^n, \quad n = 1, 2, \ldots
\end{align*}
\]  

- SLQ method

There is no CC working when the system is in the idle state. When there is a single UE in the system, that is, in state 1, the

### Table 2 Average number of backlogged users for various $\rho$ and $L$ in the JSQ system [20]

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$L$</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.1018</td>
<td>0.2132</td>
<td>0.3434</td>
<td>0.5044</td>
<td>0.7156</td>
<td>1.0141</td>
<td>1.4831</td>
<td>2.3754</td>
<td>4.9521</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.1003</td>
<td>0.2045</td>
<td>0.3196</td>
<td>0.4545</td>
<td>0.6245</td>
<td>0.8545</td>
<td>1.2015</td>
<td>1.8361</td>
<td>3.5955</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.1001</td>
<td>0.2017</td>
<td>0.3097</td>
<td>0.4321</td>
<td>0.5808</td>
<td>0.7762</td>
<td>1.0626</td>
<td>1.5892</td>
<td>2.9223</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.1000</td>
<td>0.2007</td>
<td>0.3052</td>
<td>0.4202</td>
<td>0.5559</td>
<td>0.7299</td>
<td>0.9796</td>
<td>1.407</td>
<td>2.5252</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.1000</td>
<td>0.2003</td>
<td>0.303</td>
<td>0.4133</td>
<td>0.5404</td>
<td>0.6998</td>
<td>0.9242</td>
<td>1.3011</td>
<td>2.2578</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.1000</td>
<td>0.2001</td>
<td>0.3018</td>
<td>0.409</td>
<td>0.5298</td>
<td>0.6789</td>
<td>0.8843</td>
<td>1.2251</td>
<td>2.0704</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.1000</td>
<td>0.2001</td>
<td>0.3011</td>
<td>0.4063</td>
<td>0.5226</td>
<td>0.6633</td>
<td>0.8549</td>
<td>1.1678</td>
<td>1.9249</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.1000</td>
<td>0.2000</td>
<td>0.3007</td>
<td>0.4044</td>
<td>0.5173</td>
<td>0.6516</td>
<td>0.8320</td>
<td>1.1226</td>
<td>1.8124</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.1000</td>
<td>0.2000</td>
<td>0.3005</td>
<td>0.4031</td>
<td>0.5132</td>
<td>0.6423</td>
<td>0.8135</td>
<td>1.0866</td>
<td>1.7235</td>
<td></td>
</tr>
</tbody>
</table>

---

* Fig. 3 States of the number of SCCs in four CCs IBS with SLQ and RRP

---


© The Institution of Engineering and Technology 2013
UE will have three SCCs aggregated with its PCC. When there are two UEs, one of them may have two SCCs, whereas the other one has no SCC with the SLQ scheme. When there are three UEs, only one UE may have one SCC, whereas the other two have none. When the UE number is more than 3, all UEs have no SCC. The probabilities of various numbers of SCCs are calculated as follows

\[
\begin{align*}
\omega_{\text{SLQ}}(0) &= (1 - \rho)^2 + 2 \rho^3 + 4 \rho^4/(1 - \rho) \\
\omega_{\text{SLQ}}(1) &= (1 - \rho) \rho^3 \\
\omega_{\text{SLQ}}(2) &= (1 - \rho) \rho^2 \\
\omega_{\text{SLQ}}(3) &= (1 - \rho) \rho 
\end{align*}
\]

(17)

The normalised probability of \(l\) SCCs aggregated by the SLQ method is given by

\[
P[N_{\text{SCC}} = l] = \omega_{\text{SLQ}}(l)/\sum_{m=0}^{L-1} \omega_{\text{SLQ}}(m) \quad (18)
\]

where \(P[N_{\text{SCC}} = l], \forall \geq 0\) is the probability of each UE using \(l\) SCCs \((N_{\text{SCC}} = l)\) for transmission, which is a meaningful measurement metric to assess the load balance performance of the DAC schemes and dispatching algorithms. For a general system with \(L\) CCs, \(\omega_{\text{SLQ}}(m)\) can be expressed as

\[
\omega_{\text{SLQ}}(l) = \begin{cases} 
(1 - \rho) \sum_{i=2}^{L-1} (i-1) \rho^i + L \rho^l, & l = 0 \\
(1 - \rho) \rho^{l-1}, & l = 1, \ldots, L - 1
\end{cases}
\]

(19)

• RRP method

The probabilities of various numbers of SCCs using the RRP method with four CCs are given as follows

\[
\begin{align*}
\omega_{\text{RRP}}(0) &= (1 - \rho)[2 \rho^3 + 4 \rho^4/(1 - \rho)] \\
\omega_{\text{RRP}}(1) &= (1 - \rho)(2 \rho^2 + \rho^3) \\
\omega_{\text{RRP}}(2) &= 0 \\
\omega_{\text{RRP}}(3) &= (1 - \rho) \rho 
\end{align*}
\]

(20)

Also, the normalised probability of \(l\) SCCs aggregated by RRP is

\[
P[F(n) \leq a] = 1 - e^{-a/F}, \quad a \in \{0, 1, 2, \ldots, \infty\} \quad (21)
\]

Owing to the high complexity of RRP, a general expression of \(\omega_{\text{RRP}}(m)\) is analytically unattainable.

The probability of UEs without the use of any SCC, that is, \(P[N_{\text{SCC}} = 0]\), is an important performance metric. Since UEs in the LTE system cannot support CA, a higher probability of UEs without using SCCs means better backward compatibility with LTE for LTE-advanced system.

### 5 Numerical and simulation results

In this section, the performances of various carrier scheduling schemes in the LTE-advanced system are compared. The system main parameters are listed in Table 3. The urban macro (UMa) wireless channel environment is adopted for the purpose of performance studies [21]. It targets the scenario of continuous coverage for urban areas.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of CC ((L))</td>
<td>4</td>
</tr>
<tr>
<td>number of subchannels per CC ((K))</td>
<td>25</td>
</tr>
<tr>
<td>total bandwidth</td>
<td>5 MHz</td>
</tr>
<tr>
<td>distance between sub-carriers ((W_s))</td>
<td>15 KHz</td>
</tr>
<tr>
<td>penetration loss</td>
<td>20 dB</td>
</tr>
<tr>
<td>total BS TX power ((P_{\text{total}}))</td>
<td>49 dBm for 20 MHz</td>
</tr>
<tr>
<td>noise figure of UE</td>
<td>5 dB</td>
</tr>
<tr>
<td>thermal noise spectral density</td>
<td>–174 dBm/Hz</td>
</tr>
<tr>
<td>fading model</td>
<td>Rayleigh</td>
</tr>
<tr>
<td>length of time slot</td>
<td>1 ms</td>
</tr>
<tr>
<td>average size of flows ((F_n))</td>
<td>500 Kb</td>
</tr>
<tr>
<td>carrier frequency ((f_c))</td>
<td>2 GHz</td>
</tr>
<tr>
<td>channel model</td>
<td>UMa defined in 3GPP [21]</td>
</tr>
<tr>
<td>path loss</td>
<td>(PL = 194.66 + 39.1 \log_{10} d)</td>
</tr>
</tbody>
</table>

![Fig. 4](image-url)  

**Fig. 4** AST and throughput with IC and JC

a AST  
b Average throughput
hexagonal deployment is used with an inter-site distance of 500 m and antennas mounted clearly above rooftop. Non-line-of-sight is common for this scenario. Base stations operate at 2 GHz.

5.1 Comparison of queueing performance

In Fig. 4, we firstly compare the performance of the AST and ATP of the JC and IC schemes with three dispatching algorithms. From the viewpoint of performance, the JC scheme is the best, verified by not only the simulation results, but also the analysis in (9). For the IC scheme, its performance is poorer than that of JC and depends on the chosen dispatching algorithm, attributed to the unbalanced load of each CC in the IC scheme. With the RA method, the AST of each flow is 4 (number of CCs) times that of JC, whereas the ATP is 1/4 exactly as given in (10). The performance of the JSQ dispatching algorithm is the best with the IC scheme, and very close to the analytical results, which are the approximation results given in Table 2. Moreover, the RR method can achieve a better performance than RA. The analytical results in the $E_k/M_k$ – PS system with the same mean capacity $\mu$ are used as the upper bounds of the performance achievable by the RR method. As shown in Fig. 4, the simulated performance is poorer than the analytical bound when the RR method is applied in the case of IC. The difference between the simulation result and the analytical bound increases when the traffic load becomes heavier. However, when the arrival rate of flows is low, the analysis bound is very close to the simulation result, which can be used as a good approximation.

Next, Fig. 5 shows both the simulation and analytical results of the AST and throughput of the system with the SLQ and RRP methods when employing the DAC scheme. Both the JC and two DAC schemes are able to ensure full utilisation of radio resources. Thus, they have the same performance in terms of the sojourn time and throughput, which is consistent with the analyses in Section 4. Unlike the IC scheme, the performance of DAC is independent of the underlying dispatching method.

5.2 Comparison of the number of SCCs

As discussed previously, with the JC scheme every UE is served by all the aggregated CCs simultaneously, while DAC can greatly reduce the need of CA. The power consumption per user increases with the number of CCs an UE has to receive and process [3]. The less probable each UE uses multiple SCCs, the less processing complexity and power the UE requires. We use (18) and (21) to calculate the probability of the number of SCCs that each UE may use in the IBS, which can be used as the optimal bound.
The arrival rate of flows $\lambda$ is 20 s$^{-1}$, implying that the average interval between arriving flows is 50 ms. The cumulative distribution functions of the numbers of SCCs are depicted in Fig. 6a.

Let $p(N_{SCC} = 0)$ be the probability of an UE without using any SCCs, indicating that the UE uses only the PCC for transmission, which is the same as LTE Rel-8. The SLQ method has a higher $p(N_{SCC} = 0)$ than that of RRP when using the same dispatching method. In the SLQ scheme, the SCCs help only the PCC with the longest queue, leaving other PCCs with a non-empty queue alone. Thus, SLQ is better backwards compatible with the LTE system, whose UEs are unable to communicate with multiple CCs simultaneously.

The means and variances of the numbers of SCCs under various scheduling schemes are shown in Fig. 6b. Although SLQ has higher $p(N_{SCC} = 0)$, RRP has a smaller variance, implying better UE fairness and less frequent carrier switching. In addition, the dispatching method also has an impact on the probability of the number of SCCs, as demonstrated in Fig. 6b. More balanced traffic using certain dispatching methods will lead to a lower possibility that UEs use multiple carriers. In terms of traffic balance, JSQ is the best, and performs fairly close to the IBS. In our simulation of the four-CC system, the JC scheme makes all the UEs always having three SCCs, more than the proposed DAC scheme as shown in Fig. 6. Therefore it incurs higher power consumption and processing complexity per user. Therefore it incurs higher power consumption and processing complexity per user.

6 Conclusion

In this paper, we propose a novel DAC scheme for the CA-based LTE-advanced system. The system performance, including the sojourn time and throughput, is analysed using queueing theory. Among all the schemes, the JC scheme is demonstrated to perform best in terms of the sojourn time and throughput. With the IC schemes, different dispatch algorithms lead to different performances, in which the JSQ is preferred from the performance view. On the other hand, the performances of the systems with the DAC are not independent on the dispatch algorithm. Moreover, the processing complexity of implementing CA for UEs is investigated based on the distribution of the number of secondary CCs.

The DAC scheme enjoys the same full utilisation as the JC scheme and thus provides the best transmission performance. Moreover, it is able to reduce the energy consumption and processing complexity at the UE. The SLQ and RRP methods have their respective features. The former has a high possibility of UEs without using any secondary CC such that it can support more LTE UEs, improving the backward compatibility of the LTE-advanced system. The latter results in better fairness among UEs, although it may use secondary CCs aggregated with PCCs more frequently than the SLQ scheme. In addition, the JSQ dispatching method can greatly improve the balance of the traffic load on each CC, which can enhance the performance of the IC scheme and reduce the number of secondary CCs in the DAC scheme, bringing it close to the IBS. It is concluded that the DAC scheme has the advantages of ensuring full resources utilisation while reducing energy consumption as well as processing demands at the UE.

7 Acknowledgments

This work was funded in part by China NSFC (no. 61271183), Program for New Century Excellent Talents in University (NCET-11-0600) and the National Key Technology R&D Program of China under grant 2013ZX03003005.

8 References

1 3GPP TR 36.913 V10.0.0; ‘Requirement for further advancements for E-UTRA (LTE-Advanced)’, April 2011
2 3GPP TR 36.912 V10.0.0; ‘Feasibility study for Further Advancements for E-UTRA (LTE-Advanced)’, March 2011
16 Kelly, F.P.: ‘Reversibility and stochastic network’ (Wiley, 1979)
17 Ramaswamy, V.: ‘The sojourn time in the GI/M/1 queue with processor sharing’, J. Appl. Probab., 1984, 21, pp. 437–442
21 3GPP TR 36.814, v2.0.0; ‘Further advancements for E-UTRA, physical layer aspects’. March 2010

9 Appendix

Insensitivity of queue M/G/1/1 – PS on distribution of service time

Firstly, we derive an auxiliary result, which is valid for a general G/G/1 queue. Denote

$x$ is the number of transmitted bits of a flow;

Reference:

www.ietdl.org
$\lambda [1 - F_x(x)] \xrightarrow{T(x)} N(x + \Delta x) - N(x) \xrightarrow{T(x + \Delta x)}$

Fig. 7 Subsystem defined with the amount of received bits

- $F_x(x)$ is the cumulative distribution function of the length of each flow;
- $N(x)$ is the average number of flows in the system with less than $x$ bits having been transmitted;
- $T(x)$ is the average time spent in the system by the flows of which $x$ bits have been transmitted; and
- $n(x) = (dN(x))/dx)$ is the average flow density.

We then apply Little’s theorem to a subsystem defined as follows. As shown in Fig. 7, a flow arrives at the subsystem when the amount of received bits is more than $x$; the flow has then spent $T_x$ in the $G/G/1$ system on average. A flow departs from the subsystem when the amount of received bits is more than $x + \Delta x$; the flow has then spent $T(x + \Delta x)$ in the $G/G/1$ system on average.

Then, no flows exit in the subsystem on completion of the job. Finally when $\Delta x \to 0$, the discreteness of bits becomes irrelevant. As flows arrive at the system with rate $\lambda$ and the fraction of $1 - F_x(x)$ reach the ‘service age’ $x$, the arrival rate at the box, that is, the subsystem, is $\lambda (1 - F_x(x))$. The mean delay of a flow in the box is $T(x + \Delta x) - T(x)$. According to Little’s theorem, we have

$$N(x + \Delta x) - N(x) = \lambda [1 - F_x(x)] [T(x + \Delta x) - T(x)]$$

Divided by $\Delta x \to 0$, the auxiliary result is given by

$$n(x) = \lambda [1 - F_x(x)] \frac{dT(x)}{dx}$$

(23)

On the other hand, we can directly deduce that

$$n(x) = n(0) [1 - F_x(x)]$$

(24)

This is because all the flows in a PS queue are served with an equal amount of resources. As a result, the ‘service ages’ increase at the same rate. The difference in flow density with respect to the service age rises, that is, $n(x)$, due only to the departure of the flows on completion of their transmission. By age $x$, a fraction $F_x(x)$ of the flows has departed and a fraction $1 - F_x(x)$ of them remains in the $G/G/1$ system.

By combining (23) and (24), we arrive at

$$\frac{dT(x)}{dx} = \frac{n(0)}{\lambda}$$

(25)

and $T(x) = 0$ if $x = 0$. Then, we have

$$T(x) = \frac{n(0)}{\lambda} x$$

(26)

which is the average time spent in the $G/G/1$ system by flows with age $x$, and also the total mean delay of those flows whose service demand is also $x$ bits, that is, the mean delay conditioned on the service requirement.

Furthermore, one can deduce that

$$\lim_{x \to \infty} T(x) = \frac{x}{C(1 - \rho)}$$

(27)

where $C$ is the service capability of the system, and $\rho = \lambda/\mu$. The arrival of a very big job is a rare event. Such a big job stays in the system for a very long time. Meanwhile, all other small jobs arriving at the system pass by. The big job sees effectively the service rate remaining from the other jobs, that is, $C(1 - \rho)$.

Thus, the coefficient of $x$ in (5) is $1/C(1 - \rho)$ as shown below

$$T(x) = \frac{x}{C(1 - \rho)}$$

(28)

By averaging (7) for the conditional delay with respect to the distribution of the flow length, and then applying Little’s result, we obtain the following mean results

$$E[T] = \frac{1}{\mu}, \quad 1/\mu = E[X]/C$$

(29)

$$E[N] = \frac{\rho}{1 - \rho}$$

(30)

The distribution of the serving time is not assumed in the derivation process. Therefore the mean performance of the $M/G/1$ − PS queue is insensitive.