IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 November 2023, accepted 8 December 2023, date of publication 12 December 2023,
date of current version 18 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3342073

RESEARCH ARTICLE

Dual Connectivity in Heterogeneous Cellular Networks: Analysis of Optimal Splitting of Elastic File Transfers Using Flow-Level Performance Models

JOHN O. OLAIFA, (Graduate Student Member, IEEE), AND DOGU ARIFLER, (Senior Member, IEEE)
Department of Computer Engineering, Eastern Mediterranean University, 99628 Famagusta, Turkey
Corresponding author: Dogu Arifler (dogu.arifler@emu.edu.tr)

ABSTRACT The dual connectivity feature in heterogeneous cellular networks can be used to improve the download performance for elastic applications by splitting a file transfer over two connections. We employ two parallel processor sharing queues along with a heavy-traffic approximation to develop an extended framework that allows for analysis of file download performance in dual connectivity enabled networks from a relatively less-investigated yet more tractable flow-level perspective rather than a packet-level perspective. Unlike existing models, the framework developed jointly accounts for different transmission capacities and utilizations of base stations, thus enabling a proper and comprehensive assessment of user-perceived file transfer delays. We analyze the optimum file splitting ratio for reducing download delays using convex optimization and validate our findings via both queueing network and flow-level wireless simulations. Our in-house flow-level wireless simulator takes into account user locations and macroscopic propagation characteristics of wireless channels in order to create a realistic evaluation environment; we observe that optimal splitting under heavy-traffic conditions can result in up to 60% reduction in download delays for commonly encountered wireless system specifications when the macrocell and small cell base stations operating with different transmission capacities both have high utilizations. We further illustrate that our flow-level model can successfully incorporate interfering sources and different transmit powers which can easily be subsumed into the transmission capacities used in the model. Overall, the results presented show that it is indeed crucial to consider different transmission capacities as well as utilizations of base stations when determining the optimum splitting ratio.

INDEX TERMS Dual connectivity, flow-level performance analysis, heavy-traffic approximation, heterogeneous cellular networks, processor sharing queues.

I. INTRODUCTION

In wireless networks where an area is covered by multiple cells, users can potentially improve their throughput performance by making concurrent connections to more than one base station (BS) operating at different frequency bands. The dual connectivity (DC) feature, which is available in 4th Generation (4G) and beyond cellular networks, was introduced in Release 12 of the 3rd Generation Partnership Project (3GPP) Long Term Evolution (LTE) specification to support simultaneous connections between a user equipment (UE) and two BSs in heterogeneous network deployment scenarios [1], [2]. In heterogeneous networks, DC is enabled by a macrocell BS and a small cell BS. The traffic destined to a DC user can be split at the macrocell BS or inside the core network and delivered simultaneously by the macrocell

The associate editor coordinating the review of this manuscript and approving it for publication was Faissal El Bouanani.

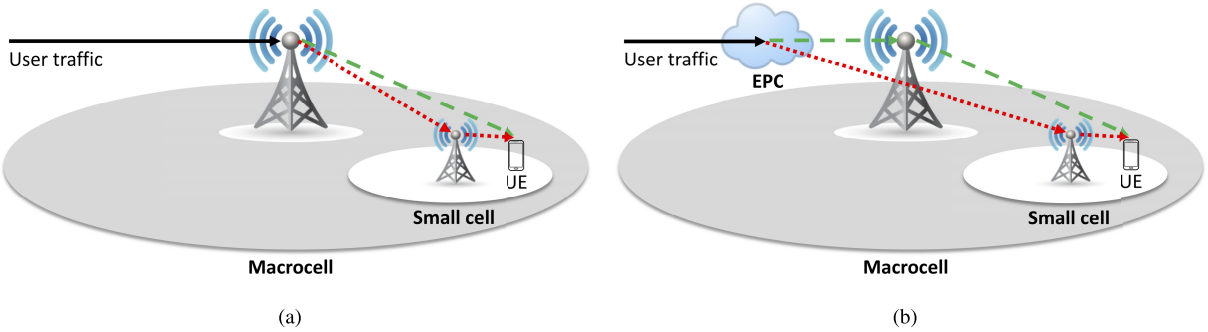

FIGURE 1. Typical dual connectivity deployment options for the downlink. User traffic can be split (a) at the macrocell base station or (b) in the Evolved Packet Core (EPC). The split traffic is reassembled at the user equipment (UE).

and small cell BSs as depicted in Fig. 1. In case the split occurs at the macrocell BS, a fraction of the traffic is diverted to the small cell BS via a very high-speed backhaul link. The split traffic is reassembled at a UE, commonly by the Packet Data Convergence Protocol (PDCP). In addition to increasing throughput performance, DC can improve reliability, especially for ultra-reliable low latency communications, by sending redundant data using multiple BSs. DC can also enhance mobility robustness with less frequent handoffs by enabling simultaneous connections to two BSs, even from different cellular generations or different radio access technologies such as 4G BSs (eNodeBs), 5G BSs (gNodeBs), and WiFi access points [3], [4], [5], [6], [7]. Recent field trials demonstrate impressive downlink and uplink rates in DC-enabled networks. In March 2022, AIS, Qualcomm, and ZTE jointly announced the world's first 5G New Radio (NR)-DC showcase in the field in Korat, Thailand to achieve 8.5 Gbps peak rate for the downlink and 2.17 Gbps peak rate for the uplink [8]. Later in January 2023, Deutsche Telekom, Ericsson, and Qualcomm implemented a priority scheduling mechanism in a 5G Standalone NR-DC network in Bonn, Germany using both millimeter wave and mid-band frequencies to satisfy quality of service requirements, particularly for the uplink. The created test bed resulted in 5 Gbps peak rate for the downlink and 700 Mbps peak rate for the uplink [9].

A. RELATED WORK

Simultaneous use of multiple connections for increasing user performance has been reviewed in detail in [10]. In their work, the authors categorize multiple connectivity solutions based on adaptability to network and traffic conditions and on the layer of the protocol stack at which they operate. In [7], the authors outline the challenges, benefits, and open issues of multi-connectivity, and test bed results demonstrate improvement in throughput for both Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) traffic. In this work, we focus on performance benefits of DC, a special case of multi-connectivity; DC has been widely adopted by the 3GPP due to a relatively manageable overhead of maintaining two concurrent connections [7].

Recently, there have been numerous simulation studies that report performance gains in the downlink through the use of DC [1], [2], [3], [11], [12], [13], [14], [15], [16], [17]. The extent of performance gains is often even higher for cell-edge users. However, the reported gains through the use of DC vary considerably among these studies since the parameters used and the assumptions made are different. In [18], the reliability of low-latency 5G downlink communication with DC through packet duplication is evaluated via simulations. Another simulation study investigates potential performance improvement that might be obtained in the uplink [19]. A disadvantage of simulation-based studies is that they generally require significant amounts of run-time if a wide range of parameters and setups is to be explored.

Despite the existence of numerous performance reports based on simulations, there have been only a few analytical modeling studies on DC. In [20], a queueing model for packet transmissions to a DC-enabled UE using multiple radio access technologies is described, and an optimization problem is formulated for finding the delay-optimal packet traffic splitting strategy. Queueing models are also used in [21] to show how packet delays experienced by a DC-enabled UE vary as the packet traffic splitting ratio between LTE and 5G links is changed. Even though some useful insights might be provided by these *packet-level* models, some networking researchers argue that the dynamics of Internet Protocol (IP) packet traffic are too complex, and the models at the packet level often ignore control mechanisms so that a realistic yet tractable analysis of network performance is generally not possible (see, for instance, [22], [23], [24] and the references therein).

Although there is no standard definition, a flow can be loosely defined as a unidirectional sequence of packets which are close to each other in time and share common identifiers such as source and destination addresses; for instance, packets corresponding to a file download constitute a flow. The traffic generated by TCP-mediated transfer or download of files such as Web pages, emails, documents, and images is called elastic traffic. Unlike packet-level models, *flow-level* performance models take into account how bandwidth is shared dynamically among a randomly fluctuating number of

concurrent users generating elastic traffic. Flow-level models often enable a more tractable analysis of user-perceived performance metrics such as user throughput and response time of elastic file transfers. In [23], the authors indicate that the quality of service perceived by users of elastic traffic is more appropriately determined at the flow level. The key argument is that the performance of elastic applications is not sensitive to the delay of each packet but is rather determined by the time taken to transfer an entire document or the response time; therefore, the perceived quality of service of TCP-mediated transfers will critically depend on flow-level dynamics. It is important to point out that flow-level models are extremely useful in traffic engineering. This is further elaborated in [24], where the authors argue in favor of these models and discuss how they leverage insights into network performance at low computational complexity by providing analytical expressions for key user performance measures of interest. Hence, in our work, we adopt a flow-level modeling approach for evaluating the performance of DC in heterogeneous cellular networks.

Dynamic bandwidth sharing of TCP flows with similar round-trip times and packet loss probabilities has successfully been modeled by a processor sharing (PS) queue, where each TCP flow receives a fraction $1/n$ of the total bandwidth when there are n active flows [22], [23]. In this model, it is assumed that fair sharing among TCP flows is immediately achieved independently of their sizes. Bandwidth sharing of heterogeneous TCP flows with different round-trip times and packet loss probabilities can be modeled by a discriminatory processor sharing (DPS) queue, where flows share the bandwidth proportionally to their weights [25]. PS and DPS models are applicable to both wired and wireless networks [26]. It should also be noted that wireless channel fluctuations ''average out'' at the flow level for different classes of users, resulting in deterministic service rates for a given class of statistically identical users with similar flow size and rate characteristics [27].

So far, only a few studies have considered flow-level PS models for analyzing the performance of DC. Previously, splitting a file download into multiple concurrent connections has been modeled by using parallel PS queues [28]. The described model, which assumes that all service stations operate with the same transmission capacity, has been used to find optimum splitting ratios for achieving minimum download times. In their work, the authors have observed that although there are correlations between sojourn times of file fragments, such correlations have negligible effects on the download times of entire files. The authors have also noted that the download times are ''nearly insensitive'' to the job size distribution function employed. Yet, we note that the study described does not consider the effects of any physical layer characteristics on the queueing model parameters employed. An analytical framework for evaluating the performance of LTE/WiFi multihoming is presented in [29]. The authors consider both network-centric and user-centric resource allocation strategies. Their user-centric network-assisted approach involves using PS queues to determine the optimal file split that maximizes user throughput in multihomed scenarios. The study described considers service stations with different transmission capacities. It is important to point out, however, that the average user throughputs reported are obtained using the reduced service rate approximation. As demonstrated through extensive numerical experiments in [28], this approximation is not accurate for obtaining optimum splitting ratios in parallel PS queues with highly asymmetric background loads; the relative errors observed in such cases may exceed 20%.

In [30], the authors compare static and dynamic traffic splitting policies for DC. Dynamic splitting policies may result in better performance since the transmission capacity allocated for a flow may change during the sojourn time of the flow and routing decisions are made for individual TCP segments. Mapping of TCP segment-level dynamics into a flow-level model is described in [31]; yet, this mapping depends on WiFi parameters and is limited to WiFi networks. Performance evaluation for dynamic traffic splitting via learning-based and heuristic methods is presented in [32] and [33], where multiple wireless networks with both the same and different capacities are considered; these studies again employ the method described in [31] for mapping WiFi and TCP parameters into a flow-level model and are hence tailored for WiFi networks.

Particularly with the adoption of 5G non-standalone deployments, there has been another line of recent research that focuses on algorithmic aspects of splitting traffic over dual connections. In [34], the authors introduce the notion of a ''maximum out-of-order depth'' metric and propose two low-complexity traffic splitting mechanisms based on UE feedback and BS observation to minimize this metric in order to achieve high throughputs in DC settings. A traffic steering mechanism that considers discontinuous reception at a UE with DC capability is described in [35]. The mechanism dynamically directs traffic over the two links to reduce power consumption without incurring additional packet delays. The study described in [36] presents a DC flow control scheme that takes into account the effect of blockages on signal quality of 5G millimeter waves for determining the best traffic splitting ratio. Two different scheduling mechanisms are devised to be coupled with the presented flow control scheme for resolving out-of-order packet deliveries to achieve high throughputs or to minimize packet buffering times. In [37], the authors propose another DC flow control mechanism that uses information on assigned radio resources and buffering delay statistics when making traffic splitting decisions. The same authors later describe a fast data recovery mechanism for DC aimed at minimizing interruption times due to packet losses or reorderings resulting from mobility and link failures [38].

Finally in [39], the throughput performance and fairness properties of a relatively newer transport protocol, namely

Quick UDP Internet Connections (QUIC), are investigated experimentally over DC settings. The network conditions and DC parameters for which performance improvements can be achieved are explored in detail. The authors conclude by pointing to possible analytical modeling of DC as part of their future work.

B. CONTRIBUTIONS AND ORGANIZATION

Despite significant progress in the field as outlined in Section I-A, there is still a need for a well-validated analytical framework that captures flow-level performance characteristics, and unlike existing work, jointly accounts for different transmission capacities and utilizations of BSs, thus enabling a proper investigation of user-perceived file transfer delays with DC. Our main aim in the research described here is to fulfill this need and present an analytical flow-level framework to be able to factor in parameters necessary to analyze heterogeneous DC scenarios. We build upon the parallel PS queueing model and the heavy-traffic approximation in [28] to develop an extended model where BSs may operate with different transmission capacities due to different physical layer characteristics. The extended model developed opens the way for analysis of file download performance in more realistic DC settings from a relatively less-investigated yet more tractable flow-level perspective rather than a packet-level perspective. In particular, our work has the following specific aims:

(i) to present a parallel PS queueing model that can be used to analyze download delays while allowing for different transmission capacities and utilizations at service stations processing TCP flows with similar characteristics;
(ii) to investigate the convexity of the expression we derive for the download delay based on a heavy-traffic approximation;
(iii) to find optimal file splitting strategies for elastic downloads via convex optimization, and suggest guidelines to determine when deployment of DC solutions is beneficial; and
(iv) to validate our extended model via both queueing network simulations and more realistic flow-level wireless simulations, with physical layer parameters subsumed into the transmission capacities used in the model.

We note that allowing service stations to have different capacities is necessary for a more realistic analysis since different cells may provide different spectral efficiencies. On the other hand, convexity of the download delay as a function of the splitting ratio guarantees that the splitting ratio obtained in the optimization problem results in a globally minimum delay. Convexity also enables us to employ the state-of-the-art polynomial-time optimization algorithms for determining the optimum splitting ratios. We further note that our analytical model allows for a rapid exploration of a wide range of parameters, a significant advantage for finding optimal DC deployment solutions.

The rest of the paper is organized as follows. Models and methods that are used for analyzing DC are described in Section II. This section is divided into three subsections; the first subsection details the development of our analytical framework which represents the main flow-level model used in the current study, whereas the other two subsections provide descriptions of queueing network and flow-level wireless simulations that are employed to validate our model. The analytical results obtained for different scenarios are presented and benchmarked against simulations in two separate subsections in Section III. Finally, Section IV concludes the paper.

II. MODELS AND METHODS

A. PARALLEL PROCESSOR SHARING QUEUEING MODEL

In cellular networks that enable DC, two independent PS queues with service stations having transmission capacities c and κc bits/s, respectively, can be used to model the download performance perceived by a DC user. As in [28], we will refer to the traffic destined to a typical DC user as the *foreground* traffic; the foreground traffic will be serviced concurrently by the two parallel PS queues. The traffic belonging to other (non-DC) users that co-exist in a given PS queue will be referred to as the *background* traffic. The model considered in this work is illustrated in Fig. 2. The fork point represents the component in the macrocell BS or the core network where splitting of a file takes place. The join point represents the component that merges the file at the DC-enabled UE and generally executes a synchronization operation in parallel processing systems [40]. We suppose that file size information can be returned by the content server at the initiation of TCP-mediated downloads (via, for example, File Transfer Protocol's (FTP) `size` command or Hypertext Transfer Protocol's (HTTP) `HEAD` request) and captured by the traffic splitting/steering component represented by the fork point. We will make the usual assumption that network performance is mainly limited by the cellular network, and the public or private backbone links that connect the servers of content providers to the cellular system have much larger capacities.

Table 1 tabulates key notation and symbols used in analytical model development. Let random variable B with a general distribution denote the size in bits of a file being transferred to the DC user. The file is split into two fragments using a splitting ratio α so that αB bits are serviced by the first queue and $(1-\alpha)B$ bits are serviced by the second queue. It is assumed that the splitting ratio α is determined upon the initiation of the file download and remains constant throughout the sojourn time.[1] File downloads making up the foreground traffic arrive according to a Poisson process

[1] Since file fragment sizes in bytes must be integers, the system may slightly adjust the value of α when implementing file splitting in practice; such slight adjustments will only have negligible influence on the outcome.

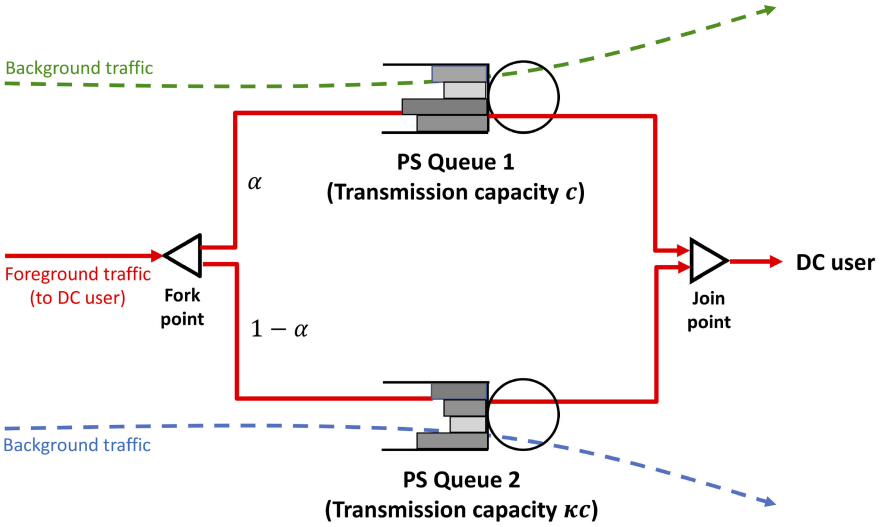

FIGURE 2. Dual connectivity (DC) model with processor sharing (PS) queues. A file belonging to the foreground traffic is split at the fork point using a splitting ratio α and reassembled at the join point. Service stations of Queue 1 and Queue 2 have transmission capacities c and κc, respectively.

TABLE 1. Key notation and symbols used in analytical model development.

α	Splitting ratio
α^*	Optimum splitting ratio
B	Random variable representing the file size in bits
b	A file size value assumed by B
β	Mean file size in bits given by $\mathbb{E}[B]$
c	Transmission capacity in bits/s of service station 1
κ	Transmission capacity factor for service station 2
λ_0	Arrival rate of foreground traffic in files/s
ϱ_0	Load due to foreground traffic and normalized by c
ρ_i	Utilization of service station $i = 1, 2$ due to background traffic
τ_i	Service time requirement in seconds of a file's ith fragment at service station $i = 1, 2$
T_i^α	Random variable representing the generic sojourn time in seconds of a file's ith fragment through queue $i = 1, 2$ for a given α
$T_i^\alpha(\tau_i)$	Random variable representing the conditional sojourn time in seconds of a file's ith fragment having service requirement τ_i through queue $i = 1, 2$ for a given α
T^α	Random variable representing the generic sojourn time in seconds of a file download for a given α
ξ_i	Total utilization of service station $i = 1, 2$
ξ	Maximum of the utilizations of service stations

with rate λ_0 per second. The arrival process modeling the background traffic for each queue is also assumed to be Poisson; file sizes making up the background traffic are assumed to be generally distributed as well. Hence, each queueing system considered in isolation is modeled as an M/G/1-PS queue.

The utilizations of the M/G/1-PS queues due to the background traffic are denoted by ρ_1 and ρ_2, respectively. The total utilizations of the queueing systems (including the ones due to the foreground traffic) are then expressed as

$$\xi_1 = \rho_1 + \alpha\varrho_0, \tag{1}$$

and

$$\xi_2 = \rho_2 + (1-\alpha)\varrho_0/\kappa, \tag{2}$$

with $\xi_1 < 1$ and $\xi_2 < 1$ for stability, and $\varrho_0 = \lambda_0\mathbb{E}[B]/c$ that can be interpreted as the load due to the foreground traffic and normalized by a nominal capacity of c bits/s. A value of α is said to be *feasible* if the stability conditions are satisfied. The factor κ accounts for different transmission capacities of the service stations.

Let T_1^α and T_2^α be the generic sojourn times of the fragments of an arbitrary file at the first and second queues for a splitting ratio α. Then, the generic sojourn time of the file through the system will be $T^\alpha = \max\{T_1^\alpha, T_2^\alpha\}$ since both fragments must arrive before declaring a download complete. Our goal is to find the minimizing α, denoted by α^*, of the following optimization problem:

$$\begin{aligned} \underset{\alpha}{\text{minimize}} \quad & \mathbb{E}[T^\alpha] \\ \text{subject to} \quad & \rho_1 + \alpha\varrho_0 < 1, \\ & \rho_2 + (1-\alpha)\varrho_0/\kappa < 1, \text{ and} \\ & 0 \le \alpha \le 1. \end{aligned} \tag{3}$$

We note that finding a closed-form solution for the distribution of sojourn times in M/G/1-PS queues is not possible [28]. Hence, an analytical solution of the optimization problem in (3) does not exist. Consequently, as in [28], we resort to heavy-traffic approximations which provide us with an analytical expression for solving the optimization problem. Unlike [28], however, we take into account different transmission capacities of the service stations of the queues. We point out that the usefulness of heavy-traffic approaches has been extensively verified in modeling of controlled queueing and communication networks even with non-heavy traffic [41].

Set $\xi = \max\{\xi_1, \xi_2\}$. Consider files of size $B = b$. Also, define the service time requirements of the fragments as $\tau_1 = \alpha b/c$ and $\tau_2 = (1-\alpha)b/(\kappa c)$. For $i = 1, 2$, let $T_i^{\alpha}(\tau_i)$ represent the conditional sojourn times of fragments with service time requirement τ_i at the ith service station; it will be assumed that these fragment sojourn times are independent. We use the heavy-traffic approximation from [28] and [42] to obtain the conditional sojourn times. At the ith queue for which $\xi_i = \xi$, the marginal distribution for the conditional sojourn time $T_i^{\alpha}(\tau_i)$ as $\xi \uparrow 1$ (heavy traffic) is given by

$$\lim_{\xi \uparrow 1} \mathbb{P}\left((1-\xi)T_i^{\alpha}(\tau_i) > t\right) = \exp(-t/\tau_i) \tag{4}$$

with a feasible value of α. Note that (4) is valid for the heavily-loaded queue. It is remarked in [28] that under heavy-traffic conditions, approximating the fragment sojourn times through the non-heavily-loaded queue using an exponential distribution has only a negligible effect on the mean sojourn time. Hence, the marginal distribution for the sojourn time $T_i^{\alpha}(\tau_i)$ at the ith queue for which $\xi_i \neq \xi$ can also be approximated as:

$$\mathbb{P}\left(T_i^{\alpha}(\tau_i) > t\right) \approx \exp\left(-\frac{1-\xi_i}{\tau_i}t\right). \tag{5}$$

As a result, the conditional sojourn time $T^{\alpha}(b)$ of a file of size $B = b$ is the maximum of two independent exponential random variables $T_1^{\alpha}(\tau_1)$ and $T_2^{\alpha}(\tau_2)$ in heavy traffic. The mean values of the aforementioned exponential random variables are

$$\mathbb{E}[T_1^{\alpha}(\tau_1)] = \frac{\alpha b/c}{1-\xi_1} \tag{6}$$

and

$$\mathbb{E}[T_2^{\alpha}(\tau_2)] = \frac{(1-\alpha)b/(\kappa c)}{1-\xi_2}. \tag{7}$$

The mean of the maximum of two independent and identically distributed (i.i.d.) exponential random variables with means $1/\mu_1$ and $1/\mu_2$ is given by $1/\mu_1 + 1/\mu_2 - 1/(\mu_1+\mu_2)$ (see Appendix A in [28]). Since the expectations in (6) and (7) are linear in b, the unconditioned expectation $\mathbb{E}[T^{\alpha}]$ can be written by replacing b with $\beta = \mathbb{E}[B]$:

$$\mathbb{E}[T^{\alpha}] = \frac{\beta}{c}\left(\frac{\alpha}{1-\xi_1} + \frac{(1-\alpha)/\kappa}{1-\xi_2} - \frac{\alpha(1-\alpha)}{(1-\alpha)(1-\xi_1)+\kappa\alpha(1-\xi_2)}\right). \tag{8}$$

Finally, substituting the utilizations in (1) and (2) into (8), we have:

$$\mathbb{E}[T^{\alpha}] = \frac{\beta}{c}\left(\frac{\alpha}{1-\rho_1-\alpha\varrho_0} + \frac{(1-\alpha)}{\kappa(1-\rho_2)-(1-\alpha)\varrho_0} + \frac{\alpha^2-\alpha}{2\varrho_0\alpha^2+(-2\varrho_0+\rho_1+\kappa(1-\rho_2)-1)\alpha+(1-\rho_1)}\right). \tag{9}$$

Note that the expression for $\mathbb{E}[T^{\alpha}]$ in (9) is the sum of three terms: two convex linear fractional functions and a quasiconvex quadratic-over-quadratic fractional function. Since the sum of convex and quasiconvex functions is not necessarily convex [43], we verify the convexity of (9) as a function of feasible α via numerical analysis. Using the fact that $f(x)$ is convex if and only if $f''(x) \geq 0$ for all x in the domain of f (and the domain of f is a convex set) [43], we use MATLAB®'s Symbolic Math Toolbox™ to evaluate the second derivative of the function in (9). For $\kappa = 1, 2, 3$, and 4, we exhaustively evaluate and check the value of the second derivative over feasible α values by varying α from 0 to 1 using a step size of 0.01, varying ρ_1 and ρ_2 from 0 to 0.95 using a step size of 0.05, and varying ϱ_0 from 0 to 1.95 using a step size of 0.05. The resulting values from the exhaustive check are all greater than or equal to zero, thus establishing the convexity of the function. Since the constraint set is convex, the optimization problem in (3) is convex. Hence, any numerical optimization algorithm that minimizes the function finds the global minimum. We employ MATLAB®'s Optimization Toolbox™ `fmincon` routine and opt for the Interior Point Algorithm [43], providing the routine with the gradient of the objective function in order to have more accurate results more efficiently.

B. QUEUEING NETWORK SIMULATIONS

We assess the validity of the analytical approximation derived in Section II-A by simulating a DC system using Java Modeling Tools 1.0.3 [44]. Simulation of the setup shown in Fig. 2 includes a fork point to create two subtasks for each foreground task (file download). The subtasks are merged at a join point. In a fork-join system, a task is considered complete when all of its subtasks are complete and the task can hence depart the join point. Throughout, we choose to ignore processing latencies related to splitting decisions and file merging in order to focus specifically on delays associated with network transmissions.

Queueing network simulations are conducted for five heavy-traffic scenarios and one relatively lighter-traffic scenario detailed in Table 2. When conducting simulations, we use a mean file size of $\beta = 0.5$ Megabytes and a

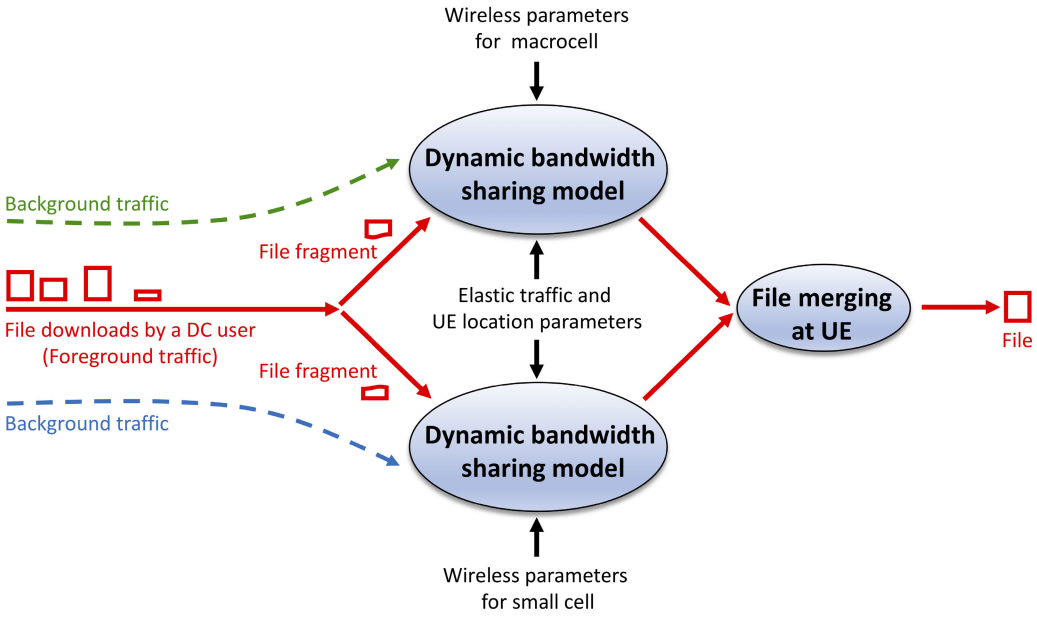

FIGURE 3. Flow-level wireless simulator (FLWS) framework that takes into account elastic traffic parameters and user locations as well as macroscopic propagation characteristics of wireless channels in macrocells and small cells.

TABLE 2. Simulation scenarios under different load conditions defined by ϱ_0, ρ_1, and ρ_2.

Scenario	ϱ_0	ρ_1	ρ_2
1	0.10	0.80	0.80
2	0.10	0.50	0.50
3	0.10	0.75	0.85
4	0.10	0.85	0.75
5	0.10	0.20	0.80
6	0.10	0.80	0.20

nominal bit rate of $c = 16.8$ Mbps.[2] File downloads by the DC user are initiated according to a Poisson process at a rate of $\lambda_0 = 0.4$ per second. In our simulations, service requirements of download fragments are generated from exponential distributions with means calculated to result in foreground utilizations $\alpha\varrho_0$ and $(1-\alpha)\varrho_0/\kappa$ at the first and second service stations, respectively. The background traffic flows at the queues are created similarly, with Poisson arrivals and exponential service time distributions resulting in background utilizations ρ_1 and ρ_2. Sojourn times of file downloads are collected by the simulation tool. Each simulation is run until either the half-width of the 95%-confidence interval is no more than 3% of the sample mean or one million sojourn time samples are collected.

C. FLOW-LEVEL WIRELESS SIMULATIONS

In order to further validate our extended model for realistic wireless environments, we have implemented a flow-level wireless simulator (FLWS) in MATLAB®. Our in-house simulator takes into account elastic traffic parameters and user locations as well as macroscopic propagation characteristics of wireless channels in macrocells and small cells. The implemented simulator framework is illustrated in Fig. 3. Here, users, whose locations are chosen uniformly randomly in the macrocell or the small cell, arrive in time according to a Poisson process, initiate a file download with a size that is exponentially distributed with mean $\beta = 0.5$ Megabytes, and leave the system. The UE locations and elastic traffic parameters related to the download request arrival process and file size distribution are inputs to the dynamic bandwidth sharing model referred to in Section I-A and implemented in FLWS. Small cell users that are designated as DC-enabled can split the file download to use two connections. The splitting ratio α denotes the fraction of a file to be serviced by the macrocell BS. The BSs apply round-robin (RR) scheduling for processing ongoing downloads.[3] With RR scheduling, the rate that can be used for transmission to a user by a BS in a given transmission time interval (TTI) is obtained by dividing the achievable bit rate at the user's location by the number of ongoing downloads at the BS in that TTI.

Before conducting simulations, FLWS generates path loss and shadowing maps using a resolution of 1 m × 1 m. These maps constitute the physical layer wireless parameters used as additional inputs to FLWS (Fig. 3). The transmission capacities c and κc of the macrocell and the small cell BSs are then computed as the harmonic average of the achievable bit rates at each possible location in the cells as justified in [45].

[2]We use some typical values to calculate the downlink capacity of an LTE BS: 100 Physical Resource Blocks (PRBs), each having 84 Resource Elements (REs), can be transmitted in 0.5 ms in 20-MHz systems. The average spectral efficiency of users is assumed to be 1 bit per RE. Different values of κ account for different spectral efficiencies.

[3]RR scheduling treats all active UEs equally and allocates the same amount of time-frequency resources sequentially to the UEs in every TTI in a cyclic manner.

TABLE 3. Flow-level wireless simulator (FLWS) parameters.

Macrocell radius	250 m
Small cell radius	10 m
Available bandwidth per cell	20 MHz
Transmission time interval (TTI)	1 ms
Macrocell BS transmit power	46 dBm
Small cell BS transmit power	20 dBm
Macrocell path loss model	$128.1 + 37.6 \log_{10}(d/1000)$, where user distance d from the BS is in km
Small cell path loss model	$127 + 30 \log_{10}(d/1000)$, where user distance d from the BS is in km
Shadowing model in macrocell	Lognormal with mean 0 and standard deviation 10 dB
Shadowing model in small cell	Lognormal with mean 0 and standard deviation 4 dB (indoor)
Noise density	-174 dBm/Hz

In order to compute the achievable bit rate r at a location, we use a "modified version" of the Shannon formula [46]:

$$r = \eta_{\text{BW}} w \log_2 \left(1 + \frac{\text{SINR}}{\eta_{\text{SINR}}}\right), \tag{10}$$

where w is the bandwidth and SINR is the received signal-to-interference-plus-noise ratio which can be calculated as

$$\text{SINR} = \frac{P_{Rx}}{N + I}. \tag{11}$$

In (11), P_{Rx} is the power received from the serving BS in the considered macrocell or small cell, noise power N is defined as the noise density multiplied by the available bandwidth, and interference I is the total power received from all BSs except the serving one operating at the same frequency. The received power in the absence of small-scale fading effects is obtained by

$$P_{Rx}(\text{dBm}) = P_{Tx}(\text{dBm}) - L(\text{dB}), \tag{12}$$

where P_{Tx} is the transmit power used by a BS and L is the combined path loss and shadowing. We note that small-scale fading effects on the received power are not considered here since such fading processes occur on a much faster timescale compared to flow-level dynamics considered in queueing models [27], [45], [47], [48]. Yet, as in the approaches adopted in [47] and [48], an appropriate function that maps SINR to achievable bit rate *can* account for other physical link-level intricacies of the wireless technology under consideration. As such, the mapping function in (10) includes the bandwidth and SINR correction factors, η_{BW} and η_{SINR}, to allow for a better match for abstraction at the system level of the physical link-level details. We use $\eta_{\text{BW}} = 0.48$ and $\eta_{\text{SINR}} = 0.81$ as reported in [49] for a 20 MHz-bandwidth, 2×2 Multiple Input Multiple Output (MIMO) system that uses RR scheduling. The primary parameters used in FLWS are mostly based on [50] and are tabulated in Table 3.

We first focus on a default case where we assume that there are no interfering sources, i.e. $I = 0$ in (11). The load conditions in this case are the same as those given in Table 2. Here, ρ_1 and ρ_2 correspond to background utilizations of the macrocell BS and the small cell BS, respectively. It is also important to consider the influence of various physical layer parameters on optimal splitting. As such, we next focus on two extended cases where we assess the effects of changing the transmit power of the small cell BS or having interfering sources under the load conditions specified for a selected scenario, namely Scenario 2:[4]

1) In the first extended case, we consider the effects of a ten-fold increase (an increase by 10 dBm) and a ten-fold decrease (a decrease by 10 dBm) in the transmit power of the small cell BS with respect to its default value of 20 dBm. These changes alter the small cell BS capacity κc which can be calculated via the use of the Shannon formula in (10) and harmonic averaging as already described. There are still no interfering sources.

2) In the second extended case, we consider the effects of having other small cell BSs nearby the small cell of interest. All the small cell BSs use the same frequency band for transmission and there is no interference management in the system. As before, we assume that the macrocell BS operates at a different transmission frequency band. In this case, we no longer have $I = 0$ in (11) due to existence of interferers. We investigate the effect of having up to three interfering small cell BSs which are assumed to be actively transmitting all the time. Each of them is located on the corners of an equilateral triangle, 20 m from the center of the small cell whose performance is of interest. Since the focus of our work is to understand the fundamentals of optimally splitting file downloads, we resort to using such a relatively simple, deterministic, and regular deployment scheme for analyzing the effects of interference on DC performance. Although small cell BSs are generally deployed in an uncoordinated manner in reality, which usually requires using stochastic geometry models for analysis, consideration of deterministic regular locations has not been uncommon in the literature [51], [52]. We also assume that the small cell BSs are located indoors and there is an additional loss (wall attenuation) of 20 dB for signals originating from these interferers [53]. Once again, these changes alter the small cell BS capacity κc.

[4]In our analysis of the influence of various physical layer parameters, we select a relatively lighter-traffic scenario (Scenario 2) to allow for substantial reductions in the transmission capacity of the small cell BS without causing a congestion collapse in the network (i.e. an unstable network).

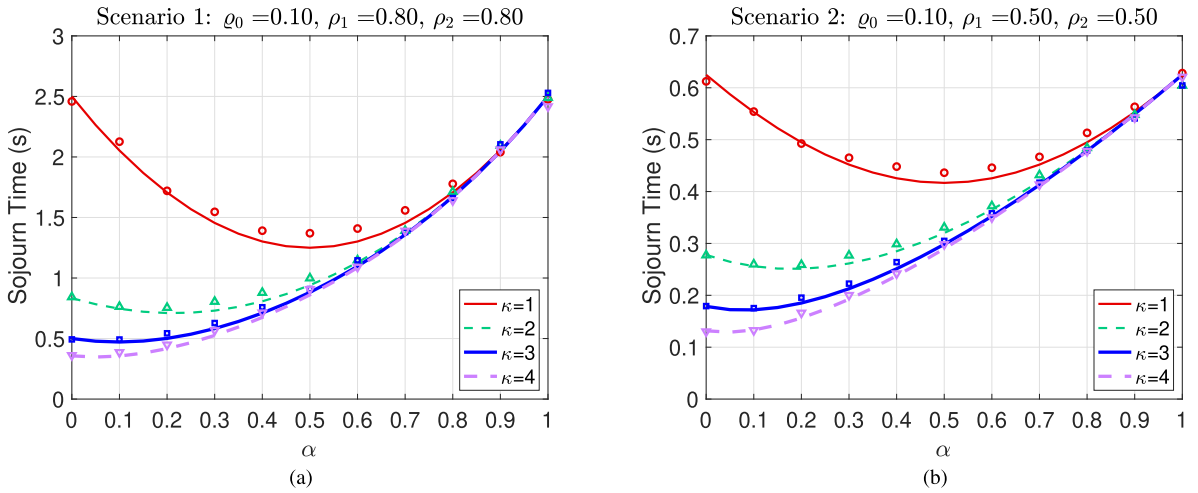

FIGURE 4. Mean sojourn times of file downloads with DC as a function of splitting ratio α for different transmission capacity factors κ of the second service station. Equal background utilizations are considered. (a) Scenario 1: $\varrho_0 = 0.10$, $\rho_1 = 0.80$, $\rho_2 = 0.80$ and (b) Scenario 2: $\varrho_0 = 0.10$, $\rho_1 = 0.50$, $\rho_2 = 0.50$. The lines correspond to analytical results, whereas the markers correspond to queueing network simulation results.

TABLE 4. Average relative errors between the analytical approximation and queueing network simulation results.

	Scenario					
	1	2	3	4	5	6
$\kappa = 1$	4.3%	2.6%	4.1%	3.8%	3.1%	1.7%
$\kappa = 2$	3.7%	2.5%	3.4%	3.9%	2.4%	2.2%
$\kappa = 3$	3.7%	2.4%	3.8%	3.0%	2.6%	2.0%
$\kappa = 4$	3.9%	1.8%	4.2%	2.7%	3.0%	1.6%

TABLE 5. Optimum splitting ratios α^* obtained via the optimization routine.

	Scenario					
	1	2	3	4	5	6
$\kappa = 1$	0.50	0.50	0.70	0.30	0.95	0.05
$\kappa = 2$	0.21	0.18	0.42	0.09	0.81	0.01
$\kappa = 3$	0.10	0.08	0.24	0.03	0.65	0.00
$\kappa = 4$	0.05	0.04	0.14	0.02	0.50	0.00

Overall, the FLWS cases described provide an extensive simulation test bed for further validation of the analytical approximation presented in Section II-A. We note that all the physical layer parameters discussed can be subsumed into the transmission capacities in our flow-level model.

III. RESULTS AND DISCUSSION

A. ASSESSMENT OF THE VALIDITY OF THE ANALYTICAL APPROXIMATION VIA QUEUEING NETWORK SIMULATIONS

The mean sojourn times with DC versus the splitting ratio α obtained with the analytical approximation in (9) and via queueing network simulations are plotted in Figs. 4 to 6 for the six scenarios given in Table 2. In all these figures, the lines correspond to analytical results, whereas the markers correspond to simulation results. Fig. 4 shows the results for Scenarios 1 and 2 where the parallel queue stations have equal background utilizations; the results for heavily-loaded queues (Scenario 1) are shown in Fig. 4(a), whereas the results for queues with lighter loads (Scenario 2) are shown in Fig. 4(b). Fig. 5 shows the results for Scenarios 3 and 4, and exhibits the effects of slight asymmetry in background utilizations on mean sojourn times under heavy-traffic conditions. Fig. 6, on the other hand, shows the results for Scenarios 5 and 6 where there is a larger asymmetry in background utilizations. It can be concluded that the derived approximation agrees well with queueing network simulations; the average relative errors between the approximation and simulation results are tabulated in Table 4 and are all below 5%. The results for Scenario 2 demonstrate that the approximation can also work well in relatively lighter-load scenarios as reported in [41]. In addition, the results for Scenarios 5 and 6 show that the approximation gives accurate results even when the queues have asymmetric loads. Figs. 4 to 6 can be used as a guide to evaluate whether splitting results in lower download times compared to a strategy that uses a single connection ($\alpha = 0$ or $\alpha = 1$). In the case of service stations with equal background utilizations (Scenarios 1 and 2) and $\kappa = 1$, for instance, one observes that the delay with optimum α is reduced by $\sim$ 50% under heavy-traffic conditions (Fig. 4(a)) and yet by $\sim$ 33% under lighter-traffic conditions (Fig. 4(b)) if DC is used as opposed to using a single connection. Hence, the benefit of DC in this case is more pronounced for heavy traffic.

The optimum splitting ratios α^* that are obtained using MATLAB®'s Optimization Toolbox™ are tabulated in Table 5. The maximum number of iterations needed for convergence to a solution is observed to be 20 in the scenarios considered. It is obvious that consideration of different transmission capacities of service stations is important as expected; the optimum splitting ratios are not only

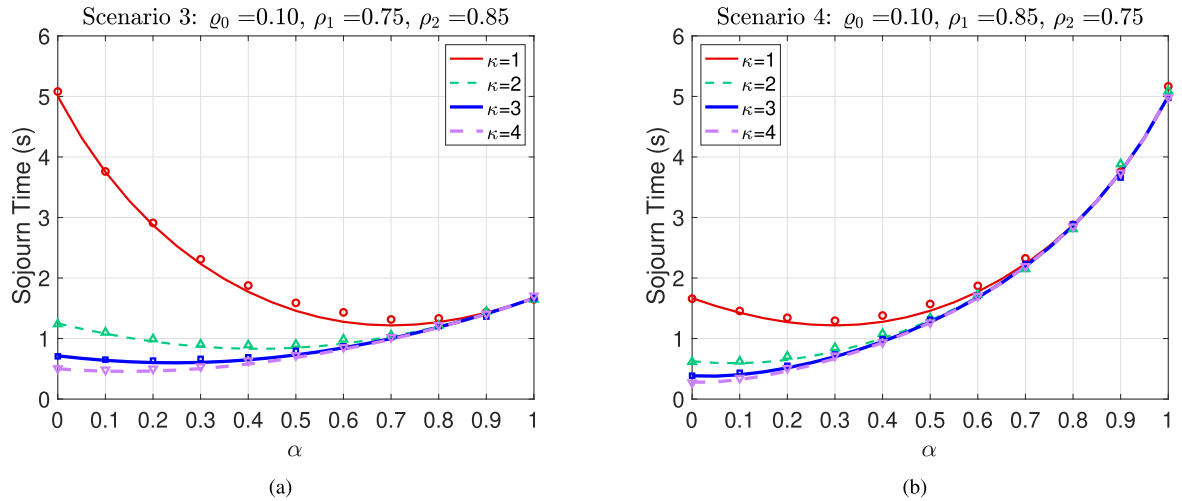

FIGURE 5. Mean sojourn times of file downloads with DC as a function of splitting ratio α for different transmission capacity factors κ of the second service station. A slight asymmetry in background utilizations is considered. (a) Scenario 3: $\varrho_0 = 0.10$, $\rho_1 = 0.75$, $\rho_2 = 0.85$ and (b) Scenario 4: $\varrho_0 = 0.10$, $\rho_1 = 0.85$, $\rho_2 = 0.75$. The lines correspond to analytical results, whereas the markers correspond to queueing network simulation results.

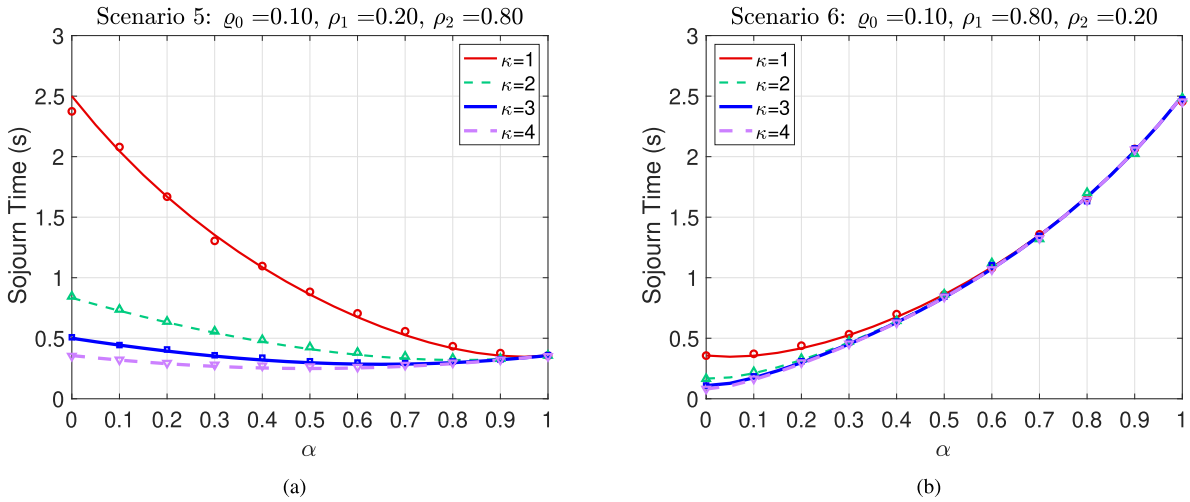

FIGURE 6. Mean sojourn times of file downloads with DC as a function of splitting ratio α for different transmission capacity factors κ of the second service station. A large asymmetry in background utilizations is considered. (a) Scenario 5: $\varrho_0 = 0.10$, $\rho_1 = 0.20$, $\rho_2 = 0.80$ and (b) Scenario 6: $\varrho_0 = 0.10$, $\rho_1 = 0.80$, $\rho_2 = 0.20$. The lines correspond to analytical results, whereas the markers correspond to queueing network simulation results.

determined by the utilizations but also by the transmission capacities of service stations. In Scenario 3, for instance, when the transmission capacities are the same ($\kappa = 1$), a larger fraction of a file (70%) is sent to the first queue with a lower utilization, whereas when the transmission capacity of the second service station is larger ($\kappa = 2$), a larger fraction of a file (58%) is sent to the second queue even though its service station has a higher utilization. Further, one can see that splitting a file may not always reduce the mean sojourn time for scenarios with a large asymmetry in background utilizations; this is illustrated in Scenario 6 where the transmission capacity and the low background utilization of the second service station offer an overall advantage over those of the first one and the entire traffic can be directed to the second service station. However, looking at the results for Scenario 5, one can argue that splitting a file does result in lower download times if one fragment is directed to a service station with a much higher background utilization but with a sufficiently compensating transmission capacity.

Overall, the decision to split a file and determination of the optimum splitting ratio can be based on the utilization information available from the BSs and achievable spectral efficiencies. Similarly, deploying a DC architecture for file downloads should be guided by the long-term congestion status of BSs and achievable spectral efficiencies.

B. VALIDATION OF THE EXTENDED MODEL VIA FLOW-LEVEL WIRELESS SIMULATIONS

We first present the results for the default case where the transmit power of the small cell BS is 20 dBm and there

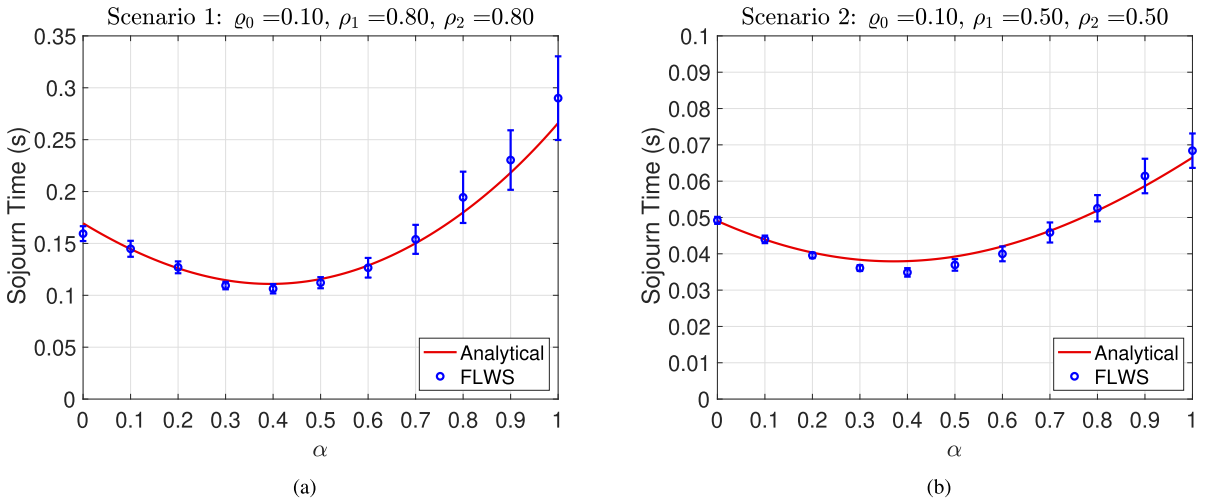

FIGURE 7. Mean sojourn times of file downloads with DC as a function of splitting ratio α for the default case. Equal background utilizations are considered. (a) Scenario 1: $\varrho_0 = 0.10$, $\rho_1 = 0.80$, $\rho_2 = 0.80$ and (b) Scenario 2: $\varrho_0 = 0.10$, $\rho_1 = 0.50$, $\rho_2 = 0.50$. The lines correspond to analytical results, whereas the markers with error bars correspond to flow-level wireless simulator (FLWS) results. For the wireless setting considered, the transmission capacity factor is $\kappa = 1.285$.

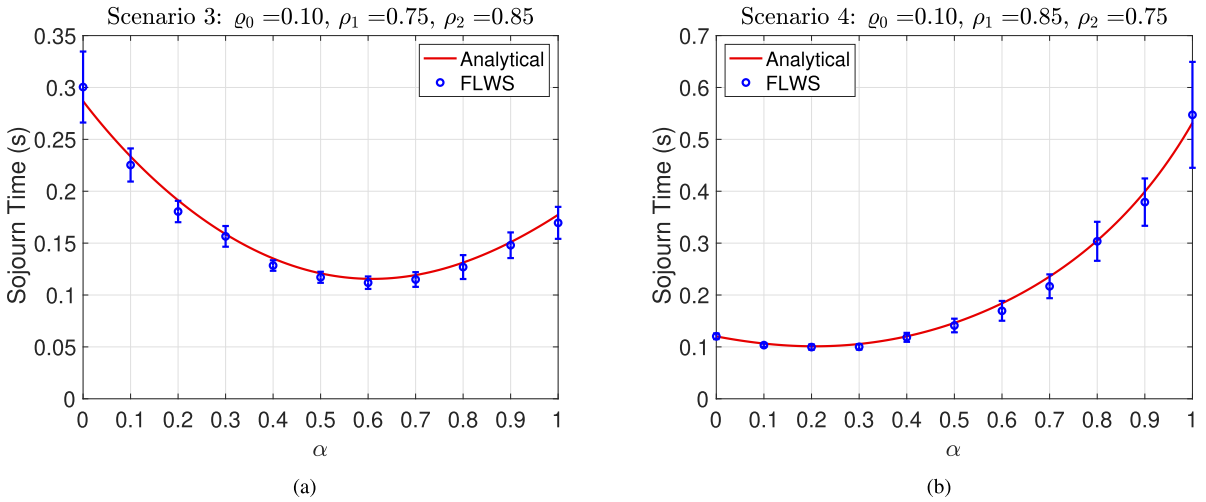

FIGURE 8. Mean sojourn times of file downloads with DC as a function of splitting ratio α for the default case. A slight asymmetry in background utilizations is considered. (a) Scenario 3: $\varrho_0 = 0.10$, $\rho_1 = 0.75$, $\rho_2 = 0.85$ and (b) Scenario 4: $\varrho_0 = 0.10$, $\rho_1 = 0.85$, $\rho_2 = 0.75$. The lines correspond to analytical results, whereas the markers with error bars correspond to flow-level wireless simulator (FLWS) results. For the wireless setting considered, the transmission capacity factor is $\kappa = 1.285$.

are no interferers. In this case, the transmission capacities calculated for the generated path loss and shadowing maps are $c = 150$ Mbps and $\kappa c = 193$ Mbps for the macrocell and the small cell BSs, respectively ($\kappa = 1.285$). The mean sojourn times with DC versus the splitting ratio α obtained via FLWS are plotted in Figs. 7 to 9 together with the results obtained using the analytical approximation in (9) for the same load conditions given in Table 2. In the figures, the lines correspond to analytical results, whereas the markers with error bars correspond to FLWS results. The FLWS results reported for each scenario are averages over 20 independent runs, with the error bars representing 95% confidence intervals based on these 20 runs. The location of the small cell within the macrocell is randomized in each of the runs in order to obtain results that are as general as possible and hence not dependent upon the specific location of the small cell. Each run sweeps the splitting ratio range from $\alpha = 0$ to $\alpha = 1$ with steps of 0.1. The duration of a simulation for a particular α in a given run is 500,000 TTIs. It can be concluded that the analytical approximation accurately captures the delay performance of DC users in simulations that incorporate wireless network characteristics. The only scenario in which there is a slight discrepancy is the case of the lightly loaded network (Scenario 2) where some analytical results are not within the confidence intervals shown (Fig. 7(b)); nevertheless, the relative differences between the FLWS and analytical sojourn times are still less than $\sim 10\%$. The benefit

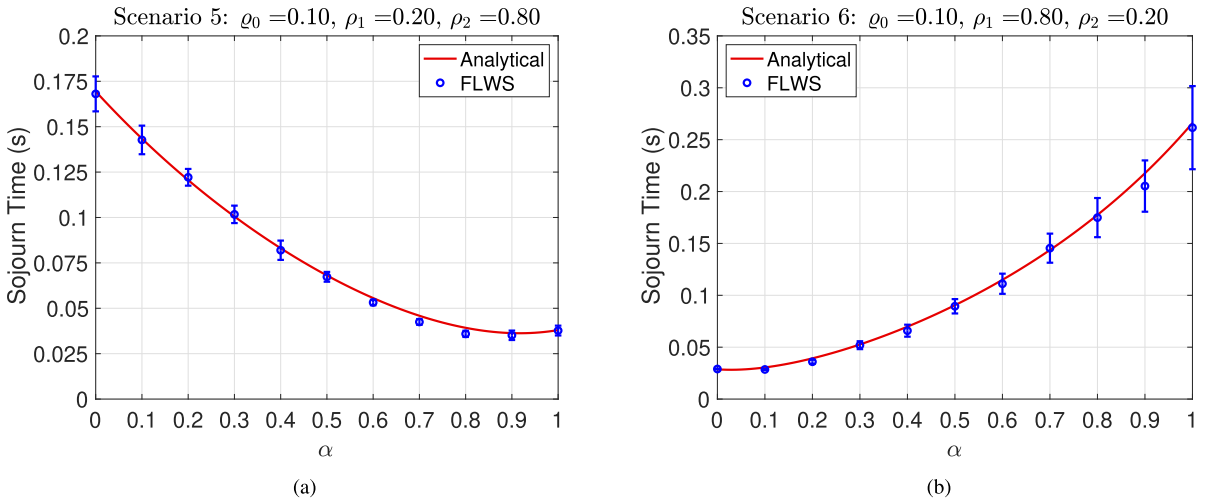

FIGURE 9. Mean sojourn times of file downloads with DC as a function of splitting ratio α for the default case. A large asymmetry in background utilizations is considered. (a) Scenario 5: $\varrho_0 = 0.10$, $\rho_1 = 0.20$, $\rho_2 = 0.80$ and (b) Scenario 6: $\varrho_0 = 0.10$, $\rho_1 = 0.80$, $\rho_2 = 0.20$. The lines correspond to analytical results, whereas the markers with error bars correspond to flow-level wireless simulator (FLWS) results. For the wireless setting considered, the transmission capacity factor is $\kappa = 1.285$.

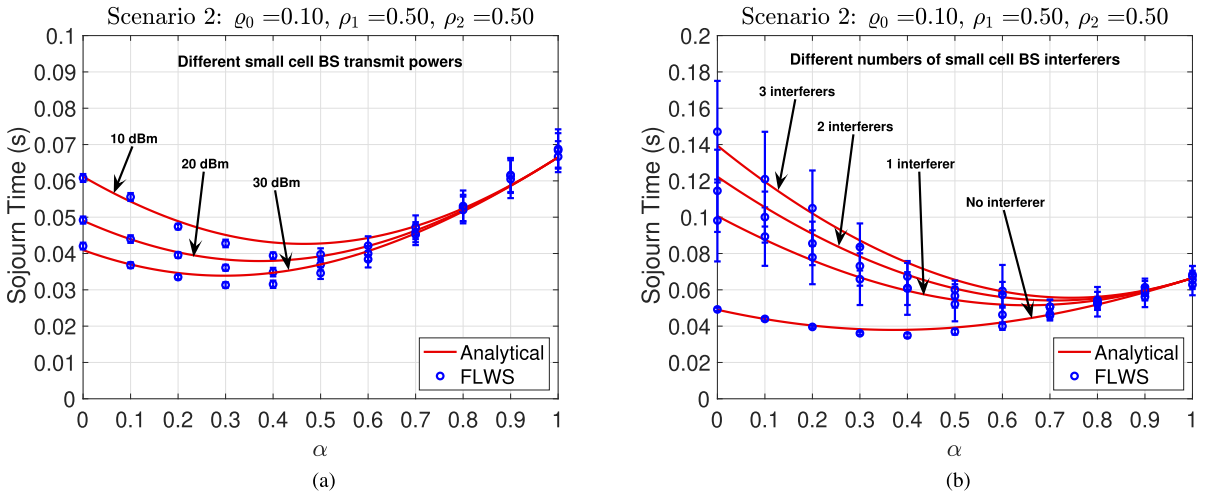

FIGURE 10. Mean sojourn times of file downloads with DC as a function of splitting ratio α for the extended cases. The load conditions are as in Scenario 2: $\varrho_0 = 0.10$, $\rho_1 = 0.50$, $\rho_2 = 0.50$. The results are shown for (a) different small cell BS transmit powers and (b) different numbers of small cell BS interferers. The lines correspond to analytical results, whereas the markers with error bars correspond to flow-level wireless simulator (FLWS) results. For the wireless settings considered in (a), the transmission capacity factors are $\kappa = 1.070$ (10 dBm), $\kappa = 1.285$ (20 dBm, default), and $\kappa = 1.500$ (30 dBm), whereas for the wireless settings considered in (b), the capacity factors are $\kappa = 1.285$ (no interferer, default), $\kappa = 0.729$ (1 interferer), $\kappa = 0.635$ (2 interferers), and $\kappa = 0.581$ (3 interferers).

of file splitting by DC users is particularly pronounced in heavily loaded symmetric and slightly asymmetric systems (Figs. 7(a), 8(a), and 8(b)). For instance, in Scenario 3, the delay is reduced by $\sim$ 60% if DC with optimum α is employed as opposed to using a single connection only to the small cell BS (Fig. 8(a)). One can again observe that splitting a file may not always reduce the mean sojourn time significantly for scenarios with a large asymmetry in background utilizations; in such cases, diverting the entire file to the macrocell BS (Fig. 9(a)) or maintaining a single connection to the small cell BS (Fig. 9(b)) will be more beneficial. As a side note, we point out that if specific small cell configurations are under investigation, then the location and size of the small cell are expected to affect the delay performance and hence the choice of the optimum splitting ratio. In this case, the DC traffic offloaded to the macrocell BS will be targeted toward users that are spatially clustered relative to the macrocell BS and hence have similar achievable bit rates.

The results obtained for the extended cases under the load conditions specified for Scenario 2 are shown in Fig. 10. As before, all the FLWS results reported are averages over 20 independent runs, each with a randomized location of the small cell within the macrocell. We again observe that there is an extremely good agreement between the analytical and FLWS results. The analytical mean sojourn times are

mostly within the confidence intervals shown; if not, they differ from the FLWS values by less than $\sim$ 10%. For the first extended case where we assess the effects of changing the transmit power of the small cell BS (Fig. 10(a)), the transmission capacity factors are $\kappa = 1.070$ (10 dBm), $\kappa = 1.285$ (20 dBm, default), and $\kappa = 1.500$ (30 dBm). If the transmit power of the small cell BS is increased from 20 dBm (default) to 30 dBm, the minimum mean sojourn time decreases and this minimum is achieved at a lower α; i.e. a smaller fragment of a file is transmitted by the macrocell BS due to a resulting increase in the capacity of the small cell BS. On the other hand, if the transmit power of the small cell BS is reduced to 10 dBm, the minimum mean sojourn time increases and this minimum is now achieved at a higher α; i.e. a larger fragment of a file is transmitted by the macrocell BS due to a resulting decrease in the capacity of the small cell BS. The capacity factors calculated for the second extended case where we assess the effects of having interfering sources (Fig. 10(b)) are $\kappa = 1.285$ (no interferer, default), $\kappa = 0.729$ (1 interferer), $\kappa = 0.635$ (2 interferers), and $\kappa = 0.581$ (3 interferers). We find that the minimum mean sojourn time increases as the number of interferers increases from zero to three; the optimum splitting ratios α^* for achieving these minimum times become larger due to a resulting decrease in the capacity of the small cell BS exposed to a greater level of interference. Close agreement between the analytical and FLWS results in all cases shown in Fig. 10 can be considered as further evidence for the applicability of our flow-level model to analyze DC performance in realistic wireless settings.

IV. CONCLUSION

In summary, we have developed an analytical framework that can be used to determine optimum file splitting ratios in DC settings by duly accounting for different transmission capacities and utilizations of BSs. All our analytical results have been validated via both queueing network and flow-level wireless simulations. Close agreement between the analytical and simulation results in all cases points to the applicability of our flow-level model to analyze DC performance. The quantitative results demonstrate that DC may indeed improve the download performance for elastic file transfers. Most importantly, our analysis provides a clear indication that factoring in both different transmission capacities and utilizations of service stations in the queueing model is a must for accurate determination of optimum splitting ratios.

We expect that the findings reported here will prove useful in implementation and operation of DC architectures in recent generations of cellular networks. Future research may involve a trade-off analysis of performance gain versus reassembly complexity at a DC user equipment as well as an evaluation of the time scale at which the determined optimum splitting ratio remains valid in highly dynamic settings. Extension of the analysis to statistically non-identical users with different flow size and rate characteristics or to spatial clusters of users will naturally be the next step in our work. Incorporation of non-ideal backhaul link latencies, consideration of complexity of distributed radio resource management, flow control in DC models, and validation against real traffic measurements are also challenges that will need to be addressed in the future.

REFERENCES

[1] C. Rosa, K. Pedersen, H. Wang, P.-H. Michaelsen, S. Barbera, E. Malkamäki, T. Henttonen, and B. Sébire, "Dual connectivity for LTE small cell evolution: Functionality and performance aspects," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 137–143, Jun. 2016.

[2] H. Wang, C. Rosa, and K. I. Pedersen, "Dual connectivity for LTE-advanced heterogeneous networks," *Wireless Netw.*, vol. 22, no. 4, pp. 1315–1328, May 2016.

[3] D. Laselva, D. Lopez-Perez, M. Rinne, and T. Henttonen, "3GPP LTE-WLAN aggregation technologies: Functionalities and performance comparison," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 195–203, Mar. 2018.

[4] M.-T. Suer, C. Thein, H. Tchouankem, and L. Wolf, "Multi-connectivity as an enabler for reliable low latency communications—An overview," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 156–169, 1st Quart., 2020.

[5] J. F. Monserrat, F. Bouchmal, D. Martin-Sacristan, and O. Carrasco, "Multi-radio dual connectivity for 5G small cells interworking," *IEEE Commun. Standards Mag.*, vol. 4, no. 3, pp. 30–36, Sep. 2020.

[6] M. Agiwal, H. Kwon, S. Park, and H. Jin, "A survey on 4G–5G dual connectivity: Road to 5G implementation," *IEEE Access*, vol. 9, pp. 16193–16210, 2021.

[7] C. Pupiales, D. Laselva, Q. De Coninck, A. Jain, and I. Demirkol, "Multi-connectivity in mobile networks: Challenges and benefits," *IEEE Commun. Mag.*, vol. 59, no. 11, pp. 116–122, Nov. 2021.

[8] Qualcomm. (Mar. 2022). *AIS, Qualcomm and ZTE Announce the World's First 5G NR-DC Showcase for 2.6 GHz and 26 GHz in Thailand*. Press Note. [Online]. Available: https://www.qualcomm.com/news/releases/2022/03/ais-qualcomm-and-zte-announce-worlds-first-5g-nr-dc-showcase-26ghz-and

[9] Ericsson. (Jan. 2023). *Deutsche Telekom, Ericsson and Qualcomm Demonstrate Millimeter Wave Technologies for QoS Managed Connectivity*. Press Release. [Online]. Available: https://www.ericsson.com/en/press-releases/3/2023/deutsche-telekom-ericsson-and-qualcomm-demonstrate-millimeter-wave-technologies-for-qos-managed-connectivity

[10] A. L. Ramaboli, O. E. Falowo, and A. H. Chan, "Bandwidth aggregation in heterogeneous wireless networks: A survey of current approaches and issues," *J. Netw. Comput. Appl.*, vol. 35, no. 6, pp. 1674–1690, Nov. 2012.

[11] G. Pocovi, S. Barcos, H. Wang, K. I. Pedersen, and C. Rosa, "Analysis of heterogeneous networks with dual connectivity in a realistic urban deployment," in *Proc. IEEE 81st Veh. Technol. Conf. (VTC Spring)*, May 2015, pp. 1–5.

[12] J. C. S. Arenas, T. Dudda, J. Schmitz, and R. Mathar, "End-user benefits of LTE dual connectivity in heterogeneous networks," *Mobilkommunikation Technologien Und Anwendungen*, vol. 263, pp. 44–48, May 2016.

[13] S. Chandrashekar, A. Maeder, C. Sartori, T. Höhne, B. Vejlgaard, and D. Chandramouli, "5G multi-RAT multi-connectivity architecture," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, May 2016, pp. 180–186.

[14] M.-S. Pan, T.-M. Lin, C.-Y. Chiu, and C.-Y. Wang, "Downlink traffic scheduling for LTE–A small cell networks with dual connectivity enhancement," *IEEE Commun. Lett.*, vol. 20, no. 4, pp. 796–799, Apr. 2016.

[15] M. Polese, M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Improved handover through dual connectivity in 5G mmWave mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2069–2084, Sep. 2017.

[16] O. N. C. Yilmaz, O. Teyeb, and A. Orsino, "Overview of LTE-NR dual connectivity," *IEEE Commun. Mag.*, vol. 57, no. 6, pp. 138–144, Jun. 2019.

[17] S. K. Ghosh and S. C. Ghosh, "Performance analysis of dual connectivity in control/user-plane split heterogeneous networks," *Comput. Commun.*, vol. 149, pp. 370–381, Jan. 2020.

[18] S. Kar, P. Mishra, and K.-C. Wang, "Dynamic packet duplication for reliable low latency communication under mobility in 5G NR-DC networks," *Comput. Netw.*, vol. 234, Oct. 2023, Art. no. 109923.

[19] S. C. Jha, K. Sivanesan, R. Vannithamby, and A. T. Koc, "Dual connectivity in LTE small cell networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2014, pp. 1205–1210.

[20] H. Yu, C. Hua, J. Li, and R. Ni, "Delay optimal concurrent transmissions in multi-radio access networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 3598–3603.

[21] V. Ramaswamy, J. T. Correia, and D. Swain-Walsh, "Modeling and analysis of multi-RAT dual connectivity operations in 5G networks," in *Proc. IEEE 2nd 5G World Forum (5GWF)*, Sep. 2019, pp. 484–489.

[22] S. B. Fredj, T. Bonald, A. Proutiere, G. Regnie, and J. W. Roberts, "Statistical bandwidth sharing: A study of congestion at flow level," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 31, no. 4, pp. 111–122, Oct. 2001. [Online]. Available: https://dl.acm.org/toc/sigcomm-ccr/2001/31/4

[23] T. Bonald and J. W. Roberts, "Congestion at flow level and the impact of user behaviour," *Comput. Netw.*, vol. 42, no. 4, pp. 521–536, Jul. 2003.

[24] A. Fehske, H. Klessig, J. Voigt, and G. Fettweis, "Flow-level models for capacity planning and management in interference-coupled wireless data networks," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 164–171, Feb. 2014.

[25] K. Avrachenkov, U. Ayesta, P. Brown, and R. Nunez-Queija, "Discriminatory processor sharing revisited," in *Proc. IEEE 24th Annu. Joint Conf. IEEE Comput. Commun. Societies.*, Mar. 2005, pp. 784–795.

[26] T. Bonald and J. Roberts, "Scheduling network traffic," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 34, no. 4, pp. 29–35, Mar. 2007.

[27] T. Bonald, S. Borst, N. Hegde, M. Jonckheere, and A. Proutiere, "Flow-level performance and capacity of wireless networks with user mobility," *Queueing Syst.*, vol. 63, nos. 1–4, pp. 131–164, Dec. 2009.

[28] G. J. Hoekstra, R. D. van der Mei, and S. Bhulai, "Optimal job splitting in parallel processor sharing queues," *Stochastic Models*, vol. 28, no. 1, pp. 144–166, Feb. 2012. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/15326349.2012.646555

[29] G. Dandachi, S. E. Elayoubi, T. Chahed, and N. Chendeb, "Network-centric versus user-centric multihoming strategies in LTE/WiFi networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4188–4199, May 2017.

[30] G. J. Hoekstra, R. D. van der Mei, and J. W. Bosman, "Traffic splitting policies in parallel queues with concurrent access: A comparison," in *Proc. 26th Int. Teletraffic Congr. (ITC)*, Sep. 2014, pp. 1–9.

[31] G. J. Hoekstra and R. D. van der Mei, "Effective load for flow-level performance modelling of file transfers in wireless LANs," *Comput. Commun.*, vol. 33, no. 16, pp. 1972–1981, Oct. 2010.

[32] S. Bhulai, G. J. Hoekstra, J. W. Bosman, and R. D. van der Mei, "Dynamic traffic splitting to parallel wireless networks with partial information: A Bayesian approach," *Perform. Eval.*, vol. 69, no. 1, pp. 41–52, Jan. 2012.

[33] J. W. Bosman, G. J. Hoekstra, R. D. van der Mei, and S. Bhulai, "A simple index rule for efficient traffic splitting over parallel wireless networks with partial information," *Perform. Eval.*, vol. 70, no. 10, pp. 889–899, Oct. 2013.

[34] J. Sun, S. Zhang, S. Xu, and S. Cao, "High throughput and low complexity traffic splitting mechanism for 5G non-stand alone dual connectivity transmission," *IEEE Access*, vol. 9, pp. 65162–65172, 2021.

[35] K. Lien, K.-H. Lin, and H.-Y. Wei, "Energy-efficient traffic steering in millimeter-wave dual connectivity discontinuous reception framework," *IEEE Access*, vol. 10, pp. 115716–115731, 2022.

[36] J. Kim and S. Bahk, "Blockage-aware flow control in E-UTRA-NR dual connectivity for QoS enhancement," *IEEE Access*, vol. 10, pp. 68834–68845, 2022.

[37] C. Pupiales, D. Laselva, and I. Demirkol, "Capacity and congestion aware flow control mechanism for efficient traffic aggregation in multi-radio dual connectivity," *IEEE Access*, vol. 9, pp. 114929–114944, 2021.

[38] C. Pupiales, D. Laselva, and I. Demirkol, "Fast data recovery for improved mobility support in multiradio dual connectivity," *IEEE Access*, vol. 10, pp. 93674–93691, 2022.

[39] D. Hasselquist, C. Lindström, N. Korzhitskii, N. Carlsson, and A. Gurtov, "QUIC throughput and fairness over dual connectivity," *Comput. Netw.*, vol. 219, Dec. 2022, Art. no. 109431.

[40] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation With Computer Science Applications*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[41] H. J. Kushner, *Heavy Traffic Analysis of Controlled Queueing and Communication Networks*. Cham, Switzerland: Springer, 2001.

[42] A. Zwart and O. Boxma, "Sojourn time asymptotics in the M/G/1 processor sharing queue," *Queueing Syst.*, vol. 35, nos. 1–4, pp. 141–166, 2000.

[43] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[44] M. Bertoli, G. Casale, and G. Serazzi, "JMT: Performance engineering tools for system modeling," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 36, no. 4, pp. 10–15, Mar. 2009.

[45] T. Bonald and A. Proutière, "Wireless downlink data channels: User performance and cell dimensioning," in *Proc. 9th Annu. Int. Conf. Mobile Comput. Netw.*, Sep. 2003, pp. 339–352.

[46] P. Mogensen, W. Na, I. Z. Kovacs, F. Frederiksen, A. Pokhariyal, K. I. Pedersen, T. Kolding, K. Hugl, and M. Kuusela, "LTE capacity compared to the Shannon bound," in *Proc. IEEE 65th Veh. Technol. Conf. VTC-Spring*, Apr. 2007, pp. 1234–1238.

[47] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 1, pp. 177–190, Feb. 2012.

[48] B. Blaszczyszyn and M. K. Karray, "Performance analysis of cellular networks with opportunistic scheduling using queueing theory and stochastic geometry," *IEEE Trans. Wireless Commun.*, vol. 18, no. 12, pp. 5952–5966, Dec. 2019.

[49] J. Baumgarten and T. Kuerner, "LTE downlink link-level abstraction for system-level simulations," in *Proc. Eur. Wireless 20th Eur. Wireless Conf.*, May 2014, pp. 1–5.

[50] *Evolved Universal Terrestrial Radio Access (E-UTRA); Further Advancements for E-UTRA Physical Layer Aspects (Release 9)*, Standard (TS) 36.814, Version 9.2.0, Technical Specification, 3GPP, 3rd Generation Partnership Project (3GPP), Mar. 2017. [Online]. Available: https://www.3gpp.org/dynareport/36814.htm

[51] E. Pateromichelakis, M. Shariat, A. Ul Quddus, and R. Tafazolli, "On the analysis of co-tier interference in femtocells," in *Proc. IEEE 22nd Int. Symp. Pers., Indoor Mobile Radio Commun.*, Sep. 2011, pp. 122–126.

[52] M. Taranetz and M. K. Müller, "A survey on modeling interference and blockage in urban heterogeneous cellular networks," *EURASIP J. Wireless Commun. Netw.*, vol. 2016, no. 1, pp. 1499–1687, Dec. 2016.

[53] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "On accurate simulations of LTE femtocells using an open source simulator," *EURASIP J. Wireless Commun. Netw.*, vol. 2012, no. 1, p. 328, Dec. 2012.

JOHN O. OLAIFA (Graduate Student Member, IEEE) received the B.Sc. degree from Olabisi Onabanjo University, Nigeria, in 2010, and the M.Sc. degree from Eastern Mediterranean University, Northern Cyprus, in 2016, where he is currently pursuing the Ph.D. degree with the Department of Computer Engineering. His research interests include performance measurement of wireless systems, heterogeneous networks, and system simulation.

DOGU ARIFLER (Senior Member, IEEE) received the B.S.E.E., M.S., and Ph.D. degrees in electrical and computer engineering from The University of Texas at Austin, USA, in 1997, 1999, and 2004, respectively. He is currently a Professor with the Department of Computer Engineering, Eastern Mediterranean University, Northern Cyprus. His research interests include network performance analysis, random networks, and nanoscale communication networks. He was a recipient of the Cyprus Fulbright Commission's Full Scholarship Award, from 1993 to 1997.