User Association in HetNets: Impact of Traffic Differentiation and Backhaul Limitations

Nikolaos Sapountzis[1], Thrasyvoulos Spyropoulos[1], Navid Nikaein[1], and Umer Salim[2]

[1]Mobile Communications Department, EURECOM, 06410, Biot, France, firstname.lastname@eurecom.fr

[2]Intel Mobile Communications, Sophia Antipolis, 06560, France, umer.salim@intel.com

***Abstract*—Operators, struggling to continuously add capacity and upgrade their architecture to keep up with data traffic increase, are turning their attention to denser deployments that improve spectral efficiency. Denser deployments make the problem of user association challenging, and much work has been devoted to finding algorithms that strike a tradeoff between user quality of service (QoS), and network-wide performance (load-balancing). Nevertheless, the majority of these algorithms typically consider simple setups with a single type of traffic, usually elastic non-GBR (Guaranteed Bit Rate). They also focus on the radio access part, ignoring the backhaul topology and potential capacity limitations. Backhaul constraints are emerging as a key performance bottleneck in future networks, partly due to the continuous improvement of the radio interface, and partly due to the need for inexpensive backhaul links to reduce capital and operational expenditures. To this end, we propose an analytical framework for user association that jointly considers radio access and backhaul network performance. Specifically, we derive an algorithm that takes into account spectral efficiency, base station load, backhaul link capacities and topology, and two traffic classes (GBR and non-GBR) in both the uplink and downlink directions. We prove analytically an optimal user association rule that ends up maximizing either an arithmetic or a weighted harmonic mean of the achieved performance along different dimensions (e.g. UL and DL performance or GBR and non-GBR performance). We then use extensive simulations to study the impact of (i) traffic differentiation, and (ii) backhaul capacity limitations and topology on key performance metrics.**

***Index Terms*—hetnets; backhaul; optimization; traffic differentiation; user-association; load balancing; spectral efficiency.**

I. Introduction

DRIVEN by the exponential growth in wireless data traffic, operators are increasingly considering denser, heterogeneous network (HetNet) deployments. In a HetNet, a large number of small cells (SC) are deployed along with macrocells to improve spatial reuse [1], [2], [3]. The higher the deployment density, the better the chance that a user equipment (UE) can be associated with a nearby base station (BS) with high signal strength, and the more the options to balance the load. At the same time, denser deployments experience high spatio-temporal load variations, and require sophisticated user association algorithms. There are two key, often conflicting concerns when assigning UEs to a BS: (i) maximizing the spectral efficiency, and (ii) ensuring that the load across BSs is balanced to improve the utilization efficiency, and preempt congestion events. The former is usually achieved by associating the UE to the BS with maximum SINR: this association rule was the base up to LTE (Long-Term Evolution)-release 8. While this rule also maximizes the *instantaneous* rate of a user (i.e., the modulation and coding scheme - MCS - supported), it reflects user QoS only when the BS is lightly loaded. However, user performance, in terms of *per flow delay*, may be severely affected if the BS offering the best SINR is congested [4], [5].

As a result, a number of research works have studied the problem of user association in heterogeneous networks, optimizing user rates [6], [7], balancing BS loads [8], or pursuing a weighted tradeoff of them [9]. For instance, a distributed user-association algorithm is proposed in [10], where the global outage probability and the long term rate maximization are well studied, in the context of load balancing. The authors in [11] propose a framework that studies the interplay of user association and resource allocation in future HetNets, by formulating a non-convex optimization problem and deriving performance upper bounds. Range-expansion techniques, where the SINR of lightly loaded BSs is biased to make them more attractive to the users are also popular [2], [3]. Finally, a framework that has received much attention is [9]. This framework jointly considers a family of objective functions, each of which directs the optimal solution towards different goals (e.g. throughput optimal, delay-optimal, load balancing, etc.), using an iterative algorithm. [12], [13], [14] extend this framework to further include energy management, e.g., by switching off under-loaded BSs.

Nevertheless, the majority of these works are relatively simplified, not taking into account key features of future networks. Firstly, most existing studies only consider homogeneous traffic profiles. For example, [9], [12] assume that all flows generated by a UE are "best-effort" (i.e. elastic). However, modern and future networks will have to deal with high traffic differentiation, with certain flows being able to require specific, *dedicated*[1] (i.e., non-elastic) resources [15]. Such dedicated flows do not share BS resources like best-effort ones, are subject to admission control, and sensitive to different performance metrics [16]. Secondly, the majority of

The research leading to these results has received funding from the European Research Council under the European Community Seventh Framework Programme (FP7/2012- 2015) under the ICT theme of DG-CONNECT n^{o} 317941 (iJOIN).

[1]In terms of LTE systems, dedicated flows are differentiated by their QoS class (QCI) with different rate, priority, and latency requirements, whereas best-effort flows experience the same treatment as they belong to the same data radio bearer.

related studies only consider downlink (DL) traffic. Uplink (UL) traffic is becoming important, due to symmetric (e.g. social networking) applications, Machine-Type Communication (MTC), etc. Yet, due to the asymmetric transmit powers of UEs and BSs, leading to different physical data rates, the BS which is optimal for DL traffic might lead to severely degraded performance for UL traffic. Summarizing, a proper user-association scheme should consider all the above dimensions, and attempt to strike an appropriate tradeoff between them.

On top of that, most related works focus on the radio access part (e.g., considering the user rate on the radio interface or BS load), ignoring the backhaul (BH) network. While this might be reasonable for legacy cellular networks, given that the macrocell backhaul is often over-provisioned (e.g., fiber), this might be quite suboptimal for future cellular networks. The considerably higher number of small cells, and related Capital Expenditure (CAPEX) and Operational Expenditure (OPEX) suggest that backhaul links will mostly be inexpensive wired or wireless (in licensed or unlicensed bands), and underprovisioned [17]. Multiple BS might also have to share the capacity of a single backhaul link due to, e.g, point-to-multipoint (PMP) or multi-hop mesh topologies to the aggregation node(s) [18]. Finally, various BS-coordinated schemes have been proposed in the literature as a promising way to better use the available spectrum and further improve system performance, e.g., enhanced Inter-Cell Interference Coordination (eICIC) [19], [20] and Coordinated Multi-Point (CoMP) transmission [21] scenarios. Such schemes are expected to further stress the backhaul network capacities. Hence, as the radio access technologies are constantly improving, it is argued that the backhaul network will emerge as a major performance bottleneck, and user association algorithms that ignore the backhaul load and topology can lead to poor performance [22].

As a result of this increasing focus on the backhaul, some recent works have appeared that attempt to jointly consider radio access and backhaul. These are mostly concerned with joint scheduling issues (for in-band or PMP backhaul links) [22], [23], signaling overhead and performance tradeoffs for cooperative multi-point communication [24], Software-Defined-Networking (SDN)-based implementation flexibility [25]. The user rate maximization problem is studied in [26] under backhaul capacity constraints, and in [27] jointly with backhaul resource allocation and flow control. Also, a distributed user association scheme was developed for maximizing the network-wide spectrum efficiency in [28] using combinatorial optimization. Some backhaul aware association heuristics include: [29] where the sum of user rates is attempted to be optimized in an energy-efficient manner, [30] where the ergodic capacity is maximized under an iterative algorithm, and [31] where resource allocation is investigated in conjunction with carrier aggregation. Other works in this context include investigation of caching capabilities to overcome the backhaul capacity limitations and enhance QoS [32], [33]. Finally, Chen et al. attempt to derive the total expected delay by considering retransmission over the wireless backhaul links [34].

Nevertheless, to our best knowledge, none of these works formally addresses the problem of optimal user association in future and potentially backhaul-limited HetNets. To this end, we revisit the user association problem, jointly considering the radio access and backhaul networks. Specifically, our main contributions can be summarized as follows

1) We use the popular framework of α-optimal user association [9] as our starting point, and considerably extend it to include (i) traffic differentiation, (ii) UL traffic, and (iii) backhaul topology and capacity constraints.

2) We then analytically prove different association rules, depending on whether UL and DL traffic of the same UE can be "split" to different BSs or not [35]. Interestingly, depending on this UL/DL "split" the derived rules end up maximizing either an arithmetic or a weighted harmonic mean of optimal association rules per problem dimension.

3) We use our framework to investigate the various tradeoffs arising in this complex association problem, and provide some initial insights and guidelines about the impact of traffic differentiation and backhaul limitations in optimal user-association policies for future HetNets.

4) Our results also highlight some shortcomings of future HetNets, and indicate potential extensions to tackle them within our framework. These include the need for joint radio access and Layer 3 routing on the transport (backhaul) network, and dynamic allocation of access as well as backhaul resources (e.g., in the context of dynamic TDD).

The remainder of the paper is organized as follows: Section II describes the system model and related assumptions. In Sections III and IV we derive the optimal user-association policies for provisioned and under-provisioned backhaul network. In Section V we simulate our proposed optimal association rules and attempt to shed some light on the impact of traffic differentiation, backhaul constraints and topology on system performance. Section VI concludes the paper.

II. System Model and Assumptions

In the following, we describe our traffic arrival model (Section II-A), the discuss our assumptions related to the access (Section II-B) and backhaul networks (Section II-C).

We use a similar problem setup as the one used in a number of related works [9], [12], [36], [13], and extend it accordingly. To keep notation consistent, for all variables considered a first superscript "D" and "U" refers to downlink (DL) and uplink (UL) traffic, respectively. A second superscript "b" or "d" refers to best-effort and dedicated traffic, respectively. For brevity, in the following *we present most notation and assumptions in terms of downlink traffic only, assuming that the uplink case and notation is symmetric*. Specific differences will be elaborated, where necessary. In Table I, we summarize some useful notation we use throughout the paper.

A. Traffic Model

(A.1 - Traffic arrival rates) Traffic at location $x \in \mathcal{L}$ consists of file (or more generally *flow*) requests arriving according to an inhomogeneous Poisson point process with arrival rate per unit area $\lambda(x)$[2]. This inhomogeneity facilitates

[2]Without loss of generality, we do not distinguish between users at location x, as we assume that all users/flows related to location x are treated similarly.

TABLE I
NOTATION

Variable	Best-Effort Flows		Dedicated Flows	
	Downlink	Uplink	Downlink	Uplink
Flow type superscript	D,b	U,b	D,d	U,d
Flow type probability	$z^D \cdot z^b$	$z^U \cdot z^b$	$z^D \cdot z^d$	$z^U \cdot z^d$
Devoted amount of bandwidth for BS i	$w_i^D \cdot \zeta^D$	$w_i^U \cdot \zeta^U$	$w_i^D \cdot (1-\zeta^D)$	$w_i^U \cdot (1-\zeta^U)$
Traffic arrival rate (flows/sec) at x	$\lambda^{D,b}(x)$	$\lambda^{U,b}(x)$	$\lambda^{D,d}(x)$	$\lambda^{U,d}(x)$
Maximum rate | servers at x associated with i-th BS	$c_i^{D,b}(x)$	$c_i^{U,b}(x)$	$k_i^D(x)$	$k_i^U(x)$
System load (utilization density) at x associated with i-th BS	$\rho^{D,b}(x)$	$\rho^{U,b}(x)$	$\rho^{D,d}(x)$	$\rho^{U,d}(x)$
Load-balancing degree parameter $\in [0,\infty)$	$\alpha^{D,b}$	$\alpha^{U,b}$	$\alpha^{D,d}$	$\alpha^{U,d}$
Total utilization (load) of the i-th BS	$\rho^{D,b}$	$\rho^{U,b}$	$\rho^{D,d}$	$\rho^{U,d}$
Probability that a specific flow arriving at x is routed to BS i	$p^{D,b}(x)$	$p^{U,b}(x)$	$p^{D,d}(x)$	$p^{U,d}(x)$
Flow size (in bits per flow) | flow demand (in bps), duration (in sec) at x	$1/S^{D,b}(x)$	$1/S^{U,b}(x)$	$B^D(x), 1/\mu^{D,d}(x)$	$B^U(x), 1/\mu^{U,d}(x)$
Capacity of BH link j	$C_h^D(j)$	$C_h^U(j)$	-	-
Congestion indicator at BH link j	$\mathcal{I}^D(j)$	$\mathcal{I}^U(j)$	-	-

the creation of "hotspot" areas. Each new arriving request is for a *downlink (DL)* flow, with probability z^D, or *uplink (UL)* flow with probability $z^U = 1 - z^D$. Each DL (or UL) flow can further be a *best-effort* flow (e.g., file download) with probability z^b, or *dedicated* flow (e.g., a VoIP call), with probability $z^d = 1 - z^b$. z^D and z^b are input parameters that depend on the traffic mix.

Using a Poisson splitting argument [37], it follows that the above gives rise to 4 independent, Poisson flow arrival processes with respective rates

$$\lambda^{D,b}(x) = z^D \cdot z^b \cdot \lambda(x), \quad \lambda^{D,d}(x) = z^D \cdot z^d \cdot \lambda(x) \tag{1}$$

$$\lambda^{U,b}(x) = z^U \cdot z^b \cdot \lambda(x), \quad \lambda^{U,d}(x) = z^U \cdot z^d \cdot \lambda(x), \tag{2}$$

($\lambda^{D,b}(x)$ for the downlink best-effort flows, $\lambda^{U,b}(x)$ for the uplink best-effort flows, etc.).

(A.2 - Best effort flow characteristics) Each *best-effort* flow is associated with a *flow-size* (in bits) drawn from a generic distribution with mean $1/S^{D,b}(x)$. This can model heterogeneous flow characteristics across locations.

(A.3 - Dedicated flow characteristics) Each *dedicated* flow has a *required data-rate* (in bits per second) that is drawn from a generic distribution with mean $B^D(x)$. This rate must be guaranteed by the network throughout the flow's duration. This duration (in seconds) is another, independent random variable with mean $1/\mu^{D,d}(x)$.

B. Access Network

(B.1 - Access network topology) We assume an area $\mathcal{L} \subset \mathbb{R}^2$ served by a set of base stations $\mathcal{B}$, that are either macro BSs (eNBs) or small cells (SCs). These together constitute the access network.

(B.2 - DL resources) Each BS $i \in \mathcal{B}$ is associated with a transmit power P_i and a total downlink bandwidth w_i^D. Out of the total bandwidth, $\zeta_i^D \cdot w_i^D$ is allocated to best-effort traffic and $(1-\zeta_i^D)\cdot w_i^D$ for dedicated traffic ($0 \le \zeta_i^D \le 1$). Throughout this paper, we will assume that this allocation is static, at least for a given time window of interest (based on long term traffic characteristics and operator policy). Dynamically updating the ζ_i^D parameters could further improve performance, but is related more to the MAC scheduler of each BS and is out of the scope of this paper.

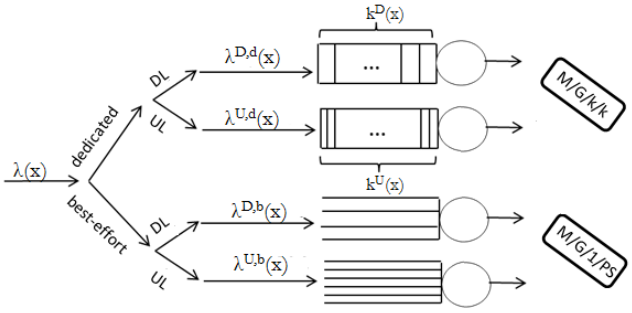


Fig. 1. The four queuing systems at the BS level. Flows, initiated from potentially different user locations x, that belong to the same traffic-class are aggregated to the same queue.

(B.3 - DL physical data rate) BS i can deliver a *maximum* physical data transmission rate of $c_i^{D,b}(x)$ to a user at location x, in absence of any other best-effort flows served, which is given by the Shannon capacity[3]

$$c_i^{D,b}(x) = \zeta_i^D \cdot w_i^D \cdot \log_2(1 + \text{SINR}_i(x)), \tag{3}$$

where $\text{SINR}_i(x) = \frac{G_i(x)P_i}{\sum_{j\neq i} G_j(x)P_j + N_0}$. [4] N_0 is the noise power, and $G_i(x)$ represents the path loss and shadowing effects between the i-th BS and the UE located at x (as well as antenna and coding gains, etc.)[5]. We assume that effects of fast fading are filtered out. Our model assumes that the total intercell interference at location x is static, and considered as another noise source, as is previously considered in most aforementioned works [9], [12].

[3]We use Shannon capacity for clarity of presentation. However, our approach could be easily adapted to include modulation and coding schemes (MCS). Furthermore, capacity improving technologies, e.g., the use of MIMO, and modifications to this capacity formula are othogonal to our framework.

[4]We have assumed that the interference caused to a BS, does not depend on the load of the (neighboring) BSs that interfere it. This is a standard practice in the queue modeling, to avoid coupled queueing systems with no closed form results, e.g. see [4],[9].

[5]In the case of UL, we assume that the Tx power of each user is P^{UE}, and slightly abuse notation for SINR, G, etc., as these don't play a major role in the remaining discussion.

The next 4 points (B.4-B.7) describe the scheduling and performance model for best effort traffic only. We return to dedicated traffic in (B.8-B.9).

(B.4 - Best effort load density) We define the service rate of best effort flows at BS i to be $\mu_i^{D,b}(x) = S^{D,b}(x) \cdot c_i^{D,b}(x)$ per unit area. Now, we can introduce the *load density* for best effort flows, at different locations x

$$\rho_i^{D,b}(x) = \frac{\lambda^{D,b}(x)}{\mu_i^{D,b}(x)} = \frac{\lambda^{D,b}(x)}{S^{D,b}(x) \cdot c_i^{D,b}(x)}, \quad (4)$$

which is the contribution of location x to the total load of a BS i, when location x is associated to BS i.

(B.5 - Best effort load) Each location x is associated with routing probabilities $p_i^{D,b}(x) \in [0,1]$, which are the probabilities that best effort DL flows generated for users at location x get associated with (i.e., are served by) BS i. Then, we can define the *total best effort load* $\rho_i^{D,b}$ for BS i as

$$\rho_i^{D,b} = \int_{\mathcal{L}} p_i^{D,b}(x) \rho_i^{D,b}(x) dx. \quad (5)$$

This is a generalization of a well known queueing result for servers with multiple traffic types (each location x corresponding to a different traffic type) [37], [4]. $\rho_i^{D,b}$ is the expected *utilization* of BS i. We'll use the terms *load* and *utilization* interchangeably for this quantity. It must be lower than 1 for the BS to be stable (or, to keep the delays finite). Note that *it is not the instantaneous load* but rather the average load this BS will experience, and is a function of the association variables $p_i^{D,b}(x)$ (which define an "association map"). Hence, similarly to [4], [9], we are interested in the system *flow-level dynamics*, and model the service of DL best-effort flows at each BS as a queueing system with load $\rho_i^{D,b}$ shown in Fig. 1. Finally, since we are interested in the aggregation of all flows at BS level (i.e. all flows from all locations x associated with BS i), even if flow arrivals at each x is not Poisson (as in A.1), the Palm-Khintchine theorem [37] suggests that Poisson assumption is a good approximation for the BS input traffic.

(B.6 - Best effort scheduling) Proportionally fair scheduling is often implemented in LTE networks for best-effort flows, due to its good fairness and spectral efficiency properties [4], [9], [15]. This can be modeled as an M/G/1 multi-class processor sharing (PS) system emulated by sharing the available capacity within time slots or frequency resource blocks. It is multi-class, because each flow might get different rates for similarly allocated resources, due to different channel quality and MCS at x. Interestingly, opportunistic scheduling can be included into our framework using a multiplicative factor in the average service rate, e.g. see [4]. Such a factor would depend on the SINR distribution, the number of served users, the type of opportunistic scheduling policy etc.

(B.7 - Performance for best effort flows) The stationary number of flows in BS i is equal to $E[N_i] = \frac{\rho_i^{D,b}}{1-\rho_i^{D,b}}$ [37]. Hence, minimizing $\rho_i^{D,b}$ minimizes $E[N_i]$, and by Little's law it also minimizes the per-flow delay for that base station [37]. Also, the throughput for a flow at location x is $c_i^{D,b}(x)(1-\rho_i^{D,b})$. This observation is important to understand how the user's physical data rate $c_i^{D,b}(x)$ (related to users at location x only) and the BS load $\rho_i^{D,b}$ (related to *all* users associated with BS i) affect the optimal association rule.

(B.8 - Dedicated traffic load density) Unlike best-effort flows which are elastic, dedicated flows are subject to admission control, since they require some resources for exclusive usage in order to be accepted in the system. Specifically, let ${c_i}^{D,d}(x)$ denote the maximum offered rate to users at location x corresponding to dedicated flows only (referred to $(1-\zeta_i)$ - see B.3 above). If each flow at x demands, on average, a rate of $B^D(x)$ (see A.3), then at most $k_i^D(x) = \frac{{c_i}^{D,d}(x)}{B^D(x)}$ dedicated flows from x could be served in parallel by BS i (assuming again *no other flows in the system*), and any additional flows would be rejected[6]. Similarly to the best effort case (B.4), we can define a system load density for dedicated traffic at x

$$\rho_i^{D,d}(x) = \frac{\lambda^{D,d}(x)}{\mu^{D,d}(x) k_i^D(x)} = \frac{\lambda^{D,d}(x) \cdot B^D(x)}{\mu^{D,d}(x) \cdot {c_i}^{D,d}(x)}. \quad (6)$$

Hence, a different number of resources $k_i^D(x)$ can be offered to different locations x, depending on the rate demand $B^D(x)$ as well as the channel quality (rate ${c_i}^{D,d}(x)$) at location x.

(B.9 - Dedicated traffic performance) Given the above heterogeneous blocking model for dedicated flows, we can approximate the allocation of BS i dedicated resources with an M/G/k/k (or k-loss) system, where the total load $\rho_i^{D,d}$ can be calculated as in (B.5) and Eq. (5), using the density of Eq. (6) and corresponding routing probability $p_i^{D,d}(x)$ for dedicated flows (see also Fig. 1). It is known that for M/G/k/k systems, minimizing $\rho_i^{D,d}$ is equivalent to minimizing the blocking probability for new flows [37]. This observation is important to understand that a similar tradeoff (as in B.7) exists between choosing a BS at x that maximizes $k_i^D(x)$ (related only to flow and channel characteristics at x) and choosing a BS whose *total* load $\rho_i^{D,d}$ (related to *all* users attached to BS i).

(B.10 - UL/DL association split) We investigate two scenarios, depending on the whether a UE is allowed to be attached to different BSs for its DL and UL traffic [35]:

Split UL/DL: Each UE can be associated to different BSs for its DL and UL traffic. This allows one to optimize UL and DL performance independently [38].

Joint UL/DL: Each UE must be associated with the same BS for both UL and DL traffic (standard practice currently).

C. Backhaul Network

(C.1 - Backhaul network topology) Each access network node (either eNB or SC) is connected to the core network through the eNB aggregation gateway via a certain number of backhaul links that constitute the backhaul network. This connection can be either direct ("star" topology) or through one or more SC aggregation gateways ("tree" topology). Fig. 2 shows such a backhaul routing topology.

Without loss of generality, we assume that there is a fiber link from the eNB to the core network, and focus on the set of capacity-limited backhaul links (wired or wireless) connecting SCs to the eNB, denoted as $\mathcal{B}_h$. We denote as routing path

[6] In fact, since the rate requirement for each flow is a random variable, using its mean $B^D(x)$ in the denominator yields a lower bound for $k_i^D(x)$ (by Jensen's inequality), which can be used as a conservative estimate.

$\mathcal{B}_h(i)$ the set of all backhaul links $j \in \mathcal{B}_h$ along which traffic is routed from BS i to an eNB aggregation point. For example, in Fig. 2, $\mathcal{B}_h(1) = \{1\}$, and $\mathcal{B}_h(3) = \{1,2,3\}$. We further denote as $\mathcal{B}(j)$ the set of all BS $i \in \mathcal{B}$ whose traffic is routed over backhaul link j. E.g., $\mathcal{B}(1) = \{1,2,3,4\}$ and $\mathcal{B}(2) = \{2,3,4\}$ in Fig. 2. In the case of a star topology, there is exactly one (unique) backhaul link used for each BS (i.e., $\|\mathcal{B}_h(i)\| = \|\mathcal{B}(j)\| = 1, \forall i,j$). We assume that the backhaul route for each BS is *given*, e.g., calculated in practice as a Layer 2 (L2) spanning tree, and is an input to our problem. In Section V, we highlight some limitations of L2 routing.

(C.2 - Backhaul capacity requirement) Each backhaul link $j \in \mathcal{B}_h$ is characterized by a DL and UL capacity, denoted as $C_h^D(j)$ and $C_h^U(j)$ bps. These capacities might be the same or different (e.g., Frequency-Division Duplex (FDD), or fixed/dynamic Time-Division Duplex (TDD) systems [39]). Backhaul links usually don't implement any particular scheduling algorithm, and can be seen as a data "pipe".

Without loss of generality, we focus on a scenario with only best-effort traffic. This not only keeps our backhaul model tractable as we shall see later, but also allows us to better understand the impact of backhaul limitations on the wide system performance. Focusing on the DL, the backhaul capacity requirement of a backhaul link $j \in \mathcal{B}_h$ in terms of bits per sec, consists of the sum of downlink loads (corresponding to best-effort traffic) of all BSs using that link

$$\sum_{i \in \mathcal{B}(j)} \rho_i^{D,b} \cdot \tilde{c}_i^D. \tag{7}$$

For example, if a single BS i only uses backhaul link j (e.g. a star topology), and i has a load of $\rho_i^{D,b} = 0.7$, i.e., is active 70% of the time on the downlink, then the average downlink *rate* on backhaul j will be $0.7 \cdot \tilde{c}_i^D$. As for $\tilde{c}_i^D$, this is a parameter tuned by the operator. It could be directly replaced with the average rate considering all possible locations (e.g. as in [4]). However, this is a rather optimistic value to use, and would lead to backhaul link capacities being violated often. Conversely, the use of peak rate (i.e. assuming the maximum MCS used for every active flow) corresponds to the most conservative choice for this parameter. However, it is well known that this is much higher than the average "busy" rate [17], and would lead to backhaul resources being wasted too often. Finally, the direct usage of $p_i(x)$ to derive $\tilde{c}_i$ would not only complicate significantly the problem at hand, but is also somewhat superfluous since in most "busy" scenarios the average rate mostly depends on the edge users [17] and does not change much. We therefore leave this to the operator as a design parameter, to set it depending on how conservative he wants to be and past statistics.Note that Eq. (7) is neither the BS load nor the backhaul link load but simply the *total rate requirement on the backhaul link* (which should not exceed capacity).

(C.3 - Backhaul provisioning) We have derived the backhaul capacity requirement ($\sum_{i \in \mathcal{B}(j)} \rho_i^{D,b} \tilde{c}_i^D$) and defined the backhaul capacity limitation ($C_h^D(j)$) for each link $j \in \mathcal{B}_h$ in DL (see C.2). Thus, each of these links shall introduce a backhaul *constraint* to avoid exceeding its maximum capacity and prohibit backhaul congestion in *either* direction

$$\sum_{i \in \mathcal{B}(j)} \rho_i^D \tilde{c}_i^D < C_h^D(j), \ \sum_{i \in \mathcal{B}(j)} \rho_i^U \tilde{c}_i^U < C_h^U(j), \ \ \forall j \in \mathcal{B}_h. \tag{8}$$

Throughout this paper, we assume that the backhaul network is either *under-provisioned* if the capacity of at least one backhaul link is exceeded, or *provisioned* otherwise. We investigate the user-association problem separately for each scenario in Sections III and IV, by focusing on different tradeoffs.

III. User-Association for Provisioned Backhaul Networks

We start our discussion for optimal user-association by assuming that the backhaul network is provisioned and so, we can safely ignore it while deriving the optimal association rules. Our aim is to focus on the radio access network performance, and traffic-differentiation involved tradeoffs.

We remind to the reader that based on our system model, the association policy consists in finding appropriate values for the routing probabilities $p_i^{l,t}(x),\ l \in \{D,U\},\ t \in \{b,d\}$, for DL and UL, best-effort and dedicated traffic, respectively (defined earlier in assumption B.5 and B.9). That is, for each location x, we would like to optimally choose to which BS i to route different flow types generated from (UL) or destined at (DL) users in x[7]. Our goal for this association problem is threefold: (i) ensure that the capacity of no BS is exceeded (later in Section IV, we will also include the constraint of no backhaul capacity is exceeded); (ii) achieve a good tradeoff between user physical data rates, user QoS and load balancing, (iii) investigate how *UL/DL association split* impacts the optimal rule derivation and the performance benefits of split UL/DL.

We define the feasible region for the aforementioned routing probabilities, by requiring that no BS capacity being exceeded.

Definition 1. *(Feasible set): Let $l \in \{U,D\}, t \in \{b,d\}$, and let ϵ be an arbitrarily small positive constant. The set $f^{l,t}$ of feasible BS loads $\rho^{\mathbf{l},\mathbf{t}} = (\rho_1^{l,t}, \rho_2^{l,t}, \ldots, \rho_{\|\mathcal{B}\|}^{l,t})$ is*

$$\begin{aligned} f^{l,t} = \Big\{ \rho^{\mathbf{l},\mathbf{t}} \mid \rho_i^{l,t} = \int_{\mathcal{L}} p_i^{l,t}(x) \rho_i^{l,t}(x) dx, \\ 0 \le \rho_i^{l,t} \le 1 - \epsilon, \\ \sum_{i \in \mathcal{B}} p_{i,t}^l(x) = 1, \\ 0 \le p_{i,t}^l(x) \le 1, \forall i \in \mathcal{B}, \forall x \in \mathcal{L} \Big\}. \end{aligned} \tag{9}$$

Lemma 3.1. *The feasible sets $f^{D,b}, f^{D,d}, f^{U,b}, f^{U,d}$ as well as the $[f^{D,b}; f^{D,d}]$, $[f^{U,b}; f^{U,d}]$, $[f^{D,b}; f^{U,b}]$, $[f^{D,b}; f^{D,d}; f^{U,b}; f^{U,d}]$, are convex.*

Proof. The proof for the feasible set $f^{D,b}$ is presented in [9]. It can be easily adapted for the other cases, too (e.g., see [40]). □

Following [9] we extend the proposed cost function to also include the DL dedicated traffic (see B.8-B.9). We introduce the parameter $\theta \in [0,1]$ that helps the operator weigh the

[7]The use of a probabilistic association rule simplifies solving the problem. As it will turn out, the optimal values will be either 0 or 1 (deterministic).

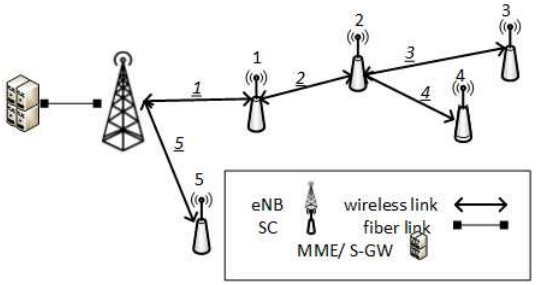


Fig. 2. Backhaul topology in future Hetnet.

importance of DL best effort versus DL dedicated traffic performance. $\alpha^{D,b}$ ($\alpha^{D,d}$) ≥ 0 controls the amount of load balancing desired in the DL best-effort (dedicated) resources. Lets denote $\alpha^{\mathbf{D}} = [\alpha^{D,b}; \alpha^{D,d}]$ and $\rho^{\mathbf{D}} = [\rho^{D,b}; \rho^{D,d}]$.

Definition 2. *(Cost function for DL) Our cost function is*

$$\phi_{\alpha^{\mathbf{D}}}(\rho^{\mathbf{D}}) = \sum_{i\in\mathcal{B}} \theta \frac{(1-\rho_i^{D,b})^{1-\alpha^{D,b}}}{\alpha^{D,b}-1} + (1-\theta)\frac{(1-\rho_i^{D,d})^{1-\alpha^{D,d}}}{\alpha^{D,d}-1}, \quad (10)$$

when $\alpha^{D,d}, \alpha^{D,b} \neq 1$. If $\alpha^{D,b}$ (or, $\alpha^{D,d}$) is equal to 1, the respective fraction must be replaced with $\log(1-\rho_i^{D,b})^{-1}$ (or, $\log(1-\rho_i^{D,d})^{-1}$).

As a final step, we want to further extend this objective to also capture the UL traffic performance and thus we introduce $\tau \in [0,1]$ to trade it off with DL. Lets assume that $\alpha = [\alpha^{D,b}; \alpha^{D,d}; \alpha^{U,b}; \alpha^{U,d}]$ and $\rho = [\rho^{D,b}; \rho^{D,d}; \rho^{U,b}; \rho^{U,d}]$.

Definition 3. *(Cost function for DL and UL) The cost function that jointly considers DL and UL performance is*

$$\phi_\alpha(\rho) = \tau \cdot \phi_{\alpha^{\mathbf{D}}}(\rho^{\mathbf{D}}) + (1-\tau)\cdot \phi_{\alpha^{\mathbf{U}}}(\rho^{\mathbf{U}}). \quad (11)$$

Lemma 3.2. *The cost function $\phi_\alpha(\rho)$ is convex in ρ.*

Proof. Since $\phi_\alpha(\rho)$ is a weighted sum of four convex functions [9], convexity is preserved [41]. □

In the rest of the section we attempt to *minimize* the cost function $\phi_\alpha(\rho)$ and derive novel user association rules.

Solution Roadmap. Our approach to solve this problem can be summarized as follows. (i) We first treat the BS loads ρ_i as the control variables and derive a sufficient condition for the optimal ρ_i values using a first order optimality condition (see e.g. Eq. (13)). (ii) Since ρ_i is a function of the association variables $p_i(x)$, we will use the optimal load to derive a condition on the optimal association decisions *conditional on the BS loads being optimal* (see e.g. Eq. (12) or Eq. (17)). These essentially say where a new flow generated at location x should go, if the BS loads where such to minimize the desired objective. Finally, (iii) we use these association rules to define an iterative algorithm (e.g. see Eq. (18)), where $p_i(x)$ are updated based on a (estimate of the) current (possibly non-optimal) load. (iv) And, we propose a distributed implementation of this iterative algorithm, and discuss why it converges to the optimal loads and thus to the optimal association map.

A. Optimal Split UL/DL User Association

We start with the Split UL/DL scenario (we come back to the Joint UL/DL scenario in the next section III-B). Here, since UL and DL traffic can be split at location x, the problem of optimal DL and UL association decouples into two independent problems: one for DL and one for UL. Specifically, $\min_\rho \phi_\alpha(\rho) = \min_{\rho^{\mathbf{D}}} \phi_{\alpha^{\mathbf{D}}}(\rho^{\mathbf{D}}) + \min_{\rho^{\mathbf{U}}} \phi_{\alpha^{\mathbf{U}}}(\rho^{\mathbf{U}})$.

Note that all DL best-effort and dedicated flows at x have to be downloaded from the same BS, i.e., $p_i^D(x) = p_i^{D,b}(x) = p_i^{D,d}(x)$. Also, that all UL best-effort and dedicated should be offloaded to the same BS, so $p_i^U(x) = p_i^{U,b}(x) = p_i^{U,d}(x)$. Nevertheless, as explained in Split UL/DL scenarios $p_i^D(x)$ and $p_i^U(x)$ can take different values (see B.10). In that case, the optimal user association rules follow. In the remainder of this section, we focus on the downlink, and we omit the superscripts $\{D, U\}$ to simplify notation.

Theorem 3.3. *If $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{\|\mathcal{B}\|}^*)$ denotes the optimal load vector, the optimal association rule at x is*

$$i(x) = \arg\max_{i\in\mathcal{B}} \frac{\left(1-\rho_i^{*b}\right)^{\alpha^b}\cdot\left(1-\rho_i^{*d}\right)^{\alpha^d}}{e^b(x)\cdot\left(1-\rho_i^{*d}\right)^{\alpha^d} + e^d(x)\cdot\left(1-\rho_i^{*b}\right)^{\alpha^b}} \quad (12)$$

where $e^b(x) = \frac{\theta z^D z^b}{S^b(x)c_i(x)}$ and $e^d(x) = \frac{(1-\theta)z^D z^d}{\mu^d(x)k_i(x)}$, optimally weight the corresponding individual rules, forming the (weighted) harmonic mean, depending on the traffic statistics.

Proof. We prove that the above association rule (Eq. 12) indeed minimizes the objective of Eq. (10). This is a convex optimization problem. Let $\rho^* = [\rho^{*b}; \rho^{*d}]$ be the optimal solution of Problem (10). (We will relax this assumption in Section III-C, as the optimal vector ρ^* is not necessarily known.) Hence, it is adequate to check for optimality if

$$\langle \nabla\phi_\alpha(\rho^*), \Delta\rho^* \rangle \geq 0 \quad (13)$$

for all $\rho \in f$, where $\Delta\rho^* = \rho - \rho^*$. Let $p(x)$ and $p^*(x)$ be the associated routing probability vectors for ρ and ρ^*, respectively. Using the deterministic cell coverage generated by(12), the optimal association rule is given by:

$$p_i^*(x) = \mathbf{1}\left\{ i = \arg\max_{i\in\mathcal{B}} \frac{\left(1-\rho_i^{*b}\right)^{\alpha^b}\cdot\left(1-\rho_i^{*d}\right)^{\alpha^d}}{e^b(x)\cdot\left(1-\rho_i^{*d}\right)^{\alpha^d} + e^d(x)\cdot\left(1-\rho_i^{*b}\right)^{\alpha^b}} \right\}. \quad (14)$$

Then the inner product in Eq. (13) can be written as:

$$\begin{aligned}
\langle \nabla\phi_\alpha(\rho^*), \Delta\rho^* \rangle &= \sum_{z=\{b,d\}} \frac{\partial\phi_\alpha}{\partial\rho_z}(\rho^*)(\rho_z - \rho_z^*) \\
&= \theta \sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^b)^{\alpha^b}}(\rho_i^b - \rho_i^{*b}) + (1-\theta)\sum_{i\in\mathcal{B}} \frac{1}{(1-\rho_i^d)^{\alpha^d}}(\rho_i^d - \rho_i^{d*}) \\
&= \sum_{i\in\mathcal{B}} \frac{\theta\int_L \rho_i^b(x)(p_i(x)-p_i^*(x))dx}{(1-\rho_i^b)^{\alpha^b}} + \frac{(1-\theta)\int_L \rho_i^d(x)(p_i(x)-p_i^*(x))dx}{(1-\rho_i^d)^{\alpha^d}} \\
&= \int_L \lambda(x)\sum_{i\in\mathcal{B}}(p_i(x)-p_i^*(x))\Big[\frac{e^b(x)(1-\rho_i^{*d})^{\alpha^d} + e^d(x)(1-\rho_i^{*b})^{\alpha^b}}{(1-\rho_i^{*b})^{\alpha^b}(1-\rho_i^{*d})^{\alpha^d}}\Big]dx,
\end{aligned} \quad (15)$$

where $e^b(x) = \frac{\theta z^{D_L} z^b}{S^b(x)c_i(x)}$ and $e^d(x) = \frac{(1-\theta)z^D z^d}{\mu^d(x)k_i(x)}$. Note that,

$$\begin{aligned}
&\sum_{i\in\mathcal{B}} p_i(x)\frac{e^b(x)(1-\rho_i^{*d})^{\alpha^d} + e^d(x)(1-\rho_i^{*b})^{\alpha^b}}{(1-\rho_i^{*b})^{\alpha^b}(1-\rho_i^{*d})^{\alpha^d}} \geq \\
&\sum_{i\in\mathcal{B}} p_i^*(x)\frac{e^b(x)(1-\rho_i^{*d})^{\alpha^d} + e^d(x)(1-\rho_i^{*b})^{\alpha^b}}{(1-\rho_i^{*b})^{\alpha^b}(1-\rho_i^{*d})^{\alpha^d}}
\end{aligned} \quad (16)$$

holds because $p^*(x)$ in (14) is an indicator for the minimizer of $\frac{e^b(x)(1-\rho_i^{*d})^{\alpha^d}+e^d(x)(1-\rho_i^{*b})^{\alpha^b}}{(1-\rho_i^{*b})^{\alpha^b}(1-\rho_i^{*d})^{\alpha^d}}$. Hence, (13) holds. □

While θ linearly weights the best effort versus dedicated flow performance (see Eq. 10), the impact of α^b, α^d is not obvious. We now discuss their impact on the system performance and refer to [9], [42] for the respective proofs.

- *Spectral Efficiency Optimization:* $\alpha^b = 0$ maximizes the average physical rate for best-effort flows (defined in B.3), whereas $\alpha^d = 0$ maximizes the average dedicated servers for dedicated flows (defined in B.8). Obviously, these optimize the user $SINR$ and spectral efficiency.
- *Optimizing related QoS metrics:* if $\alpha^b = 1$ the corresponding optimal rule tends to maximize the average user throughput. If $\alpha^b = 2$ the per-flow delay is minimized since the objective for best effort flows corresponds to the delay of an M/G/1/PS system. If $\alpha^d = 1$ the corresponding optimal rule becomes equivalent to the average *idle* dedicated servers in a k-loss system, and the actual blocking probability is minimized.
- *Load-Balancing Efficiency Optimization:* As $\alpha^b \to \infty$, we minimize the maximum BS utilization, i.e. load balancing between the ρ^b is achieved. Similar for α^d and ρ^d's. Note that, the point of α^b that all BS best-effort utilizations are equalized might be different from the one for dedicated, depending on the respective traffic statistics.

In the case of split UL/DL association, the above analysis can be applied *separately* on UL and DL traffic, and optimize UL and DL associations independently.

B. Optimal Joint UL/DL User Association

Current cellular networks (e.g. 3G/4G) suggest that a UE should be connected to a single BS for both UL and DL traffic [43], i.e. $p_i^D(x) = p_i^U(x)$. In that case, the two problems are coupled and the (single) optimal association rule at x shall appropriately weight DL and UL performance as it follows.

Theorem 3.4. *If $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{\|\mathcal{B}\|}^*)$ denotes the optimal load vector, and given the set of all flow-types $\Omega = \{(D,b),(D,d),(U,b),(U,d)\}$, the optimal rule at x is*

$$i(x) = \arg\max_{i\in\mathcal{B}} \frac{\prod_{c\in\Omega}((1-\rho^{*c})^{\alpha^c})}{\sum_{c\in\Omega} e^c(x) \prod_{l\in\Omega\neq c}((1-\rho^{*c})^{\alpha^c})}, \quad (17)$$

where $e^{D,b}(x) = \tau\frac{\theta^D z^D z^b}{S^{D,b}(x)c_i^{D,b}(x)}$, $e^{D,d}(x) = \tau\frac{(1-\theta^D)z^D z^d}{\mu^{D,d}(x)k_i^D(x)}$, $e^{U,b}(x) = (1-\tau)\frac{\theta^U z^U z^b}{S^{U,b}(x)c_i^{U,b}(x)}$ and $e^{U,d}(x) = (1-\tau)\frac{(1-\theta^U)z^U z^d}{\mu^{U,d}(x)k_i^U(x)}$ are the corresponding weight factors.

Proof. We refer the interested reader to [40]. □

Remark 1. The above optimal rule derived in Eq. (17) suggests that in the *Joint UL/DL* scenario associated with objectives that potentially conflict with each other (due to the different flow type performances), it is optimal to associate a user with the BS that maximizes a weighted version of the *harmonic mean* of the individual association rules when considering each objective alone. To better understand this, we focus on a simple scenario with only DL and UL best-effort traffic. And assume the following BS options for a user: (BS A) offers 50Mbps DL and only 1Mbps UL; (BS B) 200Mbps DL and 0.5Mbps UL; (BS C) 20Mbps DL and 5Mbps UL. If we care about UL and DL performance equally (i.e. $\tau = 0.5$), one might assume that the BS that maximizes the arithmetic mean (or arithmetic sum) of rates would be a fair choice (i.e. BS B). However, this would lead to rather poor UL performance. Maximizing the harmonic mean would lead to choosing (BS C) instead[8]. Additionally, note that in the case of *split UL/DL*, covered in Section III-A, where each user is free to be associated with two different BSs for the DL and UL traffic offloading, DL traffic would be associated with (BS B), and UL traffic with (BS C) by maximizing the arithmetic mean (or, sum) of their throughputs [9]. These simple examples intuitively explain how split UL/DL impacts the user association policies, by allowing to independently optimize each objective. This also demonstrates why UL/DL split may perform considerably better than the joint association. We will further explore this in the Simulations.

Summarizing, we showed that our framework supports both best-effort and dedicated (in both DL and UL direction) traffic demand and we analytically found the corresponding optimal rules in different scenarios. Nevertheless, we highlight that our derived formulas allow to add more dimensions in our setup and *flexibly* derive the optimal rules without any analytical calculations (e.g., by using the arithmetic or harmonic mean maximization). For instance, consider a more modern offloading technique, where different downlink, or uplink, flow types are able to be offloaded to different BSs (e.g., per flow/QCI offloading) with conflicting aims. Using our model we can consider an additional respective α-function for each flow type, and either analytically or flexibly, optimize the complete objective as showed earlier.

C. Iterative Algorithm Achieves Optimality

Eq. (12) (or, Eq. (17)) does not yet solve the problem, but only defines the conditions that should hold at the optimal point. They essentially define an optimal *association map*: *if* all BS loads have converged to the optimal values, then a new traffic flow generated at location x should be associated according to Eq. (12). If the BSs do not currently have the optimal loads ρ_i^*, a distributed iterative algorithm can be implemented by repeating the following k steps until convergence (based on Eq. (12)):

Base Station. Each BS maintains an estimate $\hat{\rho}_i$ of its average utilization load. To deal with the utilization constraint ($\rho_i < 1-\epsilon$), the parallel and potentially asynchronous updates of $p_i(x)$ variables, and non-stationarities in the traffic demand, the BS load estimate $\hat{\rho}_i$ is updated regularly as follows:

$$\hat{\rho}_i^{(k+1)} = (1-\beta^{(k)})\cdot\rho_i^{(k)} + \beta^{(k)}\cdot\hat{\rho}_i^{(k)}. \quad (18)$$

[8]While this simple example captures the main principle, the actual rule is more complex, as it weighs each objective with the complex factor $e^l(x)$.

[9]The usage of harmonic mean and arithmetic mean/sum appears in a number of physical examples, such as in the calculation of the total resistance in circuits where all resistances are set in series or in parallel.

This is an exponential moving average with parameter $\beta^{(k)} \in [0,1)$. [10] $\rho_i^{(k)}$ is the current load measurement while $\hat{\rho}_i^{(k)}$ is the current load average estimate. $\hat{\rho}_i^{(k+1)}$ is used for the next iteration broadcast message.

Mobile Device. Each user receives the BS broadcast message and updates its association variables $p_i(x)$ (the real control variables) according to Eq. (12), where ρ_i^* is now replaced with $\hat{\rho}_i$.

The above algorithm essentially implements a distributed gradient descent on the ρ_i. Starting within a feasible point $\rho^{(0)}$ it converges to the global optimum by requiring a simple modification of the proof found on the original algorithm [9]. (The descent direction at x, improving the objective at the next $k+1$ iteration i.e. satisfying

$$\langle \nabla \phi_\alpha(\rho^{(k)}), \rho^{(k+1)} - \rho^{(k)} \rangle < 0, \tag{19}$$

is now provided from Eq. (12) under joint DL dedicated/DL best-effort association. This formula appropriately projects the direction under the constraint $p_i^b(x) = p_i^d(x)$; as shown in the proof of Theorem 3.3. Similarly for the Joint UL/DL association, where the problem does not decouple, as well.)

IV. User-Association for Under-Provisioned Backhaul Networks

While the rules derived above, that try to reflect different performance tradeoffs, always lead to BS loads that are supported from the access network, they perhaps will not be supported from the backhaul link (or the corresponding backhaul link path) for that BS, since they ignore potential backhaul limitations. To that end, in this section we try to extensively consider the backhaul network and related limitations while extracting the optimal association rules, and include to our goals (i) that no backhaul link is congested, (ii) the impact of backhaul topology and capacity on key performance metrics. In order to better elucidate it at hand and without loss of generality, we focus on a simple scenario with *only best-effort traffic*. So, in the remainder of the section we drop the corresponding superscripts "b", "d" to simplify notation.

One of the main challenges when attempting to consider these backhaul constraints is to maintain the user association policy *distributed* (famous solvers for such convex problems, e.g. through the Lagrangian dual function [41], require a centralized controller entity). To that end, we chose to consider the backhaul constraints in the cost function as appropriate *penalty functions* [44]. This not only facilitates deriving a distributed implementation of the policy, but also allows us to treat the backhaul constraint as a "soft" constraint that ends up being "hard" and satisfy convergence to a feasible solution.

A. Optimal Split UL/DL User Association

We follow the same presentation as the provisioned case, and start out discussion with the Split UL/DL case. As the association problem can be decoupled, in that case, into two independent problems, we focus on the optimal DL association problem, and we omit the superscripts $\{D,U\}$. We return to the Joint UL/DL case in the next section. To better illustrate our approach, we first consider a simple BH star topology, and then generalize for tree topologies.

[10] In the Split UL/DL scenario the UL and DL loads can be independently updated, otherwise they should use the same $\beta^{(k)}$.

Optimal User Association for Star BH Topology)

In the following, since for star topologies there is exactly one backhaul link (j) associated with each BS (i), it is $i = j$ (see C.1). Let $\mathcal{I}(i)$ be an indicator variable showing if the i-th BH link is congested ($\mathcal{I}(i)$=1) or not ($\mathcal{I}(i)$=0) (see C.2)

$$\mathcal{I}(i) = \begin{cases} 0, & \text{when } \frac{\rho_i \tilde{c}_i}{C_h(i)} < 1 \\ 1, & \text{otherwise.} \end{cases} \tag{20}$$

Furthermore, the objective shall be extended to include the penalty functions for the backhaul constraints, as it follows:

$$\phi_\alpha(\rho) = \sum_{i \in \mathcal{B}} \frac{(1-\rho_i)^{1-\alpha}}{\alpha - 1} + \gamma \sum_{i \in \mathcal{B}_h} \mathcal{I}(i) \left(\frac{\rho_i \tilde{c}_i}{C_h(i)} - 1 \right)^2. \tag{21}$$

The first term is the standard α-cost function for each BS i, already analyzed in the previous section. The second sum introduces a penalty for each backhaul link i whose capacity is exceeded ($\mathcal{I}(i) = 1$). γ could be chosen as a small constant, introducing a "soft" constraint for the backhaul links (i.e., backhaul capacity could be slightly exceeded, if this really improves access performance), or, preferably, *be iteratively adapted using increasing values, so as to converge to a "hard" constraint*. This penalty function is quadratic on the amount of excess load (quadratic penalty functions are often considered in convex optimization literature [45]). The corresponding optimal backhaul-aware rules follow.

Theorem 4.1. *If $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{\|\mathcal{B}\|}^*)$ denotes the optimal load vector, the optimal association rule at location x is*

$$\arg\max_{i \in \mathcal{B}} c_i(x) \frac{(1-\rho_i^*)^\alpha}{1 + 2\gamma \cdot (1-\rho_i^*)^\alpha \cdot \tilde{c}_i \cdot \frac{\mathcal{I}(i)}{C_h(i)} \cdot \left(\frac{\rho_i^* \tilde{c}_i}{C_h(i)} - 1 \right)}. \tag{22}$$

Proof. We now prove that the above rule indeed minimizes the cost function of Eq. (21) within the penalty term γ. This is a convex optimization problem (quadratic penalty functions are convex due to the composition property of convexity [41]). Let ρ^* be the optimal solution of this problem. (We will relax this assumption in Section IV-C, as the optimal vector ρ^* is not necessarily known.) Again, it is adequate to check if

$$\langle \nabla \phi_\alpha(\rho^*), \Delta\rho^* \rangle \geq 0 \tag{23}$$

for all $\rho \in f$, where $\Delta\rho^* = \rho - \rho^*$. Let $p(x)$ and $p^*(x)$ be the associated routing probability vectors for ρ and ρ^*, respectively. Using the deterministic cell coverage generated by (22), the optimal association rule is given by:

$$p_i^*(x) = \mathbf{1}\left\{ i = \arg\max_{i \in \mathcal{B}} \frac{c_i(x)(1-\rho_i^*)^\alpha}{1 + 2\gamma \cdot (1-\rho_i^*)^\alpha \cdot \tilde{c}_i \cdot \frac{\mathcal{I}(i)}{C_h(i)} \cdot \left(\frac{\rho_i^* \tilde{c}_i}{C_h(i)} - 1 \right)} \right\}. \tag{24}$$

The i-th element of the derivative is

$$\nabla \phi_\alpha(\rho_i) = \begin{cases} (1-\rho_i)^{-\alpha}, & \text{if } \frac{\rho_i \tilde{c}_i}{C_h(i)} \leq 1 \\ (1-\rho_i)^{-\alpha} + \gamma \mathcal{I}(i) \frac{2\rho_i \tilde{c}_i^2 - 2\tilde{c}_i C_h(i)}{C_h(i)^2}, & \text{if } \frac{\rho_i \tilde{c}_i}{C_h(i)} \geq 1. \end{cases} \tag{25}$$

The inner product defined in Eq. (23), becomes:

$$\langle \nabla\phi_\alpha(\rho^*), \Delta\rho^* \rangle = \sum_{i\in\mathcal{B}} \Big\{ \frac{1}{(1-\rho_i^*)^\alpha} + \gamma\mathcal{I}(i) \frac{2\rho_i^*\tilde{c}_i^2 - 2\tilde{c}_i C_h(i)}{C_h(i)^2} \Big\} (\rho_i - \rho_i^*)$$

$$= \sum_{i\in\mathcal{B}} \frac{1 + 2\gamma\mathcal{I}(i)(1-\rho_i^*)^\alpha \frac{(\rho_i^*\tilde{c}_i^2 - \tilde{c}_i C_h(i))}{C_h(i)^2}}{(1-\rho_i^*)^\alpha} \int_{\mathcal{L}} \rho_i(x) \big(p_i(x) - p_i^*(x)\big)\, dx$$

$$= \int_L \frac{\lambda(x)}{S(x)} \sum_{i\in\mathcal{B}} \left(\frac{1 + 2\gamma(1-\rho_i^*)^\alpha \tilde{c}_i \frac{\mathcal{I}(i)}{C_h(i)} \left(\frac{\rho_i^*\tilde{c}_i}{C_h(i)} - 1 \right)}{c_i(x)(1-\rho_i^*)^\alpha} \right) \big(p_i(x) - p_i^*(x)\big)\, dx.$$

Note that,

$$\sum_{i\in\mathcal{B}} p_i(x) \left\{ \frac{1 + 2\gamma(1-\rho_i^*)^\alpha \tilde{c}_i \frac{\mathcal{I}(i)}{C_h(i)} \left(\frac{\rho_i^*\tilde{c}_i}{C_h(i)} - 1 \right)}{c_i(x)(1-\rho_i^*)^\alpha} \right\} \geq$$

$$\sum_{i\in\mathcal{B}} p_i^*(x) \left\{ \frac{1 + 2\gamma(1-\rho_i^*)^\alpha \tilde{c}_i \frac{\mathcal{I}(i)}{C_h(i)} \left(\frac{\rho_i^*\tilde{c}_i}{C_h(i)} - 1 \right)}{c_i(x)(1-\rho_i^*)^\alpha} \right\}$$

holds because $p_i^*(x)$ in (24) is an indicator for the minimizer of $\frac{1+2\gamma\cdot(1-\rho_i^*)^\alpha\cdot\tilde{c}_i\cdot\frac{\mathcal{I}(i)}{C_h(i)}\cdot\left(\frac{\rho_i^*\tilde{c}_i}{C_h(i)}-1\right)}{c_i(x)(1-\rho_i^*)^\alpha}$. Hence (23) holds. □

Regarding the optimal association rule of Eq. (22), we note that when the capacity constraint for the backhaul link i is not active (i.e., $\mathcal{I}(i) = 0$, in provisioned BH networks), the above theorem states that the optimal association rule is the same as the one found in [9], or the one defined in Eq. (12) when $\theta \to 1$. However, when the backhaul link of BS i gets congested, a second term is added in the denominator that penalizes that BS making it less preferable to UEs at location i, even if the offered radio access rate $c_i(x)$ is high, or the radio interface of i is not itself congested.

Optimal User Association for Tree BH Topology)

We now consider a more complex backhaul scenario, where a single backhaul link might route traffic from multiple BSs, and the traffic of a single BS might be routed over multiple backhaul links (multi-hop path) towards the eNB. $\mathcal{I}(j)$ is now

$$\mathcal{I}(j) = \begin{cases} 0, & \text{when } \frac{\sum_{i\in\mathcal{B}(j)} \rho_i\tilde{c}_i}{C_h(j)} < 1 \\ 1, & \text{otherwise.} \end{cases} \tag{26}$$

Similarly, the backhaul constraints shall be modified appropriately, and the cost function eventually becomes

$$\phi_\alpha(\rho) = \sum_{i\in\mathcal{B}} \frac{(1-\rho_i)^{1-\alpha}}{\alpha - 1} + \gamma \sum_{j\in\mathcal{B}_h} \mathcal{I}(j) \left(\frac{\sum_{i\in\mathcal{B}(j)} \rho_i\tilde{c}_i}{C_h(j)} - 1 \right)^2. \tag{27}$$

Theorem 4.2. *If $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{\|\mathcal{B}\|}^*)$ denotes the optimal load vector, the optimal association rule at location x is*

$$\arg\max_{i\in\mathcal{B}} c_i(x) \frac{(1-\rho_i^*)^\alpha}{1 + 2\gamma\cdot(1-\rho_i^*)^\alpha \cdot \tilde{c}_i \sum\limits_{j\in\mathcal{B}_h(i)} \frac{\mathcal{I}(j)}{C_h(j)} \cdot \left(\frac{\sum_{k\in\mathcal{B}(j)} \rho_k^*\tilde{c}_k}{C_h(j)} - 1 \right)} \tag{28}$$

Proof. The steps of this proof are similar to the star case, so we present here directly the corresponding inner product.

$$\langle \nabla\phi_\alpha(\rho^*), \Delta\rho^* \rangle =$$

$$= \sum_{i\in\mathcal{B}} \Big\{ \frac{1}{(1-\rho_i^*)^\alpha} + 2\gamma \sum_{j\in\mathcal{B}_h(i)} \mathcal{I}(j) \big[\frac{\sum_{k\in\mathcal{B}(j)} \rho_k^*\tilde{c}_k}{C_h(j)^2} \tilde{c}_i - \frac{\tilde{c}_i}{C_h(j)} \big] \Big\} (\rho_i - \rho_i^*)$$

$$\cdot \int_{\mathcal{L}} \rho_i(x) \big(p_i(x) - p_i^*(x)\big)\, dx =$$

$$= \int_L \frac{\lambda(x)}{S(x)} \sum_{i\in\mathcal{B}} \left(\frac{1 + 2\gamma(1-\rho_i^*)^\alpha \tilde{c}_i \sum\limits_{j\in\mathcal{B}_h(i)} \frac{\mathcal{I}(j)}{C_h(j)} \cdot \left(\frac{\sum_{k\in\mathcal{B}(j)} \rho_k^*\tilde{c}_k}{C_h(j)} - 1 \right)}{c_i(x)(1-\rho_i^*)^\alpha} \right) \cdot$$

$$\cdot \big(p_i(x) - p_i^*(x)\big)\, dx \geq 0, \tag{29}$$

due to the corresponding maximizer $p_i^*(x)$ derived from (28). □

There are a number of interesting differences between the optimal association rules of star and tree topology. First, the penalty term in the denominator of the rule (Eq. (28)) now considers the whole backhaul path $\mathcal{B}_h(i)$ that traffic from BS i traverses, and adds a penalty for *every* link along that path that is congested (outer sum in the denominator). This observation provides some support for the number of BH hops heuristic proposed in [46], [29]. However, our analysis also suggests that it can be suboptimal, as a path with few hops might still include one or more congested links, and provides the optimal way to weigh in the amount of congestion on each link.

Second, the actual congestion on each backhaul link j is now not only dependent on the load of the candidate BS i, but also on other BSs whose load is routed over j. Hence, a BS i which would otherwise be a good candidate for traffic at location x, might still be penalized and not selected, even if it does not impose itself a large load on a backhaul link j. This is because *other* BSs sharing the same backhaul link might be heavily loaded or congested.

In the case of split UL/DL traffic, the above analysis can be applied *separately* on UL and DL traffic, and optimize UL and DL associations independently. Finally, although we have provided separate solutions for star and tree topologies, to better illustrate our approach, the optimal rule for the tree topology is generic, and includes star topologies as well.

B. Optimal Joint UL/DL User Association

In the Joint UL/DL case, we remind the reader that each user at x shall be associated with one BS for both UL and DL traffic, i.e. $p_i^D(x) = p_i^U(x)\ \forall i \in \mathcal{B}$, as discussed in Section III-B. The penalty function should now consider both uplink and downlink capacity being exceeded on each backhaul link, and the cost function $\phi_\alpha(\rho)$ becomes

$$\tau\phi_{\alpha^{\mathbf{D}}}(\rho^{\mathbf{D}}) + (1-\tau)\phi_{\alpha^{\mathbf{U}}}(\rho^{\mathbf{U}}) + \gamma \sum_{k\in\{D,U\}} \sum_{j\in\mathcal{B}_h} \mathcal{I}^k(j) \left(\frac{\sum_{i\in\mathcal{B}(j)} \rho_i^k\tilde{c}_i^k}{C_h^k(j)} - 1 \right)^2 \tag{30}$$

Here, we present our results directly for the general case of tree backhaul topology, and we remind the reader that this is applicable to star backhaul topologies as well.

Theorem 4.3. *If $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{\|\mathcal{B}\|}^*)$ denotes the optimal load vector, the optimal user-association rule at location x is*

$$i(x) = \arg\max_{i\in\mathcal{B}} \frac{\left(1-\rho_i^{*D}\right)^{\alpha^D}\cdot\left(1-\rho_i^{*U}\right)^{\alpha^U}}{e^D(x)\cdot\left(1-\rho_i^{*U}\right)^{\alpha^U} + e^U(x)\cdot\left(1-\rho_i^{*D}\right)^{\alpha^D}}, \tag{31}$$

where if $g^D = \tau, g^U = 1-\tau$, then for $l \in \{D, U\}$:

$$e^l(x) = \frac{z^l\left(g^l + 2\gamma\left(1-\rho_i^{*l}\right)^{\alpha^l}\sum\limits_{j\in\mathcal{B}_h(i)}\frac{\mathcal{I}^l(j)}{C_h^l(j)}\left(\frac{\sum_{k\in\mathcal{B}(j)}\rho_k^{*l}\tilde{c}_k^l}{C_h^l(j)}-1\right)\right)}{S^l(x)c_i^l(x)}.$$

Proof. We refer the interested reader to [40]. □

C. Iterative Algorithm Achieves Optimality

Our proposed iterative framework stays similar in nature as the one described in Section (III-C).

We now focus on the penalty method and the convergence to the global optimal point. Using the proposed (quadratic) penalty functions we now solve a sequence of unconstrained problems (e.g. see Eq. (21)) with *monotonically increasing values* of γ at each iteration (chosen so that the solution to the next problem is "close" to the previous one; otherwise we risk getting stuck in steep valleys). Thus, let $\rho^{(\mathbf{k})} = \{\rho^{(0)}, \rho^{(1)}, ..., \rho^{(k)}\}$ be a sequence generated within iterations, where $\rho^{(l)}, 0 \le l \le k$, is the global minimum of $\phi(\rho)$ with penalty constant $\gamma^{(l)}$ at the l iteration. Then any limit point $\rho^{(k)}$ of the sequence is a solution to this problem. This is a well known result for such convex cost functions proposed from Luenberger in 1984 [47].

Remark 2. In this remark we want to underline various important properties of our derived rules (regarding both Sections III and IV). Firstly, no matter the number of traffic types and the backhaul topology design, all derived rules are "device centric". I.e., the user is able to optimally select where to associate based on (i) *its own measurements*, e.g. $c_i(x)$ in Eq. (22), and (ii) *BS broadcast information* that jointly capture its access and backhaul performance, e.g. the fraction seen in Eq. (22). [11] The latter clearly allows for distributed implementations that do not require any controller to govern the BSs and the UEs with access to all the necessary information, e.g., as in [51] . [12] The backhaul penalties can be seen as the "prices" calculated by each backhaul link and sent to all connected BSs to express their usage. Alternatively, the link cost could be measured implicitly [52] and then broadcasted together with the BS loads $\hat{\rho}_i$, by allowing for different distributed implementation setups. Finally, note that our derived rules also satisfy the following important properties for distributed frameworks: *scalable* (constant amount of the BS broadcast messages irrespective of the number of users, backhaul topology), *simple* (constant complexity of the rule with respect to the number of BSs), and offer *flexible performance optimization* (through α values).

[11] Such broadcast quantities can be easily integrated through the newly proposed Access Network Discovery and Selection Function (ANDSF) mechanism [48], or in the absolute/dedicated priority list mechanisms of LTE [49], or in IEEE 802.16m [50].

[12] Such centralized schemes usually imply a high (a) burden of collecting all the necessary information to a central location (usually implemented in a server deep in the core network), processing, and redistributing to every user, (b) computational complexity that increases exponentially in the network size.

TABLE II
SIMULATION PARAMETERS

Parameter	Variable	Value
Transm. Power of eNB/ SC/ UE	$P_{eNB}/P_{SC}/P_{UE}$	43/24/12 dBm
BS Bandwidth for DL, UL	w/W	10/10 MHz
Noise Power Density	N_0	-174 dBm/Hz
Splitting parameter for DL, UL	ζ_i^D, ζ_i^U	0.5/0.5
Average DL/UL flow sizes	$\frac{1}{S^{D,b}}/\frac{1}{S^{U,b}}$	100/20 Kbytes
Average DL/UL flow demands	$B^D(x)/B^U(x)$	512, 128 kbps
Different flow ratios	z^b, z^D	0.3,0.6

V. SIMULATIONS

In this section we briefly present some numerical results and discuss related insights. We consider a $2\times 2\ km^2$ area. Figure 3(a) shows a color-coded map of the heterogeneous traffic demand $\lambda(x)$ ($flows/hour$ per unit area) (blue implying low traffic and red high), with 2 hotspots. We assume that this area is covered by two macro BSs and eight SCs. The macro BSs that are shown with asterisks are numbered from 1-2, and the SCs that are shown with triangles are numbered from 3-10, as we can see in Fig. 3(b)-(c), Fig. 4, and in Fig. 5. We also consider standard parameters as adopted in 3GPP [53], listed in Table II[13]. If not explicitly mentioned, we assume $\theta^D = \theta^U = \tau = 0.5$, and the split UL/DL scenario as default.

Before proceeding, we need to setup a metric to evaluate load balancing (or, utilization) efficiency. Thus, we introduce the Mean Squared Error ($MSE^{D,b}$), between the DL best-effort utilization of different BSs, normalized to 1:

$$\text{MSE}^{D,b} = \frac{1}{2\cdot\left\lfloor\frac{\|\mathcal{B}\|}{2}\right\rfloor\cdot\left\lceil\frac{\|\mathcal{B}\|}{2}\right\rceil}\sum_i\sum_j(\rho_i^{D,b}-\rho_j^{D,b})^2. \tag{32}$$

We define the DL load balancing metric for best-effort traffic to be $1 - MSE^{D,b}$, that increases on the amount of load balancing . Similarly, we can define them for the other three cases $1 - MSE^{D,d}$, $1 - MSE^{U,b}$, $1 - MSE^{U,d}$.

A. Provisioned Backhaul

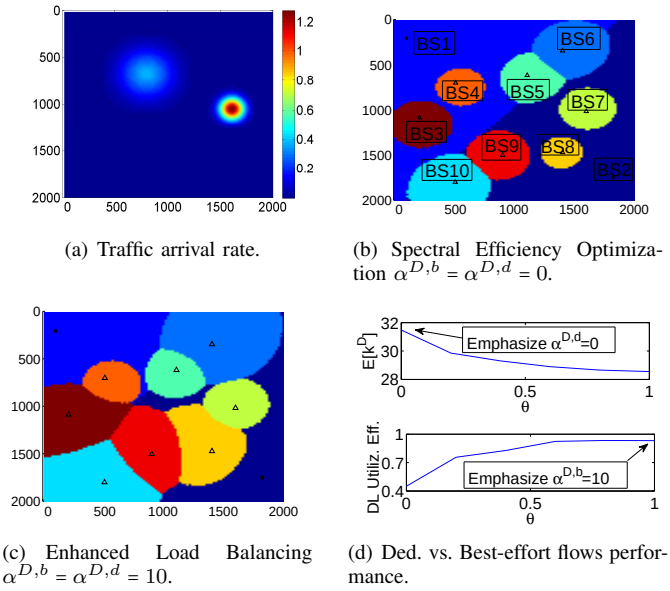
We now focus on the case of provisioned backhaul as considered in Section III and investigate the involved tradeoffs both qualitatively and quantitatively. We will present the impact of our proposed association rules via coverage snapshots to show how users associate in the considered network, while we will also provide values for related performance metrics that complete our study numerically.

Spectral efficiency vs. Load balancing. Figure 3(b) outlines the optimal DL user-associations if $\alpha^{D,b} = \alpha^{D,d} = 0$, i.e., when *spectral efficiency* is maximized. Thus, each UE at x is attached to the BS that offers the *highest DL SINR* and promises higher DL physical rate for best effort flows $c_i^{D,b}(x)$, and more "dedicated" servers $k_i^D(x)$; i.e. most of UEs are

[13] As for (i) the sizes and ratios of different flows, (ii) splitting parameters, we can use different values in order to capture different simulation scenarios, and derive similar results.

attached to macro BSs due to their high power transmission, and fewer to SCs, forming small circles around them. Consequently, macrocells are overloaded and load imbalance within the cells is sharpened (decreased $1-MSE^{D,b}$, $1-MSE^{D,d}$; see line 1 of Table III). However, in Fig. 3(c) we emphasize the *load-balancing* efficiency and set $\alpha^{D,b}=\alpha^{D,b}=10$. Now, most SCs vastly increase their coverage area in order to offload the overloaded macro BSs (e.g., BSs 6, 8, 10); "heavily" loaded (due to the hotspots) BSs, roughly maintain the same coverage (BS 4 and 7). Thus load balancing is improved, at the cost of $E[c^{D,b}], E[k^D]$ (see line 2 of Table III). For further implications of α parameters we refer the reader to [9].

Best-effort versus dedicated traffic performance. Although in the previous scenarios the best-effort- and dedicated- related traffic rules (represented from $\alpha^{D,b}, \alpha^{D,d}$) are aligned, one could ask how would two conflicting optimization objectives affect our network? The answer lays in the usage of θ^D, that judges which objective carries more importance. E.g., an operator has two main goals: (i) to maximize the average number of servers for "dedicated" traffic captured by $E[k^D]$ (set $\alpha^{D,d}=0$), (ii) to better balance the utilization of best-effort resources between BSs (set $\alpha^{D,b}=10$). As shown in Fig. 3(d), if $\theta\to 0$ $E[k^D]$ is maximized, whereas as $\theta\to 1$, $1-MSE^{D,b}$ (DL best-effort load balancing) is optimized, and each objective comes at the price of the other.



(a) Traffic arrival rate.

(b) Spectral Efficiency Optimization $\alpha^{D,b}=\alpha^{D,d}=0$.

(c) Enhanced Load Balancing $\alpha^{D,b}=\alpha^{D,d}=10$.

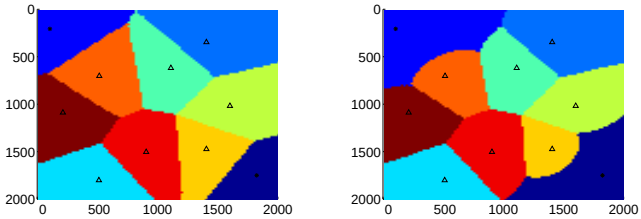
(d) Ded. vs. Best-effort flows performance.

Fig. 3. DL Optimal user-associations (Spectral efficiency vs. Load balancing and best-effort vs. ded. traffic performance)

TABLE III
NUMERICAL VALUES FOR FIGURE 3.

	Rates and Servers		**Load Balancing**	
	$E[c^{D,b}]$ (Mbps)	$E[k^D]$	1-$MSE^{D,b}$	1-$MSE^{D,d}$
Fig. 3(b)	16.3	32	0.77	0.78
Fig. 3(c)	14.3	27	0.96	0.995

DL vs. UL traffic performance is considered in Figure 3(b), 4(a)-4(b), with respective numerical performance metrics in Table IV. The first two figures depict the DL and UL optimal associations, in case of split UL/DL, for each user at x. However, if split is not available from the operator point of view, we have to weight whether the DL or UL performance is more important while selecting a *single* BS for Joint UL/DL association, using parameter τ. To that end, Figure 3(b) (also) outlines the optimal associations in the Joint UL/DL case if the whole emphasis is on the *DL performance* ($\tau=1$): this hurts the UL performance due to the asymmetric transmission powers of the UEs and BSs (see line 1 of Table IV). In Fig. 4(a) the emphasis is moved on the *UL performance* ($\tau=0$), and each UE is attached to the nearest BS, in order to minimize the path loss [38] and enhance the UL performance; this hurts its DL performance though (see line 3 of Table IV). Finally, Fig. 4(b) shows the optimal coverage areas when one assigns equal importance to the UL and DL performance (i.e. $\tau=0.5$): this moderates both DL and UL performance (line 2 of Table IV). This also corroborates the notion that split is able to simultaneously optimize UL and DL performances, as already discussed in theory.



(a) UL performance optimization $\tau=0$.

(b) Equal emphasis of DL vs. UL $\tau=0.5$.

Fig. 4. Optimal user-associations (DL vs. UL traffic performance)

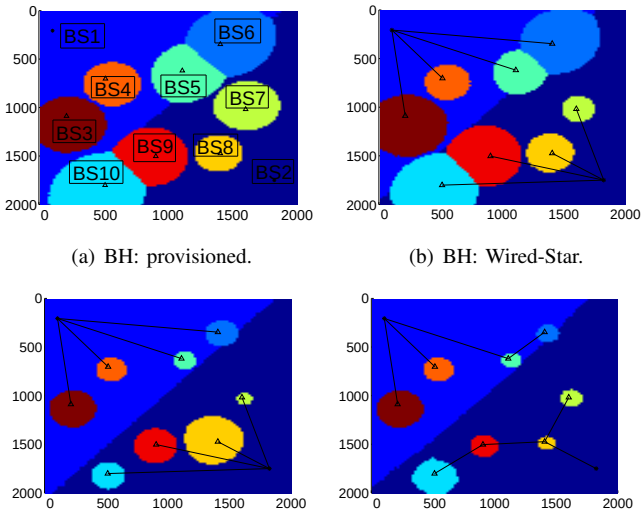
TABLE IV
NUMERICAL VALUES FOR FIGURE 4.

	DL performance		**UL performance**	
	$E[c^{D,b}]$ (Mbps)	$E[k^D]$	$E[c^{U,b}]$ (Mbps)	$E[k^U]$
Fig. 3(b)	16.3	32	2.3	18
Fig. 4(b)	14.7	28	3	24
Fig. 4(a)	13.3	26	3.6	28

B. Under-provisioned Backhaul

We now continue with some backhaul-limited network scenarios. We remind to the reader that our focus is on the backhaul links *between the macro cells and SCs* (for simplicity we assume provisioned links between the macro cells and core network). As already discussed in assumption C.1, we investigate two different backhaul topology families: (i) "star" topologies (single-hop paths), (ii) "tree" topologies (with multi-hop paths), along with two backhaul links types: *wired and wireless*.Our aim is to evaluate the derived association rules described in Section IV for different *under-provisioned* scenarios, by fixing the aforementioned trade-offs related to the traffic differentiation as it follows: $\theta^D=\theta^U=1$ (we only

focus on the best-effort flows by dropping the superscripts "b" and "d"), and $\alpha^D = \alpha^U = 1$ (throughput optimal values). Also, we assume *fixed* backhaul routing paths, pre-established with traditional Layer 2 routing, that the BH capacities on the DL and UL are the same (i.e. $C_h^D(j) = C_h^U(j) = C_h,\ \forall j \in \mathcal{B}_h$), and if not explicitly mentioned we assume them to be equal to $400Mbps$. We maintain this assumption to facilitate our discussion, although our framework works for heterogeneous backhaul links and UL/DL capacities (see C.2). We finally assumed that $\tilde{c}_i$ is the 80% of the maximum user capacity associated with BS i (we have also tried higher values for more "conservative" scenarios, and also lower for more "aggressive" with similar conclusions).



(a) BH: provisioned. (b) BH: Wired-Star.

(c) BH: Wireless-Star. (d) BH: Wireless-Tree.

Fig. 5. DL optimal associations in different scenarios.

Before proceeding, we need to make an assumption about the backhaul link capacities. In case of *wired* backhaul links, we assume that the peak backhaul capacity C_h is always guaranteed. For *wireless* backhaul links we adopt a simple model associating peak backhaul capacity to distance: if the length of the i-th link is r_i, the peak capacity drops as:

$$d(r_i) = \begin{cases} 1, & r_i \leq r_0 \\ (\frac{r_0}{r_i})^n, & \text{otherwise,} \end{cases} \quad (33)$$

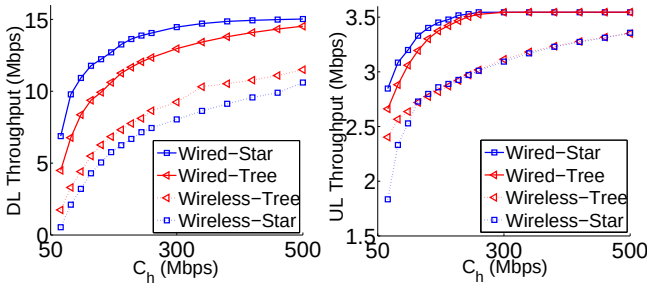
where r_0 is some threshold range within which the maximal rate is obtained (e.g. Line-of-Sight), and n is the attenuation factor. Hence, the available capacity drops to $d(r_i)C_h(j)$ ($\leq C_h(j)$). For our simulations, we assumed that $r_0 = 200m$, and $n = 3$. While the above model is perhaps oversimplifying, our main goal is to simply include a generic model for the propagation related impact on wireless backhaul, compared to wired, without getting into the details of specific backhaul implementations. For detailed path loss models for different backhaul technologies, we refer the interested reader to [34].

Coverage Snapshots. In Fig. 5(a) we depict the optimal DL user-associations for provisioned backhaul network with respect to the traffic arrival rates shown in Fig. 3(a). Compared to the associations showed in Figure 3(b) where $\alpha^D = \alpha^U = 0$, we note that now some SCs have slightly increased coverage area, in order to improve the mean user throughput [9].

In the following, we focus on different *under-provisioned* backhaul scenarios, and study the DL associations (similar behavior in the UL; we refer the interested reader to [40] for them). In Fig. 5(b) we adopt a *wired-star* backhaul topology, where SCs shrink their coverage areas, by handing-over users to other BSs, in order to offload the corresponding (under-provisioned) backhaul links; this phenomenon becomes more intense in the "hot-spot" areas (e.g., BS7 have vastly decreased their coverage areas) due to the higher traffic demand. Similarly, in Fig. 5(c), we assume a *wireless-star* backhaul topology, where SCs further decrease their coverage areas, due to the higher backhaul capacity loss caused from the long wireless links (see Eq.(33)).

In Fig. 5(d) we adopt a *wireless-tree* topology, where some SCs are required to carry also traffic of other SCs, and end up more congested. As a result, most SCs further decrease their coverage area, compared to the star-wireless topology. However, BS7 and BS10 enlarge their coverage areas, compared to the star case. This occurs because these SCs are far from the eNB, and multi-hop topology allows them to route their traffic over shorter wireless links with smaller capacity losses, compared to the star case (Fig. 5(c)). Hence, there are two main factors affecting the coverage areas in such wireless backhaul networks: (*topology*) each BS-load might traverse through multi-hop backhaul paths, by "wasting" resources from more than one backaul links (drawback for tree topologies); *(location)* the higher the η,r_0 the worse the capacity loss "wastage" over a dedicated direct backhaul link (drawback for star topologies that require longer links).

As backhaul networks become increasingly complex, e.g. "mesh" topologies, each BS has *multiple* possible routing paths to follow, beyond what is shown in the figures (we remind the reader that the above shown topologies are simply the given spanning routing trees). The above observations thus underline the shortcomings of predetermined, Layer 2 (L2) backhaul routing mechanisms, and call for a *joint* optimization of user-association on the radio access network along with dynamic, Layer 3 (L3) backhaul routing.



(a) DL (global) user throughput. (b) UL (global) user throughput.

Fig. 6. Mean throughputs overall all users in the network.

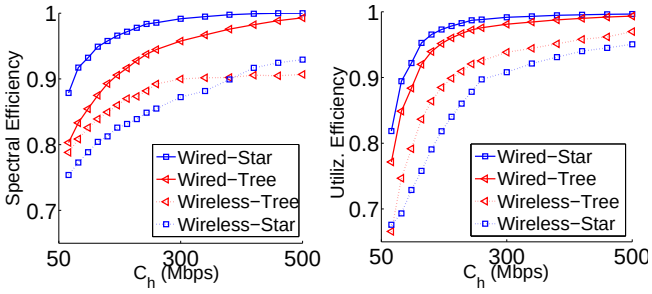
Under-provisioning impact on user performance. Figure 6(a), 6(b) depict the *average* DL and UL user throughputs, as a function of the backhaul capacity constraint C_h, on different scenarios. Generally, as C_h drops, the mean through-

puts are decreased, since users are handed over to (potentially far-away) macro BSs, causing performance degradation. Interestingly, *the slope of the dropping rate* becomes more steep for lower values of C_h, due to the logarithimic capacity formula chosen in assumption (B.2). Also, as C_h increases, the average throughputs "converge" to the value corresponding to a provisioned backhaul network. Note that the average UL throughput convergences more quickly, compared to the DL. This happens due to the asymmetry between the DL and UL traffic demand on the radio access network: the UL one is much lower, mainly due to the asymmetry between the transmission powers of BSs and UEs, as well as different file sizes assumed in each direction. Beyond this point, the UL backhaul resources will be underutilized. This calls for a *flexible* TDD duplexing scheme, that will dynamically distribute the backhaul resources accordingly, for example by giving more backhaul resources to DL when the UL demand is already satisfied (e.g. the eIMTA scheme [54]). Finally, in the wired case, star topology is always slightly better than the tree, whereas in the wireless the opposite, as explained earlier.

TABLE V
MEAN THROUGHP. FOR HANDED-OVER USERS (IN MBPS).

Topology	C_h = 50	250	500 (Mbps)
DL / UL thr.: Star-Wired	1.1 / 0.2	3.1 / 1.6	4.1 / X
DL / UL thr.: Tree-Wired	0.6 / 0.1	2.4 / 0.7	3.2 / X
DL / UL thr.: Tree-Wirel.	0.2 / 0.03	1.7 / 0.07	2.1 / 0.15
DL / UL thr.: Star-Wirel.	0.1 / 0.001	1.4 / 0.05	1.7 / 0.02

One could notice that user throughputs drop slightly on the C_h constraint, e.g. in a wired-star topology if C_h drops $500 \rightarrow 50$ Mbps (10 times), the mean user throughput only drops $15 \rightarrow 6$ Mbps (~ 3 times). This is due to the fact that, under-provisioned backhaul links do not affect the whole network, but specific groups of users associated with the cells that suffer from low backhaul capacity. To better illustrate this, in Table V we show the average throughput of the *handed-over users*, as a function of C_h. Indeed, their performance is severely affected: for the same scenario, their DL throughput drops all the way to 1.1 Mbps (~ 15 times). (In scenarios with no handovers, we mark the respective table entry with an X.)



(a) DL Spectral efficiency. (b) DL Load balancing (Utilization) Efficiency.

Fig. 7. Downlink Network Efficiencies (normalized).

Under-provisioning impact on Network Performance. Turning our attention to network-related performance, Fig. 7(a) considers spectral efficiency ($bit/s/Hz$), *normalized* by the *maximum* corresponding value when the network is provisioned. Load-balancing ("utilization") efficiency is further considered in Fig. 7(b) in terms of the MSE metric, described earlier. Both efficiencies converge to 1 as the network gets provisioned. Low C_h values will push users to handover to far-away BSs, and this will potentially decrease their $SINR$ (spectral efficiency decrease), and create steep differences between BSs loads, e.g. by congesting macro BSs and under-utilizing the SCs (load balancing decrease). Note that, joint degradation of these performances also impacts user performance negatively (e.g. throughput), as explained in B.7.

Note that regarding spectral efficiency (for the wireless scenario), there is a crucial point that judges which topology is better than the other in different under-provisioned cases (that is actually the point where the two curves meet). Obviously, this point, that is highly affected by the network topology and the given set of assumptions, can significantly vary between different performance metrics (e.g., in terms of load balancing and user throughput). Note that it is not possible to analytically derive this cross point, as this depends on the value of the objective function of the respective optimization problem. As a result, the actual cross point can only be found via sensitivity analysis. Nevertheless, we provide here the two main factors that affect this tradeoff between star and tree topology, to provide some qualitative insights. These factors are: (i) *path loss characteristics*: the higher the path loss (e.g. higher η) the worse the capacity loss on a long direct backhaul link required for a star topology, favoring multi-hop connectivity (i.e. a tree topology). (ii) *the number of hops*: the higher the number of backhaul hops the higher the total backhaul resources consumed "per BS" or "per bit of radio access", which might disfavor a tree topology.

TABLE VI
UL/DL SPLIT VS. JOINT-ASSOCIATION IMPROVEMENTS

Performance	τ = 0	τ = 0.5	τ = 1
DL / UL Throughput	6% / 32%	4% / 35%	0% / 37%
DL / UL Spectr. Eff.	4% / 29%	3% / 31%	0% / 33%
DL / UL Uiliz. Eff.	7% / 34%	4% / 38%	0% / 41%

Split UL/DL impact. As discussed earlier, while split is able to optimize the DL and UL performance, *simultaneously*, Joint UL/DL association is incapable of this; using $0 \leq \tau \leq 1$ we can trade-off which dimension is more important. Table VI illustrates the *performance improvements* that Split promises over the Joint UL/DL association, for various underprovisioned scenarios. We underline that split enhances the UL performance considerably, e.g. the average UL throughput is increased up to 37%. This is due to the *dependency* that Joint UL/DL generates between the DL and UL associations in the access network, that often makes the DL the bottleneck in the backhaul (due to aforementioned asymmetry between the peak access rates). Thus, DL will often "preempt" the backhaul constraint, and potentially (i) leave some UL resources unused, (ii) cause UL performance degradation.

Comparison with existing work. We now compare our proposed algorithm with two others user-association schemes; specifically, we run them in our considered network topology

for the wired backhaul case (similar behavior for the wireless). In Table VII we depict the performance decrease compared to our scheme in terms of: Spectral Efficiency (SE), User Performance (UP) (in terms of idle dedicated servers for dedicated traffic and throughput for best-effort) and Load Balancing (LB). We have assumed that all BSs associated with a BH link that is congested decrease their SE and UP proportionally to the amount of congestion by illustrating the effective corresponding performance.

Firstly, we investigate the algorithm proposed from Mesodiakaki et al. that focuses on dedicated traffic demand [29], [46]. This algorithm involves two stages, each of which independently considers access and backhaul performance: in the first, a subset of cells for each users' association is selected as candidates separately for DL and UL based on the radio access conditions (i.e., the DL data rate or the UL path loss without considering the BS load). Then, the best one among them is selected based on the BH conditions: *the BS with the fewest backhaul hops to the core is selected as optimal*. However, such a simplistic criterion can lead to rather suboptimal performance as explained in Section IV-A. Specifically, in our case, in both DL and UL, we see that effective SE and UP are decreased since the path with the fewest hops might include congested backhaul links (obviously, this is more intense in the tree topology where multiple BSs share the resources of a single backhaul link). LB is also slightly hurt (by also hurting UP) since the reference algorithm ignores the BS loads.

Secondly, we consider the user-association algorithm proposed by Domenico et al. for DL best-effort flows [30]. There, the ergodic capacity is attempted to be maximized through an iterative algorithm: at each step every user changes its association if the gain in terms of ergodic capacity is positive. Nevertheless, such an algorithm, strongly dependent on the initial condition and the corresponding step directions, does not necessarily converge to the best point by potentially getting stuck in subpar steps. Also, heterogeneous traffic demand and thus BS loads are not considered there. UL traffic and BH tree topology are not considered, so we (i) assume that UL associations are identical as DL, and (ii) extend the proposed resource allocation policy to evenly split the available link resources in tree topologies. Simulation results show significant performance degradation since in the (suboptimal) converged point (i) some users end up being attached to far-away BSs, and (ii) some BSs are driven to congestion while attempting to improve ergodic capacity by affecting LB and SE, correspondingly. Joint degradation of them also impacts UP negatively, as explained in (B.7).

TABLE VII
COMPARISON WITH EXISTING WORK.

	DL performance			UL performance		
Algorithm	SE	UP	LB	SE	UP	LB
[29], [46] Star-Wired	1.4	1.3	1.2	1.8	2.1	1.3
[29], [46] Tree-Wired	1.5	1.4	1.2	2.2	3.2	1.4
[30] Star-Wired	5.3	5.6	2.1	3.9	5.1	1.1
[30] Tree-Wired	5.7	5.9	2.4	4.1	5.4	1.2

VI. CONCLUSION

In this paper, we propose a user-association framework for future HetNets by investigating both (a) provisioned, and (b) underprovisioned backhaul network scenarios. We showed how traffic differentiation, different backhaul topologies and capacity limitations affect the user and network performance, with joint consideration of the access and backhaul resources. Initial simulation results corroborate the correctness of our framework, and reveal interesting tradeoffs for different network scenarios, as well as potential drawbacks of schemes operated in the backhaul, currently.

REFERENCES

[1] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, "Cell association and interference coordination in heterogeneous LTE-A cellular networks," *IEEE Journal on Selected Areas in Communications (JSAC)*, 2010.

[2] A. Khandekar, N. Bhushan, J. Tingfang, and V. Vanghi, "LTE-Advanced: Heterogeneous networks," in *Proc. European Wireless Conference*, 2010.

[3] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Communications*, 2011.

[4] T. Bonald and A. Proutiere, "Wireless downlink data channels: User performance and cell dimensioning," in *Proc. Mobile Computing and Networking (MobiCom)*, 2003.

[5] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An overview of load balancing in HetNets: Old myths and open problems," *IEEE Wireless Communications*, 2014.

[6] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing Quality of Service over a shared wireless link," *IEEE Communications Magazine*, 2001.

[7] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. Vehicular Technology Conference*, 2000.

[8] P. Hande, S. Patil, and H. Myung, "Distributed load-balancing in a multi-carrier wireless system," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, 2009.

[9] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed alpha-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Transactions on Networking*, 2012.

[10] H. Boostanimehr and V. Bhargava, "Unified and distributed QoS-driven cell association algorithms in heterogeneous networks," *IEEE Transactions on Wireless Communications*, 2015.

[11] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Transactions on Wireless Communications*, 2013.

[12] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *IEEE Journal on Selected Areas in Communications (JSAC)*, 2011.

[13] T. Han and N. Ansari, "Smart grid enabled mobile networks: Jointly optimizing BS operation and power distribution," in *Proc. IEEE International Conference on Communications*, 2014.

[14] J. Bartelt, A. Fehske, H. Klessig, G. Fettweis, and J. Voigt, "Joint bandwidth allocation and small cell switching in heterogeneous networks," in *Proc. IEEE Vehicular Technology Conference*, 2013.

[15] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Communication Surveys and Tutorials*, 2013.

[16] N. Sapountzis, S. Sarantidis, T. Spyropoulos, N. Nikaein, and U. Salim, "Reducing the energy consumption of small cell networks subject to QoE constraints," in *Proc. IEEE Globecom*, 2014.

[17] *Backhaul technologies for small cells*, Small Cell Forum, 2014.

[18] O. Tipmongkolsilp, S. Zaghloul, and A. Jukan, "The evolution of cellular backhaul technologies: Current issues and future trends," *in IEEE Communications Surveys and Tutorials*, 2011.

[19] Y. Wang and K. Pedersen, "Performance analysis of enhanced inter-cell interference coordination in LTE-Advanced heterogeneous networks," in *Vehicular Technology Conference (VTC Spring)*, 2012.

[20] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, "Algorithms for enhanced inter-cell interference coordination (eICIC) in lte hetnets," *IEEE/ACM Transasctions on Networking*, 2014.

[21] J. Lee, Y. Kim, H. Lee, B. L. Ng, D. Mazzarese, J. Liu, W. Xiao, and Y. Zhou, "Coordinated multipoint transmission and reception in LTE-advanced systems," *IEEE Communications Magazine*, 2012.

[22] J. Ghimire and C. Rosenberg, "Revisiting scheduling in heterogeneous networks when the backhaul is limited," *IEEE Journal on Selected Areas in Communications (JSAC)*, 2015.

[23] M. Shariat, E. Pateromichelakis, A. Quddus, and R. Tafazolli, "Joint TDD backhaul and access optimization in dense small cell networks," *IEEE Transactions on Vehicular Technology*, 2013.

[24] O. Somekh, O. Simeone, A. Sanderovich, B. Zaidel, and S. Shamai, "On the impact of limited-capacity backhaul and inter-users links in cooperative multicell networks," in *Proc. Conference Information Sciences and System (CISS)*, 2008.

[25] P. Rost, C. Bernardos, A. Domenico, M. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wubben, "Cloud technologies for flexible 5G radio access networks," *IEEE Communications Magazine*, 2014.

[26] N. Wang, E. Hossain, and V. K. Bhargava, "Joint downlink cell association and bandwidth allocation for wireless backhauling in two-tier hetnets with large-scale antenna arrays," *IEEE Transactions on Wireless Communications*, 2016.

[27] W. C. Liao, M. Hong, and Z. Q. Luo, "Max-min network flow and resource allocation for backhaul constrained heterogeneous wireless networks," in *Proc. IEEE ICASSP*, 2014.

[28] H. Beyranvand, W. Lim, M. Maier, C. Verikoukis, and J. A. Salehi, "Backhaul-aware user association in fiwi enhanced lte-a heterogeneous networks," *IEEE Transactions on Wireless Communications*, 2015.

[29] A. Mesodiakaki, F. Adelantado, L. Alonso, and C. Verikoukis, "Joint uplink and downlink cell selection in cognitive small cell heterogeneous networks," in *Proc. IEEE Globecom*, 2014.

[30] A. D. Domenico, V. Savin, and D. Ktenas, "A backhaul-aware cell selection algorithm for heterogeneous cellular networks," in *Proc. of IEEE PIMRC*, 2013.

[31] Z. Cui and R. Adve, "Joint user association and resource allocation in small cell networks with backhaul constraints," in *Proc. IEEE CISS*, 2014.

[32] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Cache-aware user association in backhaul-constrained small cell networks," in *in Proc. IEEE WiOpt*, 2014.

[33] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *in Proc. IEEE INFOCOM*, 2012.

[34] D. Chen, T. Quek, and M. Kountouris, "Backhauling in heterogeneous cellular networks: Modeling and tradeoffs," *IEEE Transactions on Wireless Communications*, 2015.

[35] G. T. . v.12.0.0 Rel.12, *Study on Small Cell enhancements for E-UTRA and E-UTRAN; Higher layer aspects*. Academic press, 2013.

[36] H. Kim, H. Y. Kim, Y. Cho, and S.-H. Lee, "Spectrum breathing and cell load balancing for self organizing wireless networks," in *Proc. IEEE Communications Workshops*, 2013.

[37] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems*. Imperial college press, 2010.

[38] H. Elshaer, F. Boccardi, M. Dohler, and R. Irmer, "Downlink and uplink decoupling: A disruptive architectural design for 5G networks," in *Proc. IEEE Globecom*, 2014.

[39] "http://cbnl.com/solutions-mobile-backhaul."

[40] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, "Under-provisioned backhaul: How capacity and topology impacts user and network-wide performance," Tech Report RR-16-311, Eurecom, 2016.

[41] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[42] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, "An analytical framework for optimal downlink-uplink user association in hetnets with traffic differentiation," in *Proc. IEEE Globecom*, 2015.

[43] G. 36.300, "Evolved universal terrestrial radio access (E-UTRA); further enhancements to LTE time division duplex (TDD) for downlink-uplink (DL-UL) interference management and traffic adaptation," 2012.

[44] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, "Optimal downlink and uplink user association in backhaul-limited hetnets," in *(to appear) IEEE Infocom*, 2016.

[45] Z. G. Raphael T. Haftka, *Elements of Structural Optimization*. Springer Netherlands, 1992.

[46] A. Mesodiakaki, F. Adelantado, L. Alonso, and C. Verikoukis, "Energy-efficient user association in cognitive heterogeneous networks," *IEEE Communications Magazine*, 2014.

[47] D. G. Luenberger, Y. Ye *et al.*, *Linear and nonlinear programming*. Springer, 1984, vol. 2.

[48] G. 24.312, "Access network discovery and selection function (andsf) management object (mo)," 2016.

[49] S. Sesia, I. Toufik, and B. M., *LTE - The UMTS Long Term Evolution: From Theory to Practice, 2nd Edition*. Wiley, 2011.

[50] I. S. 802.16m, "IEEE p802.16m-2007 draft standards for local and metropolitan area networks part 16: Air interface for fixed broadcast wireless access systems,," 2007.

[51] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proc. IEEE Infocom*, 2006.

[52] R. Srikant and L. Ying, *Communication networks: an optimization, control, and stochastic networks perspective*. Cambridge University Press, 2013.

[53] *3GPP, Technical Report LTE; Evolved Universal Terrestrial Radio Access (E-UTRA)*, TR 136 931, 2011.

[54] G. 36.828, "Evolved universal terrestrial radio access (E-UTRA) and radio access network (E-UTRAN); overall description," 2012.

Nikolaos Sapountzis received his Diploma in Electronic and Computer Engineering from the Technical University of Crete, Greece, and his Ph.D degree in Electrical Engineering from Eurecom/Telecom ParisTech, France. His main research interests are in modeling, optimization and performance analysis for mobile wireless networks.

Thrasyvoulos Spyropoulos received his Diploma in Electrical and Computer Engineering from the National Technical University of Athens, Greece, and his Ph.D degree in Electrical Engineering from the University of Southern California. He was a post-doctoral researcher at Inria and then, a senior researcher with the Swiss Federal Institute of Technology (ETH) Zurich. He is currently an assistant professor at Eurecom, France.

Navid Nikaein is an assistant professor in communication system department at Eurecom. He received his Ph.D. degree in communication systems from EPFL, Swiss Federal Institute of Technology, in 2003. Currently, he is leading a research group focusing on experimental system research related to wireless systems and networking. His research contributions are in the areas of wireless access layer techniques and protocols, flexible radio access and core networks, and wireless system prototyping and emulation/simulation platforms.

Umer Salim received his Ph.D. and M.S. degrees, specializing in communication theory and signal processing from Eurecom and Supelec, France, respectively. He is currently working at TCL Communications as 5G Systems Architect and 3GPP RAN1 delegate for 5G standardization activity. Before joining TCL, he worked at Intel Mobile Communications designing modems for high end smart phones and tablets. He has several years of research experience in physical layer of wireless communications.