Optimal Downlink and Uplink User Association in Backhaul-limited HetNets

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 only the radio access part, and ignore the backhaul topology and potential capacity limitations. Backhaul constraints are emerging as a key performance bottleneck in future heterogeneous networks, partly due to the continuous improvement of the radio interface, and partly due to the need for inexpensive backhaul links to reduce CAPEX/OPEX. To this end, we propose an analytical framework for user association that jointly considers radio access and backhaul performance. We derive an algorithm that takes into account spectral efficiency, base station load, backhaul link capacities and topology, and uplink and downlink traffic demand, and prove it converges to an optimal solution. We then use extensive simulations to study the impact of (i) backhaul capacity limitations and (ii) backhaul topology on key performance metrics.**

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-release 8. While this rule also maximizes the *instantaneous* rate of a user (i.e., the best 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 only consider the radio access network, namely the user rate on the radio interface and the load of BSs, 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 [15]. Multiple BS might 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) [16]. Furthermore, the increased backhaul signaling traffic required for Coordinated Multi-Point (CoMP) [17], as well as upcoming cloud-RAN (C-RAN) [18] technologies, are expected to further stress the backhaul network. 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 [19].

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) [19], [20], signaling overhead and performance tradeoffs for co-operative multi-point communication [21], Software-Defined-Networking (SDN)-based implementation flexibility [18], or

This work was supported by the project "Network-level Optimizations for Small Cell Networks", funded by Intel Mobile Communications (IMC).

propose some simple heuristics to include the impact on the backhaul of different association schemes [22]. Nevertheless, to our best knowledge, none of these works formally addresses the problem of optimal user association in future, backaul-limited HetNets.

To this end, in this paper we revisit the problem of optimal user association, 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 extend it to include backhaul constraints and topology.

(2) We analytically prove an optimal association rule for simple (e.g. star) and generic (tree) backhaul topologies, and propose an iterative algorithm based on penalty functions to converge to the optimal solution.

(3) We consider both uplink (UL) and downlink (DL) traffic characteristics, and show that our work fits well with future 5G network features like UL/DL split [23], and SDN-based implementations [18].

(4) Based on our framework, we investigate the impact of backhaul under-provisioning, in different topologies and system performance metrics. Our results also highlight some shortcomings of backhaul Layer 2 routing and suggest the need for Layer 3, joint radio access and backhaul routing.

The remainder of the paper is organized as follows: Section II describes the proposed analytical framework along with our system model assumptions, and derive the optimal user-association rules. We then sketch a practical implementation architecture, based on SDN, in Section III. In Section IV we simulate the optimal association policies and attempt to shed some light on the impact of backhaul constraints and topology. Section V discusses potential extensions of our framework, and Section VI concludes the paper.

II. User Association Problem

A. Model and Assumptions

In the following, we first describe our problem setup and assumptions. We will use a similar problem setup as the one used in a number of related works [9], [12], [24], [13], and extend it accordingly. To keep the presentation simplified, we present most notation and assumptions in terms of downlink (DL) traffic, denoted with a "D" sub/superscript. The assumptions for uplink (UL) traffic are in most cases symmetric, so one can simply replace "D" with "U" in the respective notation. Specific differences in the uplink traffic model will be elaborated, where necessary. In Table I, we summarize some useful notation we use throughout the paper.

(A.1 - BS coverage) 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.

(A.2 - Traffic Model) 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 $\lambda(x)$. A new flow can be either DL with probability z^D, or UL with $z^U = 1 - z^D$. Using a Poisson splitting argument [25], it follows that there are two *independent* Poisson arrival processes for DL and UL traffic, with respective rates $\lambda^D(x) = z^D\lambda(x)$ and $\lambda^U(x) = z^U\lambda(x)$. Flow sizes are independently and generically distributed with mean $1/\mu^D(x)$ (and $1/\mu^U(x)$ in the uplink.)

TABLE I. Notation

	Downlink	**Uplink**
Flow type sub/superscript	D	U
Traffic arrival rate (flows/sec) at location x	$\lambda^D(x)$	$\lambda^U(x)$
Mean flow size	$1/\mu^D(x)$	$1/\mu^U$
Maximum rate of the i-th BS at location x	$c_i^D(x)$	$c_i^U(x)$
Load density of the i-th BS at location x	$\rho_i^D(x)$	$\rho_i^U(x)$
BS i max rate requirement for backhaul	$\tilde{c}_i^D$	$\tilde{c}_i^U$
Utilization/Load of the i-th BS	$0 \le \rho_i^D \le 1$	$0 \le \rho_i^U \le 1$
Congestion indicator at BH link j	$\mathcal{I}^D(j)$	$\mathcal{I}^U(j)$
Capacity of backhaul link j	$C_h^D(j)$	$C_h^U(j)$
Association probability of location x to BS i	$p_i^D(x)$	$p_i^U(x)$

(A.3 - Physical Data Rate) Each BS $i \in \mathcal{B}$ is associated with a transmit power P_i and a total downlink bandwidth W_i^D. Based on this, BS i can deliver a *maximum* physical data transmission rate of $c_i^D(x)$ to a user at location x (in absense of any other users served), which is given by the Shannon capacity[1] $c_i^D(x) = W_i^D \log_2(1 + \text{SINR}_i(x))$, where

$$\text{SINR}_i(x) = \frac{G_i(x)P_i}{\sum_{j\neq i} G_j(x)P_j + N_0}. \quad (1)$$

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.)[2]. 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].

(A.4 - System load density) A *system load density* $\rho_i^D(x)$ at location x can be defined as

$$\rho_i^D(x) = \frac{\lambda^D(x)}{\mu^D(x)c_i^D(x)}. \quad (2)$$

(A.5 - BS load) Each location x is associated with association probabilities $p_i^D(x) \in [0, 1]$, which are the probabilities that a DL flow at location x gets associated with BS i. We can thus define the *total load* ρ_i^D of BS i as

$$\rho_i^D = \int_{\mathcal{L}} p_i^D(x)\rho_i^D(x)dx. \quad (3)$$

Similarly to [4], [9], we are interested in the *flow-level dynamics* of this system, and model the service of downlink flows at each BS as a queueing system with load ρ_i^D.

(A.6 - Scheduling) Proportionally fair scheduling is often implemented in 3G/4G networks, due to its good fairness and spectral efficiency properties [26]. This can be modeled as an M/G/1 multi-class processor sharing (PS) system (see,

[1] 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.

[2] 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.

e.g., [4], [9], [12]). It is multi-class, because each flow might get different rates for similarly allocated resources, due to different channel quality and modulation and coding scheme at x. Channel-based scheduling could also be included in the model and can be accounted for using a multiplicative factor in the average service rate [27].

(A.7 - Performance impact of BS load) Given the above scheduling, the stationary number of flows in BS i is known to be equal to $E[N_i] = \frac{\rho_i^D}{1-\rho_i^D}$ [25]. Hence, minimizing ρ_i^D minimizes $E[N_i]$, and by Little's law it also minimizes the per-flow delay for that base station [25]. At the same time, the throughput for a flow at location x is equal to $c_i^D(x)(1-\rho_i^D)$. This observation is important to understand how the user's physical data rate $c_i^D(x)$ (related to users at location x only) and the BS load ρ_i^D (related to *all* users associated with BS i) affect the optimal user association decision (e.g. in Eq. (7)).

(A.8 - Backhaul topology) Each BS is connected to the core network through the eNB aggregation gateway either directly ("star" topology) or through one or more SC aggregation gateways ("tree" topology). Fig. 1 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 (e.g., wireless) connecting SCs to the eNB, denoted as $\mathcal{B}_h$. We denote as routing path $\mathcal{B}_h(i)$ 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. 1, $\mathcal{B}_h(1) = \{1\}$, 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. 1. In the case of a star topology, there is exactly one (unique) backhaul link used for each BS (i.e., $\|\mathcal{B}_h(i)\| = \|\mathcal{B}(j)\| = 1, \forall i,j$). Finally, 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 IV, we highlight some limitations of L2 backhaul routing.

(A.9 - Backhaul load) Each backhaul link $j \in \mathcal{B}_h$ is characterized by a downlink capacity $C_h^D(j)$ bps. Backhaul links usually don't implement any particular scheduling algorithm, and can be seen as a data "pipe". The capacity on the UL and DL might be the same or different (e.g., Frequency-Division Duplex (FDD), or fixed/dynamic Time-Division Duplex (TDD) systems [28]). The load on a backhaul link $j \in \mathcal{B}_h$ consists of the sum of loads of all BSs using that link:

$$\sum_{i \in \mathcal{B}(j)} \rho_i^D \tilde{c}_i^D, \tag{4}$$

where $\tilde{c}_i^D$ is an estimate of the total rate delivered by BS i. A BS is usually characterized by its "peak" rate (often upper bounded by the maximum MCS available), and a "busy" rate, when a BS serves many users [15]. The latter is usually quite smaller than the former, since users near the edge of the cell tend to bring the average rate down. However, the use of channel-based scheduling and related multi-user diversity gains suggest that conservatively setting $\tilde{c}_i^D$ closer to its nominal peak value is safer. In practice, a BS could measure this load and use it directly.

Based on the above problem setup, the association policy consists in finding appropriate values for the routing probabilities $p_i^D(x)$ and $p_i^U(x)$, for DL and UL traffic, respectively

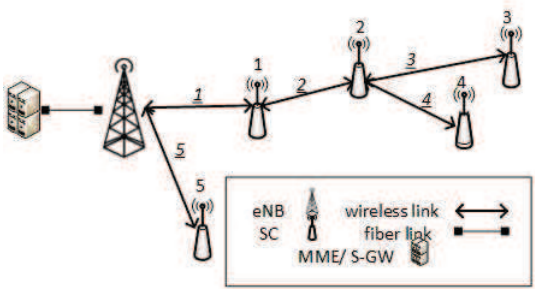


Fig. 1. Future HetNet topology.

(defined earlier in A.5). That is, for each location x, we would like to optimally choose which BS i to route to flows generated from (UL) or destined to (DL) users in x[3]. Our goals for this association problem are twofold: (i) ensure that the capacity of no network element (BS or backhaul link) is exceeded; (ii) achieve a good tradeoff between user physical data rates and load balancing. We will consider two main scenarios:

Link split or DL/UL decoupling: This allows each UE to be associated with different BSs for its DL and UL traffic, and to optimize UL and DL performance independently [29], [23].

Joint DL/UL: In current networks, a UE must be associated with the same BS for both UL and DL traffic.

B. Optimal User Association for Split UL/DL

We will first define the user association problem for the split DL/UL case. The feasible region for the variables $p_i^D(x), p_i^U(x)$ can first be delimited by the requirement that the capacity of no BS is exceeded.

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

$$\begin{aligned} f^y = \Big\{ \rho^{\mathbf{y}} \mid \rho_i^y = & \int_{\mathcal{L}} p_i^y(x)\rho_i^y(x)dx, \\ & 0 \le \rho_i^y \le 1-\epsilon, \\ & \sum_{i \in \mathcal{B}} p_i^y(x) = 1, \\ & 0 \le p_i^y(x) \le 1, \forall i \in \mathcal{B}, \forall x \in \mathcal{L} \Big\}. \end{aligned} \tag{5}$$

Lemma 2.1. *The feasible sets f^D, f^U are convex.*

Proof: The proof for the feasible DL set f^D is presented in [9]. It can be easily adapted for the UL case, as well. ■

When UL and DL traffic can be routed separately, this implies that $p_i^D(x)$ and $p_i^U(x)$ can take different values. Hence, the problem of optimal DL and UL association can be decoupled into two independent problems, one for DL and one for UL. In the remainder of this section, we focus on the optimal DL association problem, and *we omit the superscripts $\{D,U\}$ to simplify notation*. We return to the joint DL/UL association problem in the next section. To better illustrate our

[3]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, i.e. the optimal association rule will be deterministic.

approach, we first apply this for a simple star BH topology, and then generalize for a tree BH topology.

Optimal User Association for Star BH Topology)

Let $\mathcal{I}(j)$ be an indicator variable, related to backhaul link $j \in \mathcal{B}_h$, such that $\mathcal{I}(j) = 0$ when $\frac{\rho_i \tilde{c}_i}{C_h(j)} < 1$, and $\mathcal{I}(j) = 1$ when $\frac{\rho_i \tilde{c}_i}{C_h(j)} \geq 1$ (i.e, the offered load to backhaul link j exceeds the available capacity). In the following, since for star topologies there is exactly one backhaul link (j) associated with each BS (i), to simplify notation we can safely assume $i = j$.

Theorem 2.2 (User-Association in a star BH topology)**.** *The optimal user association problem with a star backhaul topology is expressed as* $\min_\rho \left\{\Phi(\rho)|\rho \in f\right\}$*, where*

$$\Phi(\rho) = \sum_{i\in\mathcal{B}} \frac{(1-\rho_i)^{1-\alpha}}{\alpha - 1} + \gamma \sum_{i\in\mathcal{B}_h} \mathcal{I}(i)\left(\frac{\rho_i \tilde{c}_i}{C_h(i)} - 1\right)^2. \quad (6)$$

If the feasible domain f of the problem is non-empty, and $\rho^* = (\rho_1^*, \rho_2^*, \cdots, \rho_{\|\mathcal{B}\|}^*)$ denotes the optimal load vector, the user-association rule at location x is

$$\arg\max_{i\in\mathcal{B}} \frac{c_i(x)(1-\rho_i^*)^\alpha}{1 + 2\gamma\cdot(1-\rho_i^*)^\alpha \cdot \tilde{c}_i \cdot \frac{\mathcal{I}(i)}{C_h(i)} \cdot \left(\frac{\rho_i^* \tilde{c}_i}{C_h(i)} - 1\right)}. \quad (7)$$

Proof: We prove here that the above association rule indeed minimizes the cost function of Eq. (6). This problem is a convex optimization problem. Its feasible set f is convex, and the objective function $\Phi(\rho)$ is also convex (the hessian matrix is positive semi-definite). Let ρ^* be the optimal solution of this minimization problem. Hence, it is adequate to check the following condition for optimality

$$\langle \nabla\Phi(\rho^*), \Delta\rho^* \rangle \geq 0 \quad (8)$$

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 (7), the optimal association rule is given by:

$$p_i^*(x) = \mathbf{1}\left\{i = \arg\max_{i\in\mathcal{B}} \frac{c_i(x)(1-\rho_i^*)^\alpha}{1 + 2\gamma\cdot(1-\rho_i^*)^\alpha \cdot \tilde{c}_i \cdot \frac{\mathcal{I}(i)}{C_h(i)} \cdot \left(\frac{\rho_i^* \tilde{c}_i}{C_h(i)} - 1\right)}\right\}. \quad (9)$$

Before proceeding to the calculation of the inner product, we analytically calculate the derivative of the corresponding cost function $\Phi(\rho)$, described in Eq. (6). The derivative is an i-th dimensional vector; the i-th element of which has value:

$$\nabla\Phi(\rho_i) = \begin{cases} (1-\rho_i)^{-\alpha}, & \text{if } \rho_i\tilde{c}_i \leq C_h(i) \\ (1-\rho_i)^{-\alpha} + \gamma\mathcal{I}(i)\frac{2\rho_i\tilde{c}_i^2 - 2\tilde{c}_i C_h(i)}{C_h(i)^2}, & \text{if } \rho_i\tilde{c}_i \geq C_h(i). \end{cases} \quad (10)$$

To that end, the inner product defined in Eq. (8), becomes:

$$\begin{aligned}
\langle \nabla\Phi(\rho^*), \Delta\rho^* \rangle &= \sum_{i\in\mathcal{B}} \left\{\frac{1}{(1-\rho_i^*)^\alpha} + \gamma\mathcal{I}(i)\frac{2\rho_i^*\tilde{c}_i^2 - 2\tilde{c}_i C_h(i)}{C_h(i)^2}\right\}(\rho_i - \rho_i^*) \\
&= \sum_{i\in\mathcal{B}} \frac{1 + 2\gamma\mathcal{I}(i)(1-\rho_i^*)^\alpha \frac{(\rho_i^*\tilde{c}_i^2 - \tilde{c}_i C_h(i))}{C_h(i)^2}}{(1-\rho_i^*)^\alpha} \int_{\mathcal{L}} \rho_i(x)\left(p_i(x) - p_i^*(x)\right)dx \\
&= \int_L \frac{\lambda(x)}{\mu(x)} \sum_{i\in\mathcal{B}} \left(\frac{1 + 2\gamma(1-\rho_i^*)^\alpha \tilde{c}_i \frac{\mathcal{I}(i)}{C_h(i)}\left(\frac{\rho_i^*\tilde{c}_i}{C_h(i)} - 1\right)}{c_i(x)(1-\rho_i^*)^\alpha}\right)\left(p_i(x) - p_i^*(x)\right)dx.
\end{aligned}$$

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 (9) is an indicator for the maximizer of $\frac{c_i(x)(1-\rho_i^*)^\alpha}{1+2\gamma\cdot(1-\rho_i^*)^\alpha\cdot\tilde{c}_i\cdot\frac{\mathcal{I}(i)}{C_h(i)}\cdot\left(\frac{\rho_i^*\tilde{c}_i}{C_h(i)}-1\right)}$. Hence (8) holds. ■

We expressed the objective (Eq. (6)) with respect to the variables ρ_i, for convenience. However, these depend on the association probabilities $p_i(x)$, which are the underlying decision variables, as shown in Definition 1. The first sum is the standard α-cost function for each BS i [9]. Parameter α controls the amount of load balancing desired. For $\alpha = 0$, minimizing this function leads to a maximum SINR user-association rule, maximizing the physical data rate for each location $c_i(x)$, and thus the spectral efficiency. As $\alpha \to \infty$, this cost function aims at equalizing the BS utilizations ρ_i, i.e. to balance the loads[4]. The second sum introduces a penalty for each backhaul link i whose capacity is exceeded ($\mathcal{I}(i) = 1$). This penalty function is quadratic on the amount of excess load (quadratic penalty functions are often considered in convex optimization literature [30]). γ could be chosen as a large constant, introducing a "soft" constraint for the backhaul links (i.e., backhaul capacity could be slightly exceeded, if this really improves access performance), or be iteratively adapted using increasing values, so as to converge to a "hard" constraint.

Regarding the optimal association rule of Eq. (7), 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]. 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. Let $\mathcal{I}(j)$ be again an indicator variable, related to congestion in backhaul link $j \in \mathcal{B}_h$. Now, $\mathcal{I}(j)$ needs to consider the load of all the BSs whose traffic it carries (see A.9): specifically, $\mathcal{I}(j) = 0$, when $\frac{\sum_{i\in\mathcal{B}(j)}\rho_i\tilde{c}_i}{C_h(j)} < 1$ and $\mathcal{I}(j) = 1$, when $\frac{\sum_{i\in\mathcal{B}(j)}\rho_i\tilde{c}_i}{C_h(j)} \geq 1$.

Theorem 2.3. *[User-Association in a tree BH topology] The optimal user association problem with a tree backhaul topology can be expressed as* $\min_\rho \left\{\Phi(\rho)|\rho \in f\right\}$*, where*

$$\Phi(\rho) = \sum_{i\in\mathcal{B}} \frac{(1-\rho_i)^{1-\alpha}}{\alpha - 1} + \gamma \sum_{j\in\mathcal{B}_h} \mathcal{I}(j)\left(\frac{\sum_{i\in\mathcal{B}(j)}\rho_i\tilde{c}_i}{C_h(j)} - 1\right)^2. \quad (11)$$

[4]Note that for $\alpha = 1$ the above α-cost function is not defined, and $\log(1-\rho_i)^{-1}$ is used instead [9].

If the feasible domain f of the problem is non-empty, the optimal user-association rule at location x is now

$$\arg\max_{i\in\mathcal{B}} \frac{c_i(x)(1-\rho_i^*)^{\alpha}}{1+2\gamma\cdot(1-\rho_i^*)^{\alpha}\cdot\tilde{c}_i\sum\limits_{j\in\mathcal{B}_h(i)}\frac{\mathcal{I}(j)}{C_h(j)}\cdot\left(\frac{\sum_{k\in\mathcal{B}(j)}\rho_k^*\tilde{c}_k}{C_h(j)}-1\right)}. \tag{12}$$

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

$$\begin{aligned}
&\langle\nabla\Phi(\rho^*),\Delta\rho^*\rangle = \\
&= \sum_{i\in\mathcal{B}}\Big\{\frac{1}{(1-\rho_i^*)^{\alpha}}+2\gamma\sum_{j\in\mathcal{B}_h(i)}\mathcal{I}(j)\Big[\frac{\sum_{k\in\mathcal{B}(j)}\rho_k^*\tilde{c}_k}{C_h(j)^2}\tilde{c}_i-\frac{\tilde{c}_i}{C_h(j)}\Big]\Big\}(\rho_i-\rho_i^*) \\
&\quad\cdot\int_{\mathcal{L}}\rho_i(x)\big(p_i(x)-p_i^*(x)\big)\,dx = \\
&= \int_L\frac{\lambda(x)}{\mu(x)}\sum_{i\in\mathcal{B}}\left(\frac{1+2\gamma(1-\rho_i^*)^{\alpha}\tilde{c}_i\sum\limits_{j\in\mathcal{B}_h(i)}\frac{\mathcal{I}(j)}{C_h(j)}\cdot\left(\frac{\sum_{k\in\mathcal{B}(j)}\rho_k^*\tilde{c}_k}{C_h(j)}-1\right)}{c_i(x)(1-\rho_i^*)^{\alpha}}\right)\cdot \\
&\quad\cdot\big(p_i(x)-p_i^*(x)\big)\,dx\geq 0,
\end{aligned} \tag{13}$$

due to the corresponding maximizer $p_i^*(x)$ derived from (12). ■

As one can see, the cost function is similar in nature. The first term corresponding to the radio access part remains unchanged. The second term again introduces a penalty for each backhaul link that is congested. However, there are a number of interesting differences between the star and tree cases. First, the penalty term in the denominator of the optimal association rule (Eq. (12)) 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 backhaul hops heuristic proposed in [31], [22]. 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 backhaul 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.

C. Optimal Joint UL and DL Association

Current cellular networks (e.g. 3G/4G) require that a UE should be connected to a single BS for both UL and DL traffic [32]. This changes the optimal association problem, as one now needs to *jointly* optimize UL and DL traffic performance. E.g., a user at location x might end up being associated with a BS offering suboptimal performance on both the downlink and uplink, because other BS candidates offer really bad UL (or really bad DL) performance.

We thus need to modify our framework accordingly. First, while deriving the association rules we will have to require $p_i^D(x) = p_i^U(x)\ \forall i\in\mathcal{B}$. Second, UL and DL performance must now be included in the same cost function. Specifically, in the part of the cost function corresponding to the radio access, the operator may weigh the importance of DL and UL traffic performance with a parameter $\tau\in[0,1]$[5]. If $\rho=[\rho^D;\rho^U]$ with corresponding feasible convex set $\mathcal{F}$[6], our objective now is

$$\phi(\rho)=\sum_{i\in\mathcal{B}}\tau\frac{(1-\rho_i^D)^{1-\alpha^D}}{\alpha^D-1}+(1-\tau)\frac{(1-\rho_i^U)^{1-\alpha^U}}{\alpha^U-1},\text{if }\alpha^D,\alpha^U\neq 1. \tag{14}$$

We also need to extend the penalty function to consider both uplink and downlink capacity being exceeded on the backhaul link. 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 2.4 (Joint UL/DL Association)**.** *The optimal association problem with a generic BH topology can be expressed as* $\min_\rho\left\{\Phi(\rho)|\rho=[\rho^{\mathbf{D}};\rho^{\mathbf{U}}]\in\mathcal{F}\right\}$*, where*

$$\Phi(\rho)=\phi(\rho)+\gamma\sum_{k\in\{D,U\}}\sum_{j\in\mathcal{B}_h}\mathcal{I}^k(j)\left(\frac{\sum\limits_{i\in\mathcal{B}(j)}\rho_i^k\tilde{c}_i^k}{C_h^k(j)}-1\right)^2. \tag{15}$$

If the feasible domain $\mathcal{F}$ of the problem is non-empty, the optimal user-association rule at location x is

$$i(x)=\arg\max_{i\in\mathcal{B}}\frac{\left(1-\rho_i^{*D}\right)^{\alpha^D}\cdot\left(1-\rho_i^{*U}\right)^{\alpha^U}}{e^D(x)\cdot\left(1-\rho_i^{*U}\right)^{\alpha^U}+e^U(x)\cdot\left(1-\rho_i^{*D}\right)^{\alpha^D}}, \tag{16}$$

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

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

Proof: The proof follows similar steps as for the split scenario, but is more involved and is omitted due to space limitations. We refer the interested reader to [33]. ■

The penalty function for the backhaul network is simply the sum of the respective penalty functions for UL and DL, described in Theorem 2.3. However, despite the similarities of the cost functions, as we can see, the resulting association policy in the joint UL/DL case is more complex. The main insights are captured in the following remark.

Remark: The above optimal rule suggests that, when jointly considering the potentially conflicting objectives of optimizing both DL and UL performance, it is optimal to associate a user with the BS that maximizes the *harmonic mean* of the

[5]If α^D or α^U is equal to 1, the respective fraction must again be replaced with $\log(1-\rho_i)$, as explained earlier.

[6]Due to space limitations we skip the analytical definition and proof of convexity for $\mathcal{F}$; we refer the interested reader to [33].

individual association rules, when considering each objective alone. Maximizing the harmonic mean presents a more "fair" way to weigh in different objectives. E.g., 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 with the highest sum (or arithmetic average) of rates should be chosen (i.e. BS B). However, this would lead to rather poor UL performance. Maximizing the harmonic mean would lead to choosing BS C instead. While this simple example captures the main principle, the actual rule is more complex, as it weighs each objective also with a complex factor $e^l(x)$ related to both radio access performance and backhaul penalties. Finally, we note, for comparison purposes, that in the case of "split" UL/DL split association, covered in Section II-B, DL traffic would be associated with BS B, and UL traffic with BS C. This simple example demonstrates why split UL/DL may perform considerably better than the joint association. We will further explore this in the simulations.

III. SDN-BASED IMPLEMENTATION

The above derived association rules tell a UE at some location x, where to associate optimally. However, as the BS loads might not be optimal at the time (i.e. equal to ρ_i^*, see proof of Theorem 2.2), this policy represents a gradient descent algorithm, that needs to be iteratively applied in practice, until it converges to the optimal loads. Here, we describe an implementation of such an algorithm facilitated by an SDN framework that offers a centralized programmable control for the underlying network. It takes as inputs (i) the overall network status, and (ii) some high level system-parameters (e.g. operator preferences). According to the SDN architecture, we consider four planes, as illustrated in Fig. 2:

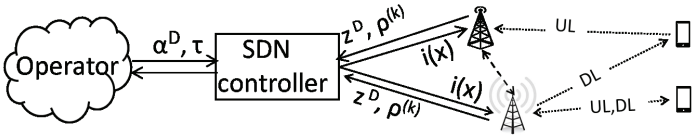


Fig. 2. Applicability to the SDN architecture.

Application tier: The operator determines and advertises to the controller some system-related parameters (e.g. α^D, τ etc.).
Controller tier: At each iteration period k, the controller receives some network-related parameters (e.g. z^U, traffic profile, etc.) as well as the 2-dimensional load vector ρ^D, ρ^U from the network tier. Then, based on the system-related parameters directed from the application tier, it determines and advertises to BSs the optimal associations (Eq. 7,12,16)[7].
Network tier: Each k-th period, BSs either apply or indicate to users the optimal rules depending on how the association is managed in the network. At the end of k, they measure and advertise to the controller their average load levels, and the network-relater parameters.

[7]The controller also handles the penalty factor γ. For "hard" backhaul constraints, it (i) starts with a small value, (ii) increases its value at each iteration, according to the magnitude of the main cost function, to avoid infeasible solutions and steep valleys, as is usually done in penalty function implementations [30].

User tier: At each k-th period, a UE at location x is associated or triggers the association procedure to the new BSs.

This iteration converges to the globally optimal point, requiring a simple modification to the proof found [9].

IV. SIMULATIONS

In this section we present some numerical results and discuss underlying insights. We consider a 2×2 km^2 area. Fig. 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 (shown with asterisks and numbered from 1-2) and eight SCs (shown with triangles and numbered from 3-10) as depicted in Fig. 3(b)-3(f). The ratio of the DL flows is z^D=0.7, and the average length sizes $1/\mu^D(x) = 100KBytes$, $1/\mu^U(x) = 20KBytes$, $\forall x \in \mathcal{L}$. We also assume that the maximum transmission powers of eNB, SC and UE are $43, 24$ and 18 dBm, whereas the access-network bandwidth $W_i^D = W_i^U = 10MHz$, $\forall i \in \mathcal{B}$ and the noise power density $N_0 = -174dBm/Hz$ [34]. We finally assume $\alpha^D = \alpha^U = 1$ (throughput-optimal values). If not explicitly mentioned we consider the split scheme (Section II-B).

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 (A.8), 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*[8]. Our aim is to evaluate the derived association rules described in Section II for different *under-provisioned* scenarios, by assuming *fixed* backhaul routing paths, pre-established with traditional Layer 2 routing. We assume that the BH capacities on the DL and UL are the same (i.e. $C_h^D(i) = C_h^U(i) = C_h$, $\forall i \in \mathcal{B}$), 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 as well (Assumption A.9).

Before proceeding, we discuss how different backhaul technologies affect the backhaul capacities, and setup a metric to evaluate the utilization efficiency. 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:

$$d(r_i) = \begin{cases} 1, & r_i \leq r_0 \\ (\frac{r_0}{r_i})^n, & \text{otherwise,} \end{cases} \quad (17)$$

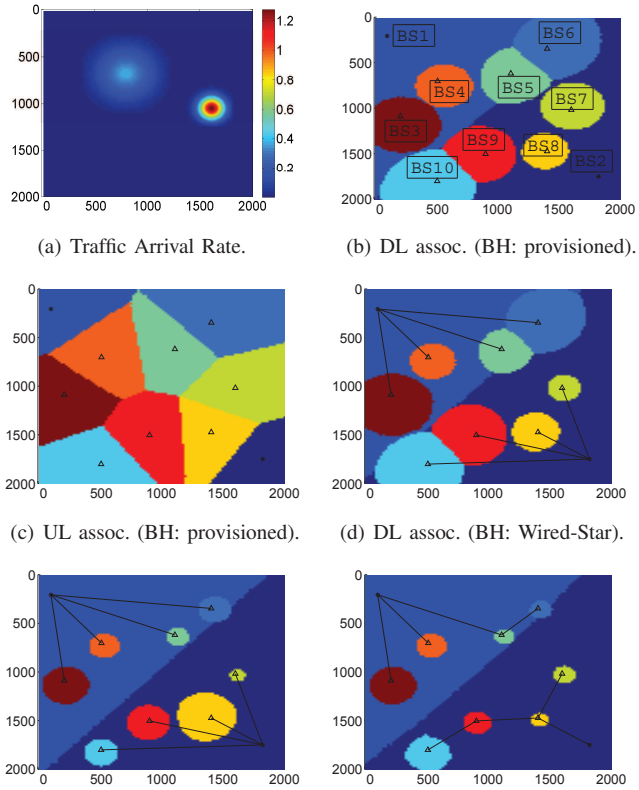
where r_0 is some threshold range within which the maximal rate is obtained (e.g. Line-of-Sight), and n is the attenuation factor. Hence, the available capacity drops to $d(r_i)C_h(j)$ ($\leq C_h(j)$). For our simulations, we assumed that $r_0 = 200m$, and $n = 3$. While the above model is perhaps oversimplifying, our main goal is to simply include the propagation related impact on wireless backhaul, compared to wired, without

[8]Note that copper and fiber access are the key technologies for wired backhaul links, and microWave and millimeter-wave P2P or P2MP access are the counterpart for the wireless backhaul links [35].

getting into the details of specific backhaul implementations. Furthermore, to evaluate the DL utilization efficiency we introduce the Mean Squared Error (MSE^D), between the DL utilization of different BSs, normalized to 1:

$$\text{MSE}^D = \frac{1}{2 * h} \sum_i \sum_j (\rho_i^D - \rho_j^D)^2, \quad (18)$$

where $h = \left\lfloor \frac{N}{2} \right\rfloor \times \left\lceil \frac{N}{2} \right\rceil$ is the normalizing factor, and N the total number of BSs (similar MSE^U for UL). We define the DL/UL utilization efficiency to be $1 - MSE^D$ and $1 - MSE^U$, respectively, that increase on the amount of load balancing[9].



(a) Traffic Arrival Rate. (b) DL assoc. (BH: provisioned).

(c) UL assoc. (BH: provisioned). (d) DL assoc. (BH: Wired-Star).

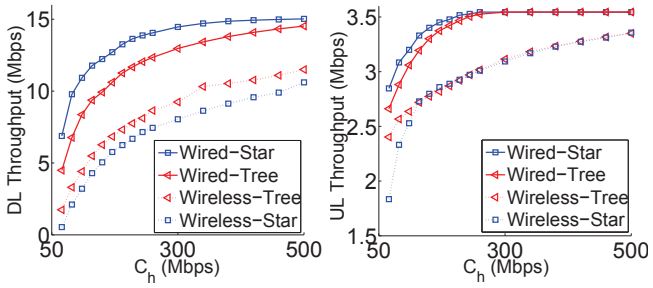
(e) DL assoc. (BH: Wireless-Star). (f) DL assoc. (BH: Wireless-Tree).

Fig. 3. Coverage snapshots with optimal associations in different scenarios.

Coverage Snapshots. Fig. 3(b)-3(c) depict the optimal DL and UL associations, with respect to the traffic arrival rates shown in Fig. 3(a), by ignoring the backhaul network (or assuming it's over-provisioned). In the DL, most users are attached to the macro BSs due to their high transmission power, whereas in the UL each user is mainly attached to the nearest BS [22]. 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 [33] for them). In Fig. 3(d) 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 (i.e. BS4 and BS7 have vastly decreased their coverage areas) due to the higher traffic demand. Similarly, in Fig. 3(e), 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.(17)).

In Fig. 3(f) 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. 3(e)). 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 (see Section V).



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

Fig. 4. Mean throughputs overall all users in the network.

User performance. Fig. 4(a), 4(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 throughputs 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 (A.3). 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,

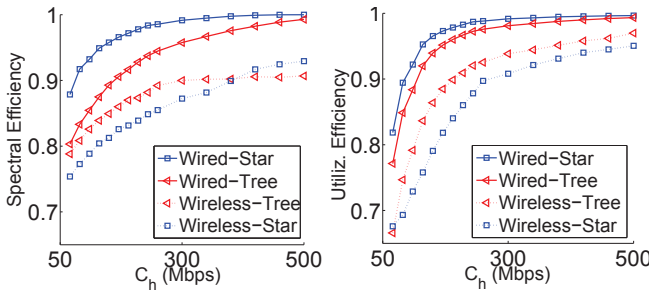
[9]We should note that different load balancing metrics could have been used, e.g. the *maximum, median and minimum* BS load; however, we chose to use MSE since it facilitates the visualization of the network efficiency.

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 [36]). Finally, in the wired backhaul links case, star topology is always slightly better than the tree, whereas in the wireless the opposite, as explained earlier.

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

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

(b) DL Utilization efficiency (normalized).

Fig. 5. Downlink Network Efficiencies.

Network Performance. Turning our attention to network-related performance, Fig. 5(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. 5(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 (utilization efficiency decrease). Note that, the joint degradation of these performances also impacts user performance negatively (e.g. user throughput), as explained in Assumption (A.7). Regarding spectral efficiency, more specifically, although in the wired scenario, star topology is always better compared to the tree, in the wireless scenario this is not the case. For low values of C_h, the star topology is worse, due to the higher capacity loss of the long and direct links. However, as C_h is increased, and some links start becoming provisioned in the star topology, the capacity loss cost due to the long wireless links in the star topology, is dominated from the capacity loss cost due to multi-hop sharing links of the tree topology, by making tree a worse choice. We highlight that this trade-off can suggest different topologies as optimal in different under-provisioned scenarios, and can affect different performance metrics.

TABLE III. SPLIT VS. NON-SPLIT IMPROVEMENTS

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

Split and non-split impact. As discussed earlier, UL/DL split (described in Section II-B) is able to optimize the DL and UL performance, *simultaneously*. Joint UL/DL association or "non-split" (Section II-C) is incapable of these parallel optimizations; however, using $0 \leq \tau \leq 1$ we can trade-off which dimension carries more importance. One would ask, what's the enhancement that split offers, given a non-split scenario with parameter τ? Table III illustrates the *performance improvements* that split promises over the non-split, in terms of various metrics, for various τ. Indeed, the higher the τ, the higher the emphasis on the DL (and less on the UL), and so the higher the gain over the UL performance that split guarantees (the inverse also holds for low τ). We remark 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 non-split 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.

V. DISCUSSION AND FUTURE WORK

In this section, we briefly discuss potential limitations of our framework, and how to possibly extend it to address them.

Additional flow-types. In our framework, we assumed all flows to be best-effort. Modern cellular networks will need to also consider *dedicated* flows that are subject to admission control, i.e., require resources for exclusive usage [26]. The user QoS related to such flows is often captured with a blocking probability, which could captured by a k-loss queueing system [25], [37]. The blocking probability in such a system again depends on the channel quality to x (since this decides how many resources must be allocated to satisfy a given performance requirement) and the load of that BS (this decides the total resources remaining unused), as we showed in [38]. Hence, one could introduce an additional term in the objective related to dedicated flow performance, and attempt to derive an optimal policy that takes both best effort and dedicated flows, as well as related access and backhaul resources into account.

Dynamic TDD scheme. In our simulations we assume that the backhaul resources are fixed, and equally distributed between the downlink and uplink (Assumption A.9). Interestingly, we showed that this scheme can result in rather suboptimal performance, and waste backhaul resources. The

design of a more flexible TDD scheme, that distributes *dynamically* the backhaul resources between the downlink and uplink dimension, can enhance the system performance.

Joint radio and L3 backhaul routing. Mesh backhaul topologies with multiple available routing paths are expected to be the rule, rather than the exception in future networks. Our assumption of fixed, L2 backhaul routing is restrictive, and as we saw in the simulations also penalizes performance. It would be interesting to consider choosing both the BS to associate to, as well as how traffic from this BS is routed towards an aggregation point (L3 routing).

VI. Conclusion

In this paper, we propose a user-association framework for future backhaul-limited HetNets. We showed how different backhaul topologies and capacity limitations affect the user and network performance, with joint consideration of the access/ backhaul resources. Initial simulation results corroborate the correctness of our framework, and reveal interesting tradeoffs for different under-provisioned networks, 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] *Backhaul technologies for small cells*, Small Cell Forum, 2014.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] T. Bonald, S. Borst, N. Hegde, M. Jonckheere, and A. Proutiere, "Flow-level performance and capacity of wireless networks with user mobility," 2009.

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

[29] 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.

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

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

[32] 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.

[33] 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.

[34] *3GPP, Technical Report LTE; Evolved Universal Terrestrial Radio Access (E-UTRA)*, TR 136 931, 2011.

[35] Alcatel-Lucet, "Mobile bakhaul architecture for hetnet, https://www.alcatel-lucent.com/solutions/mobile-backhaul," 2015.

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

[37] 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.

[38] 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.