User Association for Ultra Dense Networks with QoS Guarantees

Nikolaos Liakopoulos[1,2], Georgios Paschos[1], Thrasyvoulos Spyropoulos[2]

[1]Mathematical and Algorithmic Sciences Lab, FRC, Huawei Technologies SASU, email: firstname.lastname@huawei.com
[2]EURECOM, Sophia-Antipolis France, email: spyropou@eurecom.fr

***Abstract*—We study the problem of user association in Ultra Dense Networks (UDNs) for two network services; one requiring QoS guarantees to VIP flows, and one best effort service. The goal is to take advantage of statistical multiplexing in order to optimize the use of resources, while ensuring that the VIP flows enjoy active performance guarantees. We formulate this as an optimization problem, show that the problem is convex, and finally demonstrate that the optimum point can in fact be realized by distributed user association rules. In this way, our framework uses the fundamental problem of user association to show that both *isolation* and *statistical multiplexing* can be achieved in the context of UDNs, when different services or slices must share BS resources, as envisioned in 5G networks. We demonstrate no violations of the VIP flow constraint on real data traces for mobile network traffic, while a baseline best effort distributed policy applied to this setup inflicts up to 46.5% violations.**

I. Introduction

In order to cope with the rapid growth of data traffic demand, wireless operators move to ultra dense, heterogeneous deployments, also referred to as Ultra Dense Networks (UDNs) [1]. These consist of many low-power small cells, to maximize spatial reuse of the available bandwidth, overlaid with macro-cells which ensure coverage. With ultra dense deployment, the problem of user association becomes increasingly important, but also highly complex. Naive SINR-based schemes [2], [3], like the ones commonly used in current wireless networks, can be highly suboptimal, failing to properly balance the load across the different base stations which is necessary to translate the denser Base Station (BS) deployment into actual Quality of Service (QoS) improvement.

Furthermore, new applications and service types are emerging that need to be carried over cellular networks with diverse QoS requirements. These, intensify the need for efficient resource utilization to cope with the conflicting demands of reduced latency or increased throughput [4]. These prerequisites can be effectively treated by breaking the one-type-fits-all-services scheme and considering application specific (flow) traffic steering and prioritizing, as envisioned in 5G New Radio (NR) [4]. However, it is far from clear how to optimally allocate the common resources between different services.

Consider for example a scenario with a "VIP" and a "Best Effort" (BE) service. The former requires low delay for its applications, while the latter could correspond to standard data and voice. Naively pre-allocating specific resources, e.g., part of the bandwidth or Resource Blocks (in LTE) to each service is suboptimal [5], as part of the resources might stay underutilized, while others might become congested. On the other hand, while joint resource allocation leads to better statistical multiplexing gains, a surge of BE traffic may endanger the performance of the VIP service. *To address the challenge of isolation vs statistical multiplexing, this paper considers the basic scenario with one VIP service and one Best Effort (BE). We propose a few modifications on a simple and well-established model for BS schedulers [6], and develop an analytical optimization framework with the following goals:*

1) Provide *service isolation*, i.e., Best Effort load should not have an impact on VIP performance.
2) Provide *performance guarantees* to the VIP flows, in terms of bounding the mean number of active VIP flows at specific base stations (this readily translates to per flow delay guarantees as well).
3) Provide *statistical multiplexing gains* to both services, by optimally allocating the resources between VIP and BE flows, subject to respecting the strict priority of the VIP.
4) Ensure that when all the above goals are satisfied, loads across BSs are configured to provide *optimal network-wide performance* (in terms of different tunable metrics).

As a final note, the proposed framework achieves the set goals, while being minimally invasive to existing schedulers. It introduces a second queue, but still applies the existing scheduling policy on each queue. We believe our results offer promise for a number of envisioned 5G developments. For example, the framework can be extended to explicitly support delay-sensitive 5G service classes (e.g. URLLC [7]), as well as for resource allocation among multiple slices [5], [8], [9], corresponding to different operators or services. To extend our framework to multiple classes, a combination of priority queuing and discriminatory processor sharing [10] could be used. We defer this to future work.

A. Related Work and our Contribution

A number of more advanced schemes, beyond simple SINR-based, have been proposed to take advantage of the dense deployment of base stations and utilize all available resources [11]–[13]. These schemes aim to optimally balance different, often conflicting goals such as: giving each user a high enough rate, minimizing the average network-wide delay, and ensuring that no base station is congested. A seminal work in this direction is the framework of [12] that utilizes the α-optimal function to balance these goals, and derives optimal and distributed user association rules, assuming best effort flows. This work has since been extended to jointly

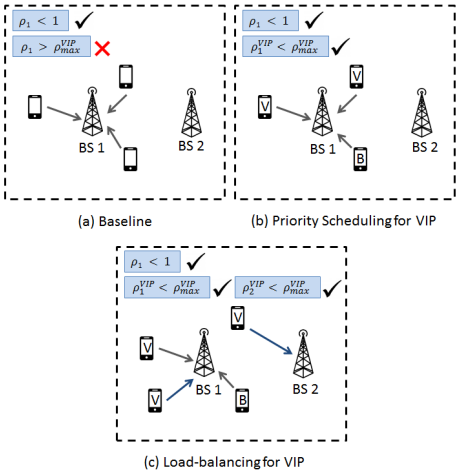


Fig. 1. Simplified example for different association policies

optimize uplink and downlink traffic [14], consider backhaul constraints [15], energy efficiency [16], and a number of other directions. In the context of slicing, the authors in [17] propose dynamic sharing of radio access network resources between (virtual) operators; while the aforementioned approach achieves capacity savings for the tenants, it cannot give performance guarantees for a class of applications (VIP).

Consider Fig.1 as a motivating example. A baseline association approach that does not considers flow differentiation (no priorities and no VIP constraints) will lead to high load for VIP flows in Fig.1a for BS 1 and low quality wireless service. Enhancing the network with a scheme that explicitly differentiates between VIP and BE flows and uses a scheduler that prioritizes VIP flows (as in Fig.1b), temporarily fixes the performance guarantee for VIP flows with no impact on total load. When, as shown in Fig.1c a 3rd VIP flow is added to BS 1 though, the QoS for BS 1 cannot be met due to the high concentration of VIP load. In this configuration, our proposed algorithm will switch one VIP flow to BS 2, achieving the VIP QoS for both base stations, with a trade-off cost of reduced SINR or instantaneous rate for the switched flow.

In our work we extend the user association framework of [12] and make the following contributions: (i) Different from [12], to ensure isolation we assume a priority-based MAC scheduler at each BS, where VIP flows are always served before BE flows. We show that such a BS scheduler can be modeled as a Processor Sharing (PS) queue with 2 priorities. (ii) We introduce a constraint on the VIP load at each base station, that is used to provide appropriate VIP performance guarantees. (iii) We derive novel distributed user association rules for both the VIP and BE services, that provably converge to optimize an α-optimal function of the *total* base station load, thus maximizing the statistical multiplexing gains within the feasible region. (iv) We show that our association policy outperforms both the original best effort policy of [12], and an improved version of the latter, adapted for the 2-class setup, using the original association rule but also giving priority to VIP flows, using telecom traces from the area of Milano [18].

II. Architecture

A. System Model

Spatial traffic. We consider a region $\mathcal{L} \in \mathbf{R}^2$ with coverage from $\mathcal{B}$ (heterogeneous) base stations. At each point $x \in \mathcal{L}$, users generate flow requests according to an inhomogeneous Poisson point process with spatial intensity $\lambda(x)$ and have independently distributed file sizes with mean $\frac{1}{\mu(x)}$.

Service Rate. If users at point $x \in \mathcal{L}$ are associated to base station $i \in \mathcal{B}$, then their flows are served with rate $C_i(x)$. In our paper, $C_i(x)$ will be a location-dependent metric that depicts the wireless signal degradation due to distance. One might be tempted to include in $C_i(x)$ channel fading and dynamic interference, which depends on user association decisions and power control, however not considering these phenomena is common in user association literature [12]–[14], [19], next we give a short justification. Fast fading is averaged out due to the time scale of association being large compared to the time of channel coherence. Furthermore, since we are considering a low mobility environment the result of slow fading or shadowing can be captured from SINR. Finally, if SINR is measured based on reference pilot signals emitted by all base stations simultaneously the measurement produces approximately equivalent results with the assumption that nearby base stations are saturated. This assumption is common in most related work [6], [12]–[15].

Here we give a specific model for $C_i(x)$, with the understanding that other models that are location-dependent are admissible in our work:[1]

$$C_i(x) = W \log(1 + (\text{SINR}_i(x))), \tag{1}$$

where W is the available frequency band, and $\text{SINR}_i(x)$ is given by:

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

Above, P_i denotes the transmission power of base station i, N_0 denotes noise power and $G_i(x)$ is the path loss between the antenna and the UE. In our analysis, powers are static.

Flow Differentiation. Depending on service requirements the flows are classified into VIP (V) and Best Effort (B). The overarching goal is to provide VIP flows with better service quality, which we will achieve via a two step scheme: (i) the service of VIP flows is strictly prioritized over Best Effort (thus they only compete for service with other VIP flows), and (ii) the load of VIP flows is regulated at each base station.

Association Rules. Let $\pi_i^V(x) = \{0, 1\}$ and $\pi_i^B(x) = \{0, 1\}$ be the association rules, indicating if point x flow of type V, B is associated with base station i. At each time instance we enforce the constraint that points are uniquely associated to one base station, hence $\sum_{i \in \mathcal{B}} \pi_i^T(x) = 1$, where $T = V, B$. The association variables $\pi_i^B(x), \pi_i^V(x), \forall x \in \mathcal{L}$ will be the means to control the performance of the system.

[1]We clarify that despite the particular modeling of $C_i(x)$, the association decisions remain coupled through the base station loads that affect the performance received at each point.

Base Station Load. The fraction of time required to deliver the traffic load destined to location $x \in \mathcal{L}$ by base station i is defined as the load density for i in x:

$$\varrho_i^T(x) = \frac{\lambda^T(x)}{\mu^T(x)C_i(x)}, \qquad T = V, B.$$

The fraction of time a base station i is busy, called the load ρ_i, due to a specific type of traffic V or B is given by:

$$\rho_i^T = \int_{\mathcal{L}} \varrho_i^T(x)\pi_i^T(x)dx, \quad T = V, B. \tag{2}$$

The total load of base station i is then

$$\rho_i = \rho_i^V + \rho_i^B.$$

The base station load vector $\boldsymbol{\rho} = (\rho_i)$ is an important performance metric of the system. In subsection Sect.II-C we will explain how we can choose association rules $\pi_i^B(x), \pi_i^V(x), \forall x \in \mathcal{L}$ to control $\boldsymbol{\rho}$ and ultimately provide differentiated levels of service to different user flow types.

B. Queue Delay Model

Even though the exact dynamics of the LTE schedulers are not standardized, it is generally accepted that in practice a proportional-fair scheduling policy ($\alpha \approx 1$) is used [2], [13], [20]. In this case and most general case of temporal fair schedulers, the dynamics of the base station queues are captured by a multi-class $M/G/1$ processor sharing system [6]. A well known result for Processor Sharing [21] is that the stationary distribution of the number of customers is insensitive to the distribution of service times, hence assuming $\rho_i < 1$ we have the following: The stationary distribution for the total number of flows in base station i is:

$$q_i(n) = \rho_i^n(1 - \rho_i).$$

By the above equation we can show that the mean number of active flows in BS i is:

$$E[N_i] = \frac{\rho_i}{1 - \rho_i}.$$

By Little's Law the Expected Delay (Response time) in Queue i is:

$$E[D_i] = \frac{1}{\lambda_i}E[N_i] = \frac{1}{\lambda_i}\frac{\rho_i}{1 - \rho_i}.$$

where the incoming arrivals at the queue are

$$\lambda_i = \int_{\mathcal{L}} (\lambda^V(x)\pi_i^V(x) + \lambda^B(x)\pi_i^B(x))dx.$$

Proposition 1 (Isolation and Performance Guarantee)**.** *By considering that the base stations also follow a preemptive priority scheduling policy in favor of the VIP flows, all of the above equations can be rewritten for ρ_i^V. By limiting the VIP load $\rho_i^V \leq c_i$ we get the upper bounds:*

$$E[N_i^V] \leq \frac{1}{1 - c_i} - 1,$$
$$E[D_i^V] \leq \frac{1}{\lambda_i}(\frac{1}{1 - c_i} - 1).$$

Therefore, in order to guarantee a certain average delay performance, below we optimize $\boldsymbol{\rho}$ subject to ensuring the constraint $\rho_i^V \leq c_i$ at each base station.

C. Dual Service Problem Formulation

First, we relax the integrality requirement of the association decisions. In practice, a fractional association could be interpreted as a time sharing of different integral associations, hence different jobs may be handled by different base stations. With this relaxation, a feasible association vector $\boldsymbol{\pi} = \{\boldsymbol{\pi}^V, \boldsymbol{\pi}^B\}$ assigns each layer of flows at location in $x \in \mathcal{L}$ with a probability $\pi_i(x) \in [0, 1]$ at base station $i \in \mathcal{B}$ such that all the flow requests are served and ensures that all the base stations are stable ($\rho_i < 1, \ \forall i \in \mathcal{B}$). Since we are also targeting performance guarantees for the service of VIP flows, we impose a threshold on the load of VIP flows per base station $\rho_i^V \leq c_i$ (Proposition 1).

Definition 1. *$\mathcal{F}$ is the convex set of flow differentiated feasible load vectors $\boldsymbol{\rho}^B, \boldsymbol{\rho}^V$:*

$$\begin{aligned}\mathcal{F} = \{\boldsymbol{\rho} \mid \ & \rho_i = \int_{\mathcal{L}} \left(\varrho_i^V(x)\pi_i^V(x) + \varrho_i^B(x)\pi_i^B(x)\right) dx \\ & 0 \leq \rho_i \leq 1 - \epsilon, \quad \forall i \in \mathcal{B} \\ & 0 \leq \rho_i^V \leq c_i, \ \forall i \in \mathcal{B} \\ & \sum_{i \in \mathcal{B}} \pi_i^T(x) = 1, \ \forall x \in \mathcal{L}, \ T = V, B \\ & 0 \leq \pi_i^T(x) \leq 1, \ \forall i \in \mathcal{B}, \ \forall x \in \mathcal{L}, \ T = V, B\},\end{aligned}$$

where c_i is the VIP load threshold for base station i.

Lemma 1. *The feasible set $\mathcal{F}$ is convex.*

Proof. Consider vectors $\boldsymbol{\rho}^1, \boldsymbol{\rho}^2 \in \mathcal{F}$ and $\boldsymbol{\rho}^1 \neq \boldsymbol{\rho}^2$. Let $\boldsymbol{\rho} = \theta\boldsymbol{\rho}^1 + (1 - \theta)\boldsymbol{\rho}^2$ with $\theta = [0, 1]$. We will show that $\boldsymbol{\rho} \in \mathcal{F}$. The elements of vector $\boldsymbol{\rho}$ are $\rho_i = \theta\rho_i^1 + (1 - \theta)\rho_i^2$ and thus

$$\begin{aligned}\rho_i = \theta & \int_{x \in \mathcal{L}} (\varrho_i^V(x)\pi_i^{1V}(x) + \varrho_i^B(x)\pi_i^{1N}(x))dx \\ & + (1 - \theta) \int_{x \in \mathcal{L}} (\varrho_i^V(x)\pi_i^{2V}(x) + \varrho_i^B(x)\pi_i^{2N}(x))dx \\ , \rho_i = & \int_{x \in \mathcal{L}} \varrho_i^V(x)(\theta\pi_i^{1V}(x) + (1 - \theta)\pi_i^{2V}(x))dx \\ & + \int_{x \in \mathcal{L}} \varrho_i^B(x)(\theta\pi_i^{1N}(x) + (1 - \theta)\pi_i^{2N}(x))dx.\end{aligned}$$

Considering $\pi_i^V(x) = \theta\pi_i^{1V}(x) + (1-\theta)\pi_i^{2V}(x)$ and $\pi_i^B(x) = \theta\pi_i^{1B}(x) + (1-\theta)\pi_i^{2B}(x)$, we may check that vector $\boldsymbol{\rho}$ satisfies all the equations in $\mathcal{F}$, thus $\boldsymbol{\rho}$ is feasible and $\mathcal{F}$ is convex. □

In the optimization scope of the user association problem, the objective is to select from the feasible vectors $\mathcal{F}$ the vector that optimizes a selected network performance. As we have defined in section II-B, we would like to optimize for some metric defined by $\boldsymbol{\rho}$, the load vector of the base station queues. For this purpose we select the α-optimal objective functions.

Definition 2. *The α-optimal functions, for $\alpha \in [0, \infty)$ are:*

$$\phi_\alpha(\boldsymbol{\rho}) = \begin{cases} \sum_i \frac{(1-\rho_i)^{1-\alpha}}{\alpha - 1} & \alpha \neq 1 \\ \sum_i \log(\frac{1}{1-\rho_i}) & \alpha = 1 \end{cases}$$

The optimization of $\phi_\alpha(\boldsymbol{\rho})$ shares some analogies with the commonly used in resource allocation optimization family of α-fair functions. It is shown in [12] that selecting: (i)

$\alpha = 0$ maximizes the throughput of the system, (ii) $\alpha = 1$ maximizes the geometric mean of base station's idle time (1-ρ_i), (iii) $\alpha = 2$ minimizes the average number of flows in the base station queues, and (iv) $\alpha \to \infty$ leads to max-min load fairness described in [22]. We can now formulate our primal problem.

Problem 1 (P1: The Service Differentiation User Association Problem)**.**

$$\underset{\boldsymbol{\rho}\in\mathcal{F}}{\textit{minimize}}\ \phi_\alpha(\boldsymbol{\rho}) = \sum_{i\in\mathcal{B}} \frac{(1-\rho_i)^{1-\alpha}}{\alpha-1}. \tag{3}$$

Since $\phi_\alpha(\boldsymbol{\rho})$ is a convex function and $\mathcal{F}$ is a convex set, P1 is a convex optimization problem. Below we exploit the convexity of P1 to design an algorithm that finds the optimal solution $\boldsymbol{\rho}^\star$ in a distributed manner. We will show that our algorithm yields integral association rules that converge to the optimal solution of P1, hence it also solves the integral counterpart of P1.

III. Distributed Constrained User Association

In this section we will solve P1 in a distributed fashion by using the theory of Lagrangian relaxation for constrained convex optimization. Initially, we relax the box VIP load constraint allowing its violation at a price γ_i. This will allow us to derive the optimal association rules for given user class, $x \in \mathcal{L}$, $\boldsymbol{\gamma}$ prices and $\boldsymbol{\rho}$ vector of loads. The derived rules can be used to iteratively solve the user association problem for fixed $\boldsymbol{\gamma}$. We will then show how to update price vector $\boldsymbol{\gamma}$ in order to converge to the optimal solution of P1. The full algorithm in steps is presented in subsection III-D.

A. Partial Lagrangian Relaxation

After relaxing the load constraint the feasible load vectors are $\boldsymbol{\rho} \in \mathcal{F}'$, where $\mathcal{F}'$ allows $\boldsymbol{\rho}^V \in [0,1)$ and the objective function is the partially relaxed Lagrangian:

$$\Phi_\alpha(\boldsymbol{\rho},\boldsymbol{\gamma}) = \sum_{i\in B} \frac{(1-\rho_i)^{1-\alpha}}{\alpha-1} + \sum_{i\in B} \gamma_i(\rho_i^V - c_i). \tag{4}$$

Problem 2 (P2: Relaxed User Association Problem)**.**

$$\underset{\boldsymbol{\gamma}\geq 0}{\textit{maximize}} \left\{ \underset{\boldsymbol{\rho}\in\mathcal{F}'}{\textit{minimize}} \left\{ \Phi_\alpha(\boldsymbol{\rho},\boldsymbol{\gamma}) \right\} \right\}. \tag{5}$$

We will prove that the solutions of the relaxed problem (P2) will be primal optimal and primal feasible.

B. Optimal User Association Rules

Lemma 2. *If P2 is feasible, the optimal association decision for each user is given by the following rule, depending on user type:*

$$\pi_i^{\star V}(x) = 1\left\{ i^V(x) = \underset{j\in B}{\textit{argmax}} \left\{ \frac{C_j(x)(1-\rho_j^\star)^\alpha}{1+\gamma_j^\star(1-\rho_j^\star)^\alpha} \right\} \right\}, \tag{6}$$

$$\pi_i^{\star B}(x) = 1\left\{ i^B(x) = \underset{j\in B}{\textit{argmax}} \left\{ C_j(x)(1-\rho_j^\star)^\alpha \right\} \right\}, \tag{7}$$

where $\rho_j^\star$ and $\gamma_j^\star$ are an optimal load and price vector of the problem above.

Proof. (Optimality of $\boldsymbol{\pi}^{\star V}, \boldsymbol{\pi}^{\star B}$ given $\boldsymbol{\rho}^\star, \boldsymbol{\gamma}^\star$). We have by using Eq.(2) and Eq.(4):

$$\begin{aligned}
&\langle \nabla_{\boldsymbol{\rho}} \Phi_\alpha(\boldsymbol{\rho}^\star) \cdot \Delta\boldsymbol{\rho}^\star \rangle = \\
&= \sum_{i\in B} \left(\frac{\partial \Phi_\alpha}{\partial \rho_i^V} \Delta\rho_i^{\star V} + \frac{\partial \Phi_\alpha}{\partial \rho_i^B} \Delta\rho_i^{\star B} \right) = \\
&= \sum_{i\in B} \left(\frac{1}{(1-\rho_i^\star)^\alpha} + \gamma_i^\star \right) (\rho_i^V - \rho_i^{\star V}) \\
&\quad + \sum_{i\in B} \left(\frac{1}{(1-\rho_i^\star)^\alpha} \right) (\rho_i^B - \rho_i^{\star B}) = \\
&= \int_{\mathcal{L}} \varrho^V(x) \sum_{i\in B} \frac{1+\gamma_i^\star(1-\rho_i^\star)^\alpha}{C_i(x)(1-\rho_i^\star)^\alpha} \left(\pi_i^V(x) - \pi_i^{\star V}(x) \right) dx \\
&\quad + \int_{\mathcal{L}} \varrho^B(x) \sum_{i\in B} \frac{1}{C_i(x)(1-\rho_i^\star)^\alpha} \left(\pi_i^B(x) - \pi_i^{\star B}(x) \right) dx,
\end{aligned}$$

since $\pi_i^{\star T}(x)$ satisfy Eq.(6) and Eq.(7):

$$\sum_{i\in B} \frac{1+\gamma_i^\star(1-\rho_i^\star)^\alpha}{C_i(x)(1-\rho_i^\star)^\alpha} \pi_i^V(x) \geq \sum_{i\in B} \frac{1+\gamma_i^\star(1-\rho_i^\star)^\alpha}{C_i(x)(1-\rho_i^\star)^\alpha} \pi_i^{\star V}(x),$$

$$\sum_{i\in B} \frac{\pi_i^B(x)}{C_i(x)(1-\rho_i^\star)^\alpha} \geq \sum_{i\in B} \frac{\pi_i^{\star B}(x)}{C_i(x)(1-\rho_i^\star)^\alpha}.$$

Hence, the first order convex optimality criterion is met [23]:

$$\langle \nabla_{\boldsymbol{\rho}} \Phi_\alpha(\boldsymbol{\rho}^\star) \cdot \Delta\boldsymbol{\rho}^\star \rangle \geq 0,$$

and $\boldsymbol{\pi}^{\star V}, \boldsymbol{\pi}^{\star B}$ are optimal association vectors. □

The association rules produce two association maps[2], one for each service, where a user at location $x \in \mathcal{L}$ is associated to base station $i \in \mathcal{B}$ if this is prescribed by the corresponding map. The aggregate of all the assigned load densities to the serving Base Station is equivalent to the optimal $\rho_i^\star$. Moreover, the optimal association rules Eq.(6) and (7) are deterministic, which proves that the optimal values for $\pi_i(x)$ are integral and that there is no loss of accuracy due to the continuous relaxation. More intuition about the proof is given in [12].

Starting from an initial load vector $\boldsymbol{\rho}^{(0)}$ we can iteratively find the $\boldsymbol{\rho}$ that minimizes Φ for fixed $\boldsymbol{\gamma}$, by using the association rules we derived. The process is described in steps in III-D. In the following subsection we will show how to update the prices $\boldsymbol{\gamma}$ to reach the primal optimal $\boldsymbol{\rho}^\star$.

C. Maximization Method

The maximization step of P2 depends on the selection of the α-objective. By selecting $\alpha > 0$ the partially relaxed Lagrangian is differentiable and we can solve the master problem by gradient method. When $\alpha = 0$, then we use a subgradient method for the maximization.

1) Gradient Ascent for $\alpha > 0$: The Hessian matrix of our α-optimal cost function Eq.(2) for $\alpha > 0$ is positive definite, thus is strictly convex. By Proposition 6.1.1 and Appendix B of Nonlinear Programming [24], since the load constraint

[2]An optimal map at a time instance assigns the total traffic of a location of the grid $\mathcal{L}$ to the optimal serving base station i. Since, we differentiate how we assign locations based on flow class, we have two Association Maps, one for BE traffic and one for VIP. The maps are direct represenations of their respective association vectors $\boldsymbol{\pi}^{B\star}(x), \boldsymbol{\pi}^{V\star}(x)$.

function is linear, the set $\mathcal{F}$ is convex and compact (closed, bounded and subset of R^n) and ϕ is strictly convex, the minimized partial Lagrangian function is differentiable and

$$\nabla_{\boldsymbol{\gamma}}\Phi_\alpha(\boldsymbol{\rho}^\star,\boldsymbol{\gamma}) = \boldsymbol{\rho}^{\star V} - c.$$

The gradient ascent algorithm will update the prices $\boldsymbol{\gamma}$ of the outer problem iteratively:

$$\boldsymbol{\gamma}^{(k+1)} = \boldsymbol{\gamma}^{(k)} + s^{(k)}\nabla_{\boldsymbol{\gamma}}\Phi_\alpha(\boldsymbol{\rho}^{(k)},\boldsymbol{\gamma}),$$

with a constant step size $s^{(k)} = c$.

2) Subgradient Method for $\alpha = 0$: The α-optimal cost function Eq.(2) is affine (non-strictly convex):

$$\phi_0(\boldsymbol{\rho}) = \sum_{i\in\mathcal{B}} \rho_i.$$

The minimized partial Lagrangian over $\boldsymbol{\rho} \in \mathcal{F}'$ is non-differentiable. The subgradient iteration is similar to the gradient but since the gradient may not exist, a subgradient $g^{(k)}$ is used instead. One of the subgradients of the minimized over $\boldsymbol{\rho}$ Lagrangian function is the load constraint function at a minimizer of the Lagrangian $\boldsymbol{\rho}^\star$. By selecting a sufficiently small step size it is shown that the subgradient method minimizes the distance to the optimal solution and converges to the optimal. The subgradient method will update the prices $\boldsymbol{\gamma}$ of the outer problem according to the iteration:

$$\boldsymbol{\gamma}^{(k+1)} = \boldsymbol{\gamma}^{(k)} + s^{(k)}g^{(k)}$$

with a subgradient step $s^{(k)} = \frac{1}{\sqrt{k}}$.

D. Distributed Constrained User Association Algorithm

Below, we present the Distributed Constrained User Association Algorithm (DCUAA) that solves P2 by combining the optimal user association rules described in Sect.III-B and the maximization methods described in Sect.III-C. The algorithm is trigger based and will iterate until convergence is reached. The output will be an optimal association map for both flow types connecting location $x \in \mathcal{L}$ to serving base station $i \in \mathcal{B}$.

Distributed Constrained
User Association Algorithm (DCUAA)

Iterate over t until convergence

Base Stations calculate $\gamma_i^{(t+1)} \leftarrow \left[\gamma_i^{(t)} + s^{(t)}\nabla_g\Phi_\alpha\right]^+$

Broadcast $\gamma_i^{(t+1)}$

Iterate over k until convergence

User at location $x \in \mathcal{L}$ calculates $\pi_i(x)$:

$$\pi_i^V(x) = 1\left\{ i^V(x) = \underset{j\in\mathcal{B}}{\text{argmax}} \left\{ \frac{C_j(x)(1-\rho_j^{(k)})^\alpha}{1+\gamma_j^{(t+1)}(1-\rho_j^{(k)})^\alpha} \right\} \right\}$$

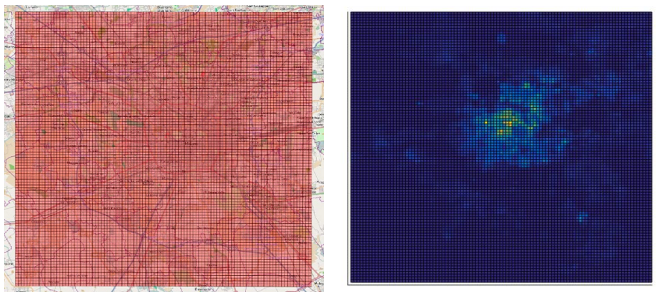
$$\pi_i^B(x) = 1\left\{ i^B(x) = \underset{j\in\mathcal{B}}{\text{argmax}} \left\{ C_j(x)(1-\rho_j^{(k)})^\alpha \right\} \right\}$$

Base Station $i \in \mathcal{B}$ measures utilization:

$U_i^{(k)} = \min\left[\int_{\mathcal{L}} (\varrho_i^V(x)\pi_i^V(x) + \varrho_i^B(x)\pi_i^B(x))dx, 1-\epsilon\right]$

$\rho_i^{(k+1)} = \beta\rho_i^{(k)} + (1-\beta)U_i^{(k)}$

Broadcast $\rho_i^{(k+1)}$



(a) (b)

Fig. 2. Milano Data Set. (a) Overview of Milano City Grid and (b) Color map of arrival intensity per square of the grid on a busy time instance

Lemma 3. *For $\alpha > 0$, the algorithm presented above will converge on the optimal association maps for P1 ($\boldsymbol{\pi}^{V\star}, \boldsymbol{\pi}^{N\star}$).*

Proof. For $\alpha > 0$, the objective in P2 is strictly convex over $\boldsymbol{\pi}$, the hessian matrix is positive definite ($\nabla^2_{\boldsymbol{\pi}}\Phi > 0$). Hence, there exist a unique solution (association maps) to the dual problem, which is primal feasible and cost equivalent to the primal [23]. This completes the proof. □

IV. Numerical Evaluation

For the numerical evaluation of our algorithm we consider a simulation scenario that verifies the performance guarantees for the VIP flows achieved by the DCUAA for delay and also the VIP isolation from BE traffic. We use telecommunication traces from the Milano dataset [18], to confirm the results on real traffic input.

A. Simulation Setup

We assume that the user receives data at Shannon Capacity Eq.(1) and we model the propagation loss $G_i(x)$ with a path loss exponent 3:

$$G_i(x) = \left(\frac{1}{\text{dist(BS}_i \text{ to } x)}\right)^3.$$

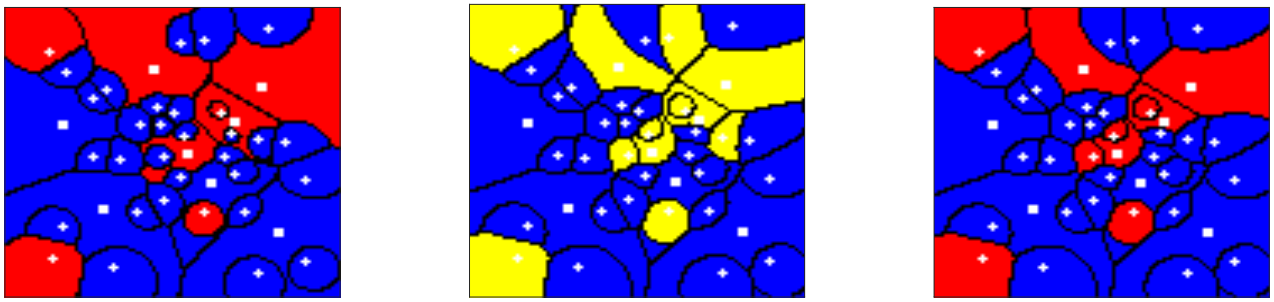
The LTE parameters for the transmission power of each base station tier and the transmission rate approximation are taken according to table I.

B. Validation of the results on the Milano dataset

To evaluate on algorithm on real telecommunication traffic traces, we will use the publicly available Milano dataset [18]. The Milano dataset provides spatially aggregated data about the telecommunication activity. The data are grouped on a regular grid overlaying the territory of Milano with 100×100 squares. Consequently, the grid designates the area $\mathcal{L}$ and every square is a location $x \in \mathcal{L}$ to be associated with base

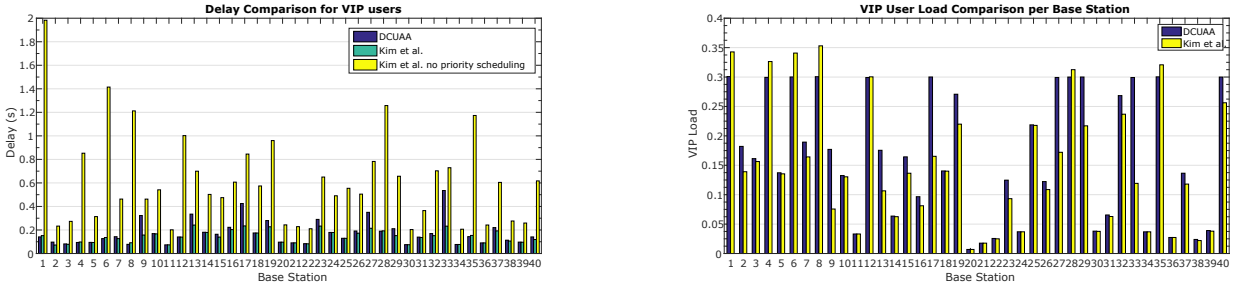
TABLE I
Simulation Parameters [14]

Parameter	Variable	Value
Transmission Power Macro BS	P_M	43 dbm
Transmission Power Micro BS	P_m	29 dbm
System Bandwidth	W	10 MHz
Noise Density	No	-174dbm/Hz



(a) (b) (c)

Fig. 3. Constraint violation comparison between the algorithms, red means violation, yellow is tight, blue means below threshold. (a) MaxSINR Association Map (b) DCUAA VIP Association Map (c) Kim et al VIP Association Map. (b) and (c) are calculated for $\alpha = 1$.



(a) (b)

Fig. 4. Average Delay and Load metric comparison per base station

stations. For every square of this grid the data set contains the aggregate per ten minutes telecommunication events in the period of 01/11/13-01/01/14.

In the experiments we consider a weekday (Tuesday 3/12/2013) during peak traffic hour at midday. An overview of the Milano Grid and the midday arrival intensity per square of the grid can be see in Fig.2a and Fig.2b, respectively. Since there is no priority differentiation for traffic in the dataset, we arbitrarily select 2, out of the 6, 10 min samples in an hour as VIP traffic. This is based on the fact that we expect VIP flow requests to manifest in (smaller) proportion to the expected total flow requests in an area.

We consider a two-tier Heterogeneous deployment of 40 Base Stations with fixed positions. The wireless network in the Milano area is simulated by 8 eNBs and 32 microcells, which are deployed with increased density in the area corresponding to the city center. We specifically design this subset of base stations, to accurately simulate a simplified environment of a dense deployment in a city, bringing in the front all the aspects of the user association problem.

We will show that our algorithm guarantees the performance of VIP flows and we will use as baseline, our implementation of the algorithm described in [12]. The association objective is set to maximize throughput ($\alpha = 1$) and the utilization cap for VIP flows is set to $c_{i\in\mathcal{B}} = 30\%$. Each of the plots in Fig.3 demonstrate location $\mathcal{L}$, the Base Stations are the white dots, squares are eNBs, circles are microcells, the black lines are the borders for the area of service of each corresponding base station and colors indicate whether an area satisfies the performance constraint. Blue means below threshold, yellow means utilization is at the threshold, red means that utilization is above the requirement. In Fig.3a, Fig.3c and Fig.3b we notice the areas of $\mathcal{L}$ that achieve the $\rho_i^V < c_i$ constraint, depending on the algorithm and user association policy selected.

A short summary of the most important results of this simulation can be found in table II. In table II we can see that DCUAA achieves the set VIP performance guarantee, while not greatly degrading the overall performance in average delay of the total traffic. In contrast Kim et al. [12] best effort only algorithm does not achieve the VIP constraints for up to 46.5% of the incoming arrivals.

Finally, in Fig.4, we can see per base station, the average delay and load performance of the two algorithms in comparison. We can see that our algorithm effectively moves flows from the crowded base stations to less congested ones, with a trade off of reduced average delay on the congested base stations but slightly increased load in average.

TABLE II
SIMULATION RESULTS ON THE MILANO DATASET

Algorithm	Constr. Violation %	Av. Delay Tot (s)
DCUAA $\alpha = 1$	0	0.2575
Kim et al. $\alpha = 1$	46.5	0.2471
DCUAA $\alpha = 2$	0	0.2814
Kim et al. $\alpha = 2$	38.9	0.2874
DCUAA $\alpha = 5$	0	0.4404
Kim et al. $\alpha = 5$	9.5	0.4368

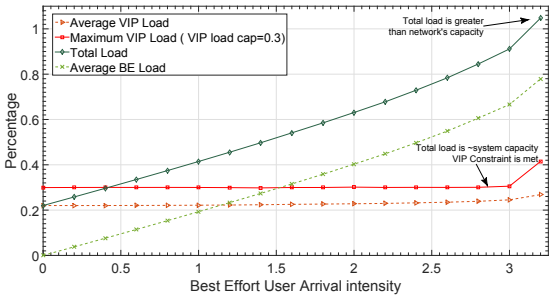


Fig. 5. Service Isolation. Increasing BE arrival intensity, while not changing VIP arrival intensity, has no effect on VIP guaranteed performance inside the Feasibility Region.

C. Service Isolation

In order to demonstrate that our algorithm provides service isolation, which is an important aspect of future wireless network slicing architecture, we successively scale up the arrival intensity of Best Effort users, from 0 up to the point that the average total incoming traffic exceeds the wireless network's capacity.

For each successive scaling of the BE input traffic we solve the optimal network maps (user association maps for VIP and BE flows) with the DCUAA. In Fig.5 we plot the best effort arrival intensity scaling factor and over all the base stations according to the optimal maps created by the DCUAA: (i) the average VIP load, (ii) the maximum VIP load (which has to be bellow the 0.3 load guarantee threshold), (iii) the total load of the system and (iv) the average BE effort load. From the Fig.5 it is clear that, the algorithm converges to configurations that guarantee the performance of the VIP flows, as long as the average total traffic is admissible.

V. Conclusion

In this paper we have proposed a framework for user association based on distributed constrained optimization for 5G New Radio (NR) in the context of future Ultra Dense Networks (UDNs). We have derived distributed association rules, that provably converge to the optimum point of operation. The resulting association guarantees performance to the VIP flows whilst balancing the load between both service types. The method is based on non-invasive extensions to current wireless networks, while it can be generalized in future work for multiple class priorities and applications in wireless network slicing. Initial Simulation results demonstrate the capabilities of the framework in bounding the mean number of VIP flows at the base stations and also the improved performance over best effort only policies.

References

[1] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer, "Network Densification: The Dominant Theme for Wireless Evolution into 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, February 2014.

[2] S. Sesia, I. Toufik, and M. Baker, "LTE-the UMTS Long Term Evolution: From Theory to Practise," 2015.

[3] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. K. Wong, R. Schober, and L. Hanzo, "User Association in 5G Networks: A Survey and an Outlook," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 1018–1044, 2016.

[4] "Making 5G a reality: Addressing the strong mobile broadband demand in 2019 & beyond," Qualcomm and Nokia, 2018. [Online]. Available: https://www.qualcomm.com/documents/whitepaper-making-5g-nr-reality

[5] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From Network Sharing to Multi-Tenancy: The 5G Network Slice Broker," *IEEE Communications Magazine*, July 2016.

[6] T. Bonald and A. Proutière, "Wireless Downlink Data Channels: User Performance and Cell Dimensioning." ACM.

[7] A. Destounis, G. S. Paschos, J. Arnau, and M. Kountouris, "Scheduling URLLC Users with Reliable Latency Guarantees," *WiOpt*, May 2018.

[8] N. Nikaein, E. Schiller, R. Favraud, K. Katsalis, D. Stavropoulos, I. Alyafawi, Z. Zhao, T. Braun, and T. Korakis, "Network Store: Exploring Slicing in Future 5G Networks." ACM.

[9] S. Vassilaras, L. Gkatzikis, N. Liakopoulos, I. N. Stiakogiannakis, M. Qi, L. Shi, L. Liu, M. Debbah, and G. S. Paschos, "The algorithmic aspects of network slicing," *IEEE Communications Magazine*, 2017.

[10] E. Altman, K. Avrachenkov, and U. Ayesta, "A Survey on Discriminatory Processor Sharing," *Queueing Systems*, 2006.

[11] D. Fooladivanda and C. Rosenberg, "Joint Resource Allocation and User Association for Heterogeneous Wireless Cellular Networks," *IEEE Transactions on Wireless Communications*, January 2013.

[12] 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*, Feb 2012.

[13] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User Association for Load Balancing in Heterogeneous Cellular Networks," *CoRR*, 2012.

[14] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, "An Analytical Framework for Optimal Downlink-Uplink User Association in HetNets with Traffic Differentiation," 2015.

[15] ——, "Optimal Downlink and Uplink User Association in Backhaul-limited HetNets," in *IEEE INFOCOM*, 2016.

[16] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Toward Energy-Efficient Operation of Base Stations in Cellular Wireless Networks," *Green Communications: Theoretical Fundamentals, Algorithms, and Applications*, 2012.

[17] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Prez, "Multi-tenant radio access network slicing: Statistical multiplexing of spatial loads," *IEEE/ACM Transactions on Networking*, Oct 2017.

[18] Telecom Italia, "Telecommunications - SMS, Call, Internet - MI," 2015. [Online]. Available: http://dx.doi.org/10.7910/DVN/EGZHFV

[19] K. Shen and W. Yu, "Distributed Pricing-Based User Association for Downlink Heterogeneous Cellular Networks," *CoRR*, 2014.

[20] "eLTE2.3 DBS3900 LTE TDD Basic Feature Description," Huawei Technologies Co. Ltd., 2014. [Online]. Available: http://www.huawei.com/ilink/cnenterprise/download/HW_328213

[21] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*, 1st ed. New York, NY, USA: Cambridge University Press, 2013.

[22] J. Mo and J. Walrand, "Fair End-to-End Window-based Congestion Control," *IEEE/ACM Transactions Networking*.

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

[24] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1995.