Distributed Alpha-Fair Throughput Aggregation in Multi-RAT Wireless Networks

Ehsan Aryafar[1] and Alireza Keshavarz-Haddad[2]
[1]Portland State University, Department of Computer Science, Portland, OR, USA
[2]Shiraz University, School of Electrical and Computer Engineering, Shiraz, Iran

***Abstract*—Today's wireless devices can be simultaneously connected to multiple communication networks based on different radio access technologies (RATs) such as WiFi, 3G, and LTE. Simultaneous aggregation of each client's traffic across multiple such RATs or the corresponding base stations (BSs) can substantially improve the quality of experience for each client and improve the overall network utilization and efficiency.**

Our goal in this paper is to design distributed resource allocation algorithms that can be independently executed by each BS and achieve alpha-fairness among the clients. In particular, we derive a simple water filling based solution and study its theoretical aspects such as convergence, optimality, evolution, and convergence time. We also characterize its performance through a comprehensive multi-RAT simulator, and show that it achieves superior performance to alternative distributed solutions due to its fast and low-overhead operation.

I. Introduction

The explosive growth in the number of wireless devices and emerging bandwidth intensive applications (*e.g.*, augmented reality) has led the wireless industry to develop 5G New Radio (NR) technologies for different parts of the wireless spectrum, including sub-1 GHz, 1-6 GHz, and mmWave bands [1]. These new wireless technologies will augment existing heterogeneous cellular networks (HetNets), which already leverage a wide variety of RATs (*e.g.*, 3G, LTE, and WiFi) to serve their clients. In parallel, client device (*e.g.*, smartphone) manufacturers are also equipping their devices with an increasing number of RATs. Simultaneous aggregation of each client's traffic across multiple such RATs can substantially enhance the client performance, *e.g.*, boost its capacity, increase its resiliency when one RAT fails or becomes congested, and enhance its mobile operation (*e.g.*, a mobile client can continue to receive traffic through LTE while switching its WiFi radio).

A. Related Work

To realize the performance gains associated with multi-RAT HetNets, several standardization organizations have been developing multi-RAT integration solutions. For example, IETF (Internet Engineering Task Force) Internet area working group is developing a multi-access management protocol to support multi-RAT combination on the core network side [2]. On a parallel front, the 3GPP has already standardized several PHY layer aggregation (*e.g.*, LTE-U, LAA [3]) and MAC layer aggregation (*e.g.*, LWA [4]) techniques. More recently, the 3GPP is standardizing different methods to support 5G NR access to unlicensed bands as well as mechanisms to support MAC layer aggregation across any combination of RATs (*e.g.*, 3G, WiFi, LTE, and NR) [5]. These efforts in totality will support multi-RAT traffic aggregation at different network layers (including PHY, MAC, transport, and application layers) with diverse performance tradeoffs.

We consider MAC-level aggregation as it is generally more efficient than transport/application layer aggregation solutions due to the availability of instantaneous channel information at the MAC layer. It can also be applied to existing HetNets, whereas majority of consumer and network side wireless equipment do not yet support PHY layer aggregation.

In particular, our goal is to design a distributed resource allocation algorithm for each base station (BS)[1] such that the total throughput achieved by each client across its RATs satisfies α-fairness. Alpha-fairness [6] is a unifying mathematical formulation to achieve fair throughput assignment. The degree of fairness is defined by a parameter $\alpha \in [0, \infty)$, which controls the tradeoff between *fairness* and *efficiency* (*i.e.*, total throughput maximization). Several special cases of α correspond to well-known fairness metrics, *e.g.*, $\alpha \rightarrow \infty$ corresponds to max-min fair allocation (which may be considered as the most fair allocation), $\alpha = 2$ corresponds to delay minimization, $\alpha \rightarrow 1$ corresponds to proportional fairness, $\alpha = \frac{1}{2}$ corresponds to harmonic fairness, and $\alpha = 0$ corresponds to throughput maximization (without any fairness consideration). Alpha-fairness has been applied to a wide variety of networking problems, but this is the first paper to propose algorithms that achieve it in HetNets.

Several research works have shown the performance gains associated with multi-RAT traffic aggregation, *e.g.*, in the case of LTE-WiFi aggregation [7]. However, none of these works has considered fairness among the clients. In our recent work [8], we presented a water filling based algorithm that can achieve proportional fairness. This paper builds on the algorithm in [8] by showing that water filling-based algorithms can also enforce generic α-fairness, presents the theoretical underpinning of the water filling solution (Theorem 1) and its connection to game theory, introduces new proof approaches to analyze algorithm convergence, optimality, and speed, and uses a comprehensive simulator (calibrated with 3GPP specifications) to demonstrate its practical performance.

B. Research Contributions

Our key contributions can be summarized as follows:

[1]**We use BS as a generic term that refers to NB in 3G, eNB in LTE, AP in WiFi, and gNB in 5G.**

- **Algorithm Design:** We first study the best resource allocation policy for each BS (assuming that other BSs do not change their resources) to gain intuition for our problem. We show that if a BS wants to *unilaterally* maximize the system objective, it should allocate its resources according to an α-dependent water filling operation [Theorem 1]. We use the learnings from this observation to design a distributed water filling algorithm (named "WFRA") for generic HetNets.
- **Theoretical Analysis:** We study the theoretical aspects of WFRA. We show that as BSs autonomously execute it, the system converges to an equilibrium [Theorem 2]. We next characterize some of the useful properties of the equilibria [Lemma 1] and prove that any outcome of WFRA is globally optimal [Theorem 3]. We also show that while the vector of optimal throughput allocations across all the clients is unique, there could be infinitely many resource allocations that can realize them [Proposition 1]. We finally derive tight bounds on WFRA's convergence time [Theorem 4] and show that while WFRA results in an optimal allocation for any finite value of α, the resulting outcome is not necessarily optimal as $\alpha \to \infty$.
- **Performance Evaluation.** We use a comprehensive multi-RAT simulator to quantify WFRA's performance. We show that our algorithm converges to the optimal solution within a few iterations. We also compare its performance against alternative distributed solutions and show that WFRA can get very close to the ideal fair outcome even in the presence of channel fluctuations and packet losses.

The paper is organized as follows. We formulate our problem and discuss the details of our algorithm in Section II. We discuss the theoretical aspects of WFRA in Section III. We present the results of our extensive simulations in Section IV. Finally, we conclude the paper in Section V. Due to page limitations, we (i) limit the discussion of related work, (ii) omit proofs of all lemmas, and (iii) omit comparisons against DComp, an alternative distributed algorithm based on well-known dual decomposition techniques. These details are provided in the technical report [9].

II. System Model

We first present our network model and problem formulation. Next, we derive a theorem that provides us with necessary intuition to design a distributed resource allocation algorithm for each BS. Finally, we present the details of our algorithm.

A. Network Model

We consider a HetNet which is composed of a set of BSs $\mathbf{M} = \{1, ..., M\}$, and a set of clients $\mathbf{N} = \{1, ..., N\}$. Each BS has a limited transmission range and only serves the clients within its range. Each client has a specific number of RATs, and therefore has access to a subset of BSs[2]. We assume that clients split their traffic over their RATs and focus on the traffic splitting problem for each client. It is itself a challenging problem to associate each client's RAT to one of the corresponding BSs (*e.g.*, choosing the optimal WiFi BS if a client has a WiFi RAT). We assume there exists a rule to pre-determine client RAT-BS association. The association rule could be any load balancing algorithm (*e.g.*, [10]), or based on the received signal strength, among others.

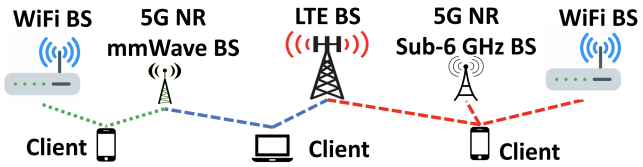

Fig. 1. **Each client is connected to a set of BSs (dotted lines) and can split or aggregate its traffic across the corresponding BSs (RATs).**

Similar to [10], [11], we assume that the transmission in one BS does not interfere with an adjacent BS. This can be achieved by means of spectrum separation between BSs that belong to different access networks, and frequency reuse among same kind BSs. We consider a multi-rate system and use $R_{i,j}$ to denote the PHY rate of client i from BS j. Since each BS generally serves more than one client, clients of the same BS need to share resources such as time/frequency slots (*e.g.* in 3/4G) or transmission opportunities (*e.g.* in WiFi). The service rate experienced by client i from BS j thus depends on the load of the BS and will therefore be a fraction of $R_{i,j}$. We assume that each BS employs a TDMA throughput sharing model and let $\lambda_{i,j}$ denote the fraction of time allocated to client i by BS j[3]. Hence, the throughput achieved by client i from BS j is equal to $\lambda_{i,j}R_{i,j}$ and its total throughput across all its RATs (denote by r_i) would be equal to $\sum_{j=1}^{M} \lambda_{i,j}R_{i,j}$.

B. Problem Formulation

Our objective is to design a distributed resource allocation algorithm for each BS such that the vector of total throughput values across all the clients $(r_1, r_2, ..., r_N)$ satisfies α-fairness. Specifically, in this paper we derive a generic algorithm and analyze its performance assuming α is a number between 0 and 1, or a number greater than 1. We discuss the theoretical aspects of our algorithm for $\alpha = 0, \to 1$, and $\to \infty$ in Section III-D. Mathematically, our problem can be formulated as

$$
\begin{aligned}
\mathcal{P}_1: \quad & \max f(\boldsymbol{\lambda}) = \textstyle\sum_{i=1}^{N} \omega_i U_\alpha(r_i) \\
s.t. \quad & r_i = \textstyle\sum_{j=1}^{M} \lambda_{i,j} R_{i,j} && \forall i \in \mathbf{N} \\
& U_\alpha(r_i) = \frac{r_i^{1-\alpha}}{1-\alpha} && \forall i \in \mathbf{N} \\
& \textstyle\sum_{i=1}^{N} \lambda_{i,j} \leq 1 && \forall j \in \mathbf{M} \\
\text{variables:} \quad & \lambda_{i,j} \geq 0 && \forall i \in \mathbf{N}, j \in \mathbf{M} \\
\text{input:} \quad & \alpha \in (0,1) \quad \text{or} \quad \alpha > 1
\end{aligned}
$$

Here, ω_i is a positive number that denotes client i's weight or priority. The second constraint captures the definition of

[2]We model clients that can aggregate traffic across same technology BSs with multiple such RATs. For example, the LTE Dual Connectivity architecture allows an LTE client to connect to two eNBs that are on different frequencies, and simultaneously utilize the radio resources that belong to both of them. Thus, we model such a client with two LTE RATs.

[3]**In Section IV, we discuss how we can extend our model and algorithm to capture practical issues such as WiFi contention and channel dynamics.**

α-fairness and the third constraint ensures that the sum of time fractions at each BS is less than 1. The feasible region in problem $\mathcal{P}_1$ is convex. Since all the $R_{i,j}$ and $\lambda_{i,j}$ values are non-negative and finite, it follows that the feasible region is also bounded, and therefore compact. Since the objective function is also concave, it follows that the α-fair allocation vector (*i.e.*, $\mathbf{r}^{opt} = (r_1, ..., r_N)$) exists and is **unique**.

Thus, problem $\mathcal{P}_1$ can be solved optimally in a centralized manner by using solvers such as CVX [12] in $O(\max(N^3, N^2M))$ computational complexity. However, a centralized solution requires information regarding the **entire topology** (*i.e.*, $R_{i,j}$s and $\lambda_{i,j}$s between every BS and its clients) for **every scheduling interval** in order to solve the problem. This significantly increases the problem complexity (particularly as the network size increases) and also introduces significant communication overhead between every BS and the central entity. Further, HetNet operators use a wide variety of backhauling technologies with diverse delay and capacity characteristics between different BSs and the central entity. The backhauling constraints limit the amount and freshness of the information that can be passed between the BSs and the central entity. The backhauling delay (*e.g.*, roundtrip latencies of 28 ms in cable and 62 ms in DSL [13]) also make the centralized solution not adaptable to system dynamics. These challenges in HetNets necessitate design of distributed resource allocation algorithms that can be independently executed by each BS.

C. Theoretical Underpinnings of the Water Filling Operation

There are two approaches to design a distributed resource allocation algorithm for problem $\mathcal{P}_1$ that can be autonomously executed by each BS. One approach, is to use dual decomposition [14] to derive a distributed algorithm[4]. The resulting algorithm converges to the optimal solution; however, it cannot be easily implemented in practice, and it incurs long convergence time and significant over-the-air (OTA) message passing overhead between every BS and its clients [9]. The second approach, is to study the optimal action by each BS in isolation to gain intuition for distributed algorithm design[5]. In particular, in our first theorem we show that if a single BS (*e.g.*, j) wants to unilaterally maximize $f(\boldsymbol{\lambda})$ (*i.e.*, the objective function in $\mathcal{P}_1$), then it must allocate its resources according to a water filling strategy. We use the intuition gained from this result to design a distributed resource allocation algorithm for HetNets in Section II-D. We show that the resulting algorithm still converges to the optimal outcome (Section III-B), but much faster and with far less OTA overhead than the dual decomposition based solution [9]. Our first theorem shows this water filling operation:

[4]Dual decomposition is a standard optimization theory technique, which is based on decomposing the Lagrangian dual problem [14]. The method is appropriate to solve $\mathcal{P}_1$, because as the constraints are relaxed through the dual formulation, the problem decouples into several subproblems, which can be solved distributedly by the clients and BSs.

[5]**This approach to solve the problem and derive a fast and low-overhead distributed algorithm is inspired by game theory. In fact, we can model the resource allocation problem in $\mathcal{P}_1$ as a game, with BSs as its players. Identifying the best resource allocation at a BS (assuming other BSs do not make a change), is referred to as the "best response strategy" in the game theory literature. Interestingly, we show later in the paper that by using the best response strategy, the system will converge to a globally optimal equilibrium (which is also a Nash equilibrium). While we highlighted the connection to game theory here, we do not discuss the issue any further in order to not confuse the reader.**

TABLE I
MAIN NOTATION

$\mathbf{N}$ and N: Set and number of all clients in the network
$\mathbf{M}$ and M: Set and number of all BSs in the network
$R_{i,j}$: PHY rate of client i to BS j
R_{max} (R_{min}): Maximum (non-zero minimum) $R_{i,j}$ across all i, j
$\lambda_{i,j}$: Fraction of time allocated to client i by BS j
$\boldsymbol{\lambda}$: Vector of $\lambda_{i,j}$s across all clients and BSs
r_i: Total throughput of client i across all its RATs
ω_i: A positive number that represents client i's weight or priority
$\omega_{\max}$ ($\omega_{\min}$): Maximum (non-zero minimum) ω_i across all clients
θ_j: Water fill level at BS j

Theorem 1 *Let $\mathbf{n}_j = \{1, ..., n\}$ denote the set of clients with non-zero PHY rate to BS j (i.e., $\forall i \in \mathbf{n}_j,\ R_{i,j} > 0$). Let r'_i denote the total throughput of client i from all BSs other than j. Then, if BS j wants to unilaterally maximize $f(\boldsymbol{\lambda})$ through allocation of its time resources (i.e., assuming no other BS changes its $\lambda_{i,j}$s), it must choose its $\lambda_{i,j}$s such that the following conditions are satisfied:*

(A) $\forall i \in \mathbf{n}_j \mid \lambda_{i,j} > 0$, *then* $\frac{r'_i + \lambda_{i,j} R_{i,j}}{(\omega_i R_{i,j})^{\frac{1}{\alpha}}} = \theta_j$

(B) $\forall i' \in \mathbf{n}_j \mid \lambda_{i',j} = 0$, *then* $\theta_j \leq \frac{r'_{i'}}{(\omega_{i'} R_{i',j})^{\frac{1}{\alpha}}}$

(C) $\sum_{i=1}^{n} \lambda_{i,j} = 1,\ \lambda_{i,j} \geq 0\ \forall i, j$

Here, θ_j is the water fill level at BS j. **Condition (A) implies that every client that receives time resources from j should reach the same water fill level at the BS. Condition (B) implies that any client (*e.g.*, i') that does not get any time resources from j, should have a higher $\frac{r_{i'}}{(\omega_{i'} R_{i',j})^{\frac{1}{\alpha}}}$ than the water fill level at j (*i.e.*, θ_j). Fig. 2 shows this water filling operation.**

Proof: Let $\lambda^*_{1,j}, \lambda^*_{2,j}, ..., \lambda^*_{n,j}$ denote the optimal time fractions at BS j that maximize $f(\boldsymbol{\lambda})$, assuming no other BS changes its time fractions. Now, if $\lambda^*_{i,j} > 0$, then for any $i' \in \mathbf{n}_j$ we must have

$$f(\boldsymbol{\lambda})\Big|_{\lambda^*_{1,j},...,\lambda^*_{i,j},...,\lambda^*_{i',j},...} \geq f(\boldsymbol{\lambda})\Big|_{\lambda^*_{1,j},...,\lambda^*_{i,j}-\epsilon,...,\lambda^*_{i',j}+\epsilon,...}$$

where ϵ is a small positive number. Leveraging the definition of $f(\boldsymbol{\lambda})$ from $\mathcal{P}_1$ (*i.e.*, $f(\boldsymbol{\lambda}) = \sum_{i=1}^{N} \omega_i U(r_i)$) in the above inequality and canceling the common terms we have

$$\omega_i U(r'_i + \lambda^*_{i,j} R_{i,j}) + \omega_{i'} U(r'_{i'} + \lambda^*_{i',j} R_{i',j}) \geq$$
$$\omega_i U(r'_i + (\lambda^*_{i,j} - \epsilon) R_{i,j}) + \omega_{i'} U(r'_{i'} + (\lambda^*_{i',j} + \epsilon) R_{i',j})$$

Therefore, if $\epsilon \to 0$ we would have

$$\omega_i \frac{\partial U(r'_i + \lambda R_{i,j})}{\partial \lambda}\Big|_{\lambda=\lambda^*_{i,j}} \geq \omega_{i'} \frac{\partial U(r'_{i'} + \lambda R_{i',j})}{\partial \lambda}\Big|_{\lambda=\lambda^*_{i',j}}$$

$$\Rightarrow \omega_i R_{i,j}(r'_i + \lambda^*_{i,j} R_{i,j})^{(-\alpha)} \geq \omega_{i'} R_{i',j}(r'_{i'} + \lambda^*_{i',j} R_{i',j})^{(-\alpha)}$$

$$\Longrightarrow \frac{r'_i + \lambda^*_{i,j} R_{i,j}}{(\omega_i R_{i,j})^{\frac{1}{\alpha}}} \leq \frac{r'_{i'} + \lambda^*_{i',j} R_{i',j}}{(\omega_{i'} R_{i',j})^{\frac{1}{\alpha}}} \quad (1)$$

If we also have that $\lambda^*_{i',j} > 0$, then we would have

$$\frac{r'_{i'} + \lambda^*_{i',j} R_{i',j}}{(\omega_{i'} R_{i',j})^{\frac{1}{\alpha}}} \leq \frac{r'_i + \lambda^*_{i,j} R_{i,j}}{(\omega_i R_{i,j})^{\frac{1}{\alpha}}} \quad (2)$$

and hence, from Eqs. (1) and (2) we would have

$$\frac{r'_i + \lambda_{i,j} R_{i,j}}{(\omega_i R_{i,j})^{\frac{1}{\alpha}}} = \frac{r'_{i'} + \lambda_{i',j} R_{i',j}}{(\omega_{i} R_{i',j})^{\frac{1}{\alpha}}} = \theta_j \quad (3)$$

On the other hand, if $\lambda^*_{i',j} = 0$, from Eq. (1) we would have

$$\frac{r'_i + \lambda^*_{i,j} R_{i,j}}{(\omega_i R_{i,j})^{\frac{1}{\alpha}}} = \theta_j \leq \frac{r'_{i'} + \lambda^*_{i',j} R_{i',j}}{(\omega_{i'} R_{i',j})^{\frac{1}{\alpha}}} = \frac{r'_{i'}}{(\omega_{i'} R_{i',j})^{\frac{1}{\alpha}}} \quad (4)$$

Finally, property (C) is because BS j can always increase $f(\boldsymbol{\lambda})$ by giving its unused time resources to its clients. Note that a client RAT is at most associated to one BS. Hence, even a single associated client can use all the BS's resources. ■

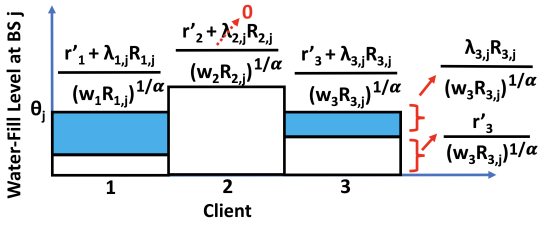

Fig. 2. **Water filling with three clients. BS j allocates its $\lambda_{i,j}$s such that clients that receive time resources reach the same water fill level (θ_j). Clients that do not receive time resources would have a higher r_i divided by $(\omega_i R_{i,j})^{1/\alpha}$ than θ_j. Here, client 2 is one such example ($\lambda_{2,j} = 0$).**

D. Distributed Resource Allocation in HetNets

We extend the water filling observation in Theorem 1 to design an autonomous resource allocation algorithm in HetNets. Algorithm WFRA depicted in Fig. 3 summarizes the steps that are executed by each BS (*e.g.*, j). There are three main steps in the algorithm: (i) clients are first sorted based on their total throughput from all BSs other than j (Line 3); (ii) the BS finds the water fill level and the corresponding $\lambda_{i,j}$s for each client associated with it; and (iii) a randomization parameter (p_j) is introduced to limit concurrent resource adaptation of a single client by multiple BSs (Line 5).

The key computational complexity of the algorithm arises from the sorting of the clients, which can be implemented in $O(n\log(n))$ complexity (n denotes the number of clients associated to BS j). Finding the water fill level and the corresponding $\lambda_{i,j}$s can be achieved in $O(n\log(n))$ complexity [8]. Thus, the total algorithm complexity is $O(n\log(n))$.

The only OTA overhead introduced by the algorithm is that each client announces to its associated BSs its total throughput (across all its RATs), whenever there is a change in that.

Algorithm WFRA: Water Filling Resource Allocation
Input: r_i , $R_{i,j}$, and $\lambda_{i,j}$ $\forall$ client i for which $R_{i,j} > 0$
randomization parameter $p_j \in (0,1)$
Output: Updated $\lambda_{i,j}$
1. Let n denote the number of clients *s.t.* $R_{i,j} > 0$
2. Find the throughput of each client i provided to it by all other BSs (*i.e.*, r'_i)
3. Sort clients based on their throughput from **other** BSs divided by their weight times PHY rate, *i.e.*,
$$\frac{r'_1}{(w_1 R_{1,j})^{1/\alpha}} \leq \ldots \leq \frac{r'_k}{(w_k R_{k,j})^{1/\alpha}} \leq \frac{r'_{k+1}}{(w_{k+1} R_{k+1,j})^{1/\alpha}} \leq \ldots \leq \frac{r'_n}{(w_n R_{n,j})^{1/\alpha}}$$
4. Run water fill, *i.e.*, find θ_j and $\lambda_{i,j}$ for each client
5. **if** *(rand* $< p_j$*)* **then** *Update* $\lambda_{i,j}$

Fig. 3. **Resource allocation algorithm autonomously run by each BS j.**

III. Convergence, Optimality, and Speed of WFRA

In this section, we investigate the convergence properties of WFRA. We first show that as BSs autonomously execute WFRA, the system converges to an equilibrium. Next, we prove that this equilibrium is optimal. Finally, we derive tight upper bounds on WFRA's convergence time.

A. Convergence to an Equilibrium

We first present a formal definition of an equilibrium.

Definition 1 *Equilibrium: The vector of time fractions across all the BSs and clients is an equilibrium outcome if none of the BSs can increase its water fill level through unilateral change of its time resource allocations.*

Our next theorem guarantees the convergence of WFRA.

Theorem 2 *Let each BS autonomously execute WFRA. Then, the system converges to an equilibrium.*

Proof: Consider $f(\boldsymbol{\lambda})$ as defined in problem $\mathcal{P}_1$. Let at each time slot a single BS execute the water filling algorithm. From Theorem 1, it follows that each time a BS makes a change to its resource allocations it increases $f(\boldsymbol{\lambda})$. Since the number of clients, BSs, and all the PHY rates ($R_{i,j}$s) are all finite numbers, $f(\boldsymbol{\lambda})$ would be upper bounded. Since at each time slot $f(\boldsymbol{\lambda})$ increases and it is a bounded function, the value of $f(\boldsymbol{\lambda})$ must converge. Let $\mathbf{r}^t = (r_1^t, ..., r_N^t)$ denote the clients' throughput sequence corresponding to the evolution of f over time. We will later show in Theorem 3 that any equilibrium outcome of WFRA converges to the same point. Hence, any converging subsequence of $\mathbf{r}$ converges to the same point. Now, leveraging the fact that $f(.)$ is a continuous function, all r_is are bounded, and $\mathbf{r}$ is compact, this results in the convergence of the $\mathbf{r}^t$ sequence to an equilibrium *i.e.*, $\mathbf{r}^t \rightarrow \mathbf{r}^{eq} = (r_1^{eq}, ..., r_N^{eq})$. Since $\forall i \in \mathbf{N}$, r_i converges to an equilibrium (r_i^{eq}), and we allow only one of the BSs change its $\lambda_{i,j}$s, therefore the changes on $\lambda_{i,j}$s must converge to zero, *i.e.*, $\lambda_{i,j} \rightarrow \lambda_{i,j}^{eq}$. This also results in the convergence of the water fill levels at each BS, *i.e.*, $\theta_j \rightarrow \theta_j^{eq}$. ■

B. Optimality and Other Properties of the Equilibria

In Theorem 2, we proved that as BSs autonomously execute WFRA, the system converges to an equilibrium. In this section, we first derive some useful properties of the equilibria and prove that the optimal solution ($\mathbf{r}^{\text{opt}}$) is also an equilibrium. Next we prove that any equilibrium outcome ($\mathbf{r}^{\text{eq}}$) is also optimal. Finally, we show that while $\mathbf{r}^{\text{opt}}$ is unique, there could be infinitely many resource allocations (*i.e.*, $\lambda_{i,j}$s) that can realize it. We start by analyzing the equilibria properties.

Lemma 1 *Consider an equilibrium outcome of WFRA. Let r_i^{eq}, θ_j^{eq}, and $\lambda_{i,j}^{eq}$, denote the corresponding throughput of client i, water fill level at BS j, and the fraction of time allocated to client i by BS j, respectively. Then*

(*I*) $\frac{\omega_i R_{i,j}}{(r_i^{eq})^\alpha} \leq \frac{1}{(\theta_j^{eq})^\alpha} \quad \forall i \in \mathbf{N}, j \in \mathbf{M}$

(*II*) $\sum_{i=1}^{N} \lambda_{i,j}^{eq} = 1 \quad \forall j \in \mathbf{M}$

(*III*) $\sum_{i=1}^{N} \omega_i (r_i^{eq})^{1-\alpha} = \sum_{j=1}^{M} \frac{1}{(\theta_j^{eq})^\alpha}$

(*IV*) *The optimal outcome is also an equilibrium. Hence, we can re-write properties (I)-(III) by replacing all r_i^{eq}, θ_j^{eq}, and $\lambda_{i,j}^{eq}$ with $r_i^{opt}, \theta_j^{opt}$, and $\lambda_{i,j}^{opt}$, respectively.*

Proof: We use the properties in Theorem 1 as well as contradiction methods to derive all the properties [9]. ∎

In our next theorem, we prove that any equilibrium is the unique optimal solution to problem $\mathcal{P}_1$.

Theorem 3 *Let $\mathbf{r}^{eq} = (r_1^{eq}, ..., r_N^{eq})$ be an equilibrium and $\mathbf{r}^{opt} = (r_1^{opt}, ..., r_N^{opt})$ be an optimal solution to problem $\mathcal{P}_1$. Then, $\mathbf{r}^{eq} = \mathbf{r}^{opt}$.*

Proof: Leveraging Lemma 1 we can write:

$$\overbrace{\sum_i \frac{\omega_i r_i^{eq}}{(r_i^{opt})^\alpha}}^{\text{left side}} = \sum_{i,j} \frac{\omega_i R_{i,j} \lambda_{i,j}^{eq}}{(r_i^{opt})^\alpha} \overset{\text{(I)}}{\leq} \sum_{i,j} \frac{\lambda_{i,j}^{eq}}{(\theta_j^{opt})^\alpha} = \sum_j \sum_i \frac{\lambda_{i,j}^{eq}}{(\theta_j^{opt})^\alpha} \overset{\text{(II)}}{=} \sum_j \frac{1}{(\theta_j^{opt})^\alpha} \overset{\text{(III)}}{=} \overbrace{\sum_i \omega_i (r_i^{opt})^{(1-\alpha)}}^{\text{right side}} \quad (5)$$

By taking similar steps to Eq. (5) and replacing all equilibrium variables with optimal ones and vice versa, we have

$$\overbrace{\sum_i \frac{\omega_i r_i^{opt}}{(r_i^{eq})^\alpha}}^{\text{left side}} = \sum_{i,j} \frac{\omega_i R_{i,j} \lambda_{i,j}^{opt}}{(r_i^{eq})^\alpha} \overset{\text{(I)}}{\leq} ... = \overbrace{\sum_i \omega_i (r_i^{eq})^{(1-\alpha)}}^{\text{rigth side}} \quad (6)$$

By multiplying the right sides and left sides of Eqs. (5) and (6), and subtracting the left side from the right side we have

$$\sum_i \omega_i (r_i^{opt})^{1-\alpha} \sum_i \omega_i (r_i^{eq})^{1-\alpha} - \sum_i \frac{\omega_i r_i^{eq}}{(r_i^{opt})^\alpha} \sum_i \frac{\omega_i r_i^{opt}}{(r_i^{eq})^\alpha} \geq 0$$

$$\implies \sum_{k,l} \omega_k \omega_l \Big[(r_k^{opt})^{1-\alpha} (r_l^{eq})^{1-\alpha} + (r_l^{opt})^{1-\alpha} (r_k^{eq})^{1-\alpha} - \frac{r_l^{eq}}{(r_l^{opt})^\alpha} \times \frac{r_k^{opt}}{(r_k^{eq})^\alpha} - \frac{r_l^{opt}}{(r_l^{eq})^\alpha} \times \frac{r_k^{eq}}{(r_k^{opt})^\alpha} \Big] \geq 0 \implies \sum_{k,l} \Big(\omega_k \omega_l$$

$$\Big[\overbrace{(r_k^{opt} r_l^{eq} - r_l^{opt} r_k^{eq}) \Big(\frac{1}{(r_k^{opt} r_l^{eq})^\alpha} - \frac{1}{(r_l^{opt} r_k^{eq})^\alpha} \Big)}^{\text{Product of these two parentheses is always } \leq 0} \Big] \Big) \geq 0 \quad (7)$$

From linear algebra, if a and b are two positive numbers, then $(a-b) \times (\frac{1}{a^\alpha} - \frac{1}{b^\alpha})$ (where α is > 0) is always ≤ 0, with equality happening when $a = b$. Since ω_k and ω_l in Eq. (7) are two positive numbers, it follows that each of the multiplications of the parentheses in Eq. (7) is zero, *i.e.*,

$$\forall k, l \in \mathbf{N} \quad r_k^{opt} r_l^{eq} = r_l^{opt} r_k^{eq} \implies \frac{r_k^{opt}}{r_k^{eq}} = \frac{r_l^{opt}}{r_l^{eq}} = \gamma \quad (8)$$

Since the inequality in Eq. (7) turns into equality to zero, it follows that inequalities in both Eqs. (5) and (6) are also *equalities*. By taking one of these (*e.g.*, Eq. (5)) we have

$$\overbrace{\sum_i \frac{\omega_i r_i^{eq}}{(r_i^{opt})^\alpha}}^{\text{left side of Eq. (5)}} \overbrace{= \sum_i \frac{\omega_i r_i^{opt}}{\gamma (r_i^{opt})^\alpha}}^{\text{Eq. (8): } r_i^{eq} = r_i^{opt}/\gamma} = \overbrace{\sum_i \omega_i (r_i^{opt})^{1-\alpha}}^{\text{right side of Eq. (5)}} \Rightarrow \gamma = 1$$

In other words (using Eq. (8)), $\forall i \in N \; r_i^{eq} = r_i^{opt}$. ∎

Finally, while $\mathbf{r}^{\text{opt}}$ is unique, we prove that there could be infinitely many resource allocations ($\lambda_{i,j}$s) to realize it.

Proposition 1 *Consider problem $\mathcal{P}_1$ and let $\mathbf{r}^{opt}$ denote the unique optimal outcome. Then, there could be infinitely many possible resource allocations to realize $\mathbf{r}^{opt}$.*

Proof: Our proof is based on an example. Consider a HetNet topology with two BSs and two clients. Let $\omega_i = 1 \; \forall i$ and $R_{i,j} = 1 \; \forall i, j$. Then, $\lambda_{1,1} = \lambda_{2,2} = \beta$ and $\lambda_{1,2} = \lambda_{2,1} = 1 - \beta$ results in the optimal allocation (*i.e.*, $r_1 = r_2 = 1$) for any value of $\beta \in [0, 1]$ and any desired α. ∎

C. Bounds on the Objective Function and Convergence Time

In this section, we analyze the objective function ($f(\boldsymbol{\lambda})$) of problem $\mathcal{P}_1$. We first derive upper and lower bounds on $f(\boldsymbol{\lambda})$. The bounds are useful for various purposes, *e.g.*, a HetNet operator could use these bounds and the current state of the network to estimate the gap from optimality and optimize network operation as needed (*e.g.*, aggregate some of calculations across BSs with better backhauling between them). We can also use these bounds to derive upper bounds on WFRA's convergence time. We start by analyzing $f(\boldsymbol{\lambda})$:

Lemma 2 *Let $f(\boldsymbol{\lambda})$ be $\mathcal{P}_1$'s objective function. Then,*

$$f(\boldsymbol{\lambda}) \leq \begin{cases} \frac{\omega_{\max} R_{\max}^{1-\alpha}}{1-\alpha} \times N^\alpha \times M^{1-\alpha} & 0 < \alpha < 1 \\ \frac{\omega_{min}^\alpha \omega_{\max}^{1-\alpha} R_{\max}^{1-\alpha}}{1-\alpha} \times N^\alpha \times M^{1-\alpha} & \alpha > 1 \end{cases}$$

$$f(\boldsymbol{\lambda}) \geq \begin{cases} \frac{N^\alpha \omega_{\min}}{1-\alpha} \times \left(\frac{R_{\min}}{R_{\max}^{1-\alpha}} \right)^{\frac{1-\alpha}{\alpha}} & 0 < \alpha < 1 \\ \frac{N^\alpha \omega_{\max}}{1-\alpha} \times R_{\min}^{1-\alpha} & \alpha > 1 \end{cases}$$

Proof: We use the properties in Theorem 1 and Lemma 1, and Jensen inequality [15] to bound $f(\boldsymbol{\lambda})$ [9]. ■

We next proceed to bound WFRA's convergence time. However, before that we need to define a discretization factor on the time fractions (*i.e.*, $\lambda_{i,j}$s). This is due to the fact that $\lambda_{i,j}$s in our model are continuous variables, which can cause some BSs to continuously make infinitesimal adjustments to them. These adjustments converge to 0 as time goes to infinity.

In practice, operations always happen in discretized levels. For example, consider the following discretization policy:

Definition 2 *Discretization Policy: During water fill calculation by a BS j in WFRA, the time fraction allocated to the client with minimum $\frac{r_i}{\omega_i R_{i,j}}$ should increase by at least ϵ. Otherwise, the BS would not update its time fractions.*

Leveraging Lemma 2 and the above discretization policy, we can derive the following bound on WFRA's speed.

Theorem 4 *Consider a HetNet with N clients and M BSs. Let $R_{\max}$ and $R_{\min}$ denote the maximum and non-zero minimum PHY rates across all the clients and BSs. Then, WFRA's convergence time is upper bounded by*

$$\begin{cases} \frac{1}{\alpha(1-\alpha)\epsilon} \times \frac{M^2}{N} \times \frac{\omega_{\max}}{\omega_{\min}} \times (\frac{R_{\max}}{R_{\min}})^2 & 0<\alpha<1 \\ \frac{1}{(\alpha-1)\alpha\epsilon} \times \frac{M^{\alpha+1}}{N} \times \left(\frac{R_{\max}}{R_{\min}}\right)^{\alpha} & \alpha>1 \end{cases}$$

Proof: Consider $f(\boldsymbol{\lambda})$ as defined in $\mathcal{P}_1$. From Theorem 1, each time a BS adjusts its water fill level, f increases. The key idea to bound the convergence time, is to find a lower bound on f's increments. Let Δ denote the lower bound on f's increments. The convergence time is then upper bounded by the difference between maximum and minimum values of f, divided by Δ. Since we have already derived upper and lower bounds on f (Lemma 2), the key remaining step is to find Δ. To derive that, we use the following lemma:

Lemma 3 *Let $U(x)=\frac{x^{1-\alpha}}{1-\alpha}$ with $x>0$ and $\alpha>0$. Let ϵ be a small positive number. Then*

(D) $U(x)-U(x-\epsilon) \geq \epsilon\, U'(x) - \frac{\epsilon^2}{2}U''(x)$

(E) $U(x)-U(x+\epsilon) \geq -\epsilon\, U'(x) - \frac{\epsilon^2}{2}U''(x+\epsilon)$

Here, U' and U'' denote the 1^{st} and 2^{nd} derivative of U.

We use the Taylor theorem [16] and our objective function properties to derive the lemma properties [9].

We now proceed to derive a lower bound on f's increments (*i.e.*, Δ). With abuse of notation, let $\{1,2,...,k\}$ denote the set of clients with non-zero PHY rates to BS j. Let client 1 be the one with minimum $\frac{r_i}{(\omega_i R_{i,j})^{\frac{1}{\alpha}}}$ across all these clients. Hence, as BS j executes WFRA (by choosing a new $\boldsymbol{\lambda}$ vector), it increases the throughput of client 1 and we would have

$$\frac{r_1+\epsilon_1 R_{1,j}}{(\omega_1 R_{1,j})^{\frac{1}{\alpha}}} \leq \frac{r_2-\epsilon_2 R_{2,j}}{(\omega_2 R_{2,j})^{\frac{1}{\alpha}}}, ..., \frac{r_k-\epsilon_k R_{k,j}}{(\omega_k R_{k,j})^{\frac{1}{\alpha}}} \qquad (9)$$

In Eq. (9), ϵ_1 is a positive number equal to $\epsilon_2+...+\epsilon_k$. The above inequality is because as j executes WFRA, the level of the water fill at client 1 cannot exceed other clients' level. Further, since $\epsilon_1=\sum_{i=2}^{k}\epsilon_i$, from Eq. (9) we would have

$$\forall i \in \{2,...,k\} \quad \frac{\omega_1 R_{1,j}}{(r_1+\epsilon_1 R_{1,j})^{\alpha}} \geq \frac{\omega_i R_{i,j}}{(r_i-\epsilon_i R_{i,j})^{\alpha}}$$

$$\implies \epsilon_1 \frac{\omega_1 R_{1,j}}{(r_1+\epsilon_1 R_{1,j})^{\alpha}} \geq \sum_{i=2}^{k} \epsilon_i \frac{\omega_i R_{i,j}}{(r_i-\epsilon_i R_{i,j})^{\alpha}} \qquad (10)$$

We derive a lower bound on f's increments through:

$$f(\boldsymbol{\lambda})^{\text{new}} - f(\boldsymbol{\lambda}) =$$

$$\omega_1(U(r_1+\epsilon_1 R_{1,j})-U(r_1)) + \sum_{i=2}^{k}\omega_i\Big(U(r_i-\epsilon_i R_{i,j})-U(r_i)\Big)$$

$$\overset{(D),(E)}{\geq} \omega_1\Big[\epsilon_1 R_{1,j}U'(r_1+\epsilon_1 R_{1,j}) - \frac{(\epsilon_1 R_{1,j})^2}{2}U''(r_1+\epsilon_1 R_{1,j})\Big] -$$

$$\sum_{i=2}^{k}\omega_i\Big[\epsilon_i R_{i,j}U'(r_i-\epsilon_i R_{i,j}) + \frac{(\epsilon_i R_{i,j})^2}{2}U''(r_i)\Big] =$$

$$\epsilon_1\omega_1 R_{1,j}U'(r_1+\epsilon_1 R_{1,j}) - \sum_{i=2}^{k}\epsilon_i\omega_i R_{i,j}U'(r_i-\epsilon_i R_{i,j}) -$$

$$\frac{\omega_1(\epsilon_1 R_{1,j})^2}{2}U''(r_1+\epsilon_1 R_{1,j}) - \sum_{i=2}^{k}\omega_i\frac{(\epsilon_i R_{i,j})^2}{2}U''(r_i) =$$

$$\overbrace{\epsilon_1\frac{\omega_1 R_{1,j}}{(r_1+\epsilon_1 R_{1,j})^{\alpha}} - \sum_{i=2}^{k}\epsilon_i\frac{\omega_i R_{i,j}}{(r_i-\epsilon_i R_{i,j})^{\alpha}}}^{\text{from Eq. (10) this term is } \geq 0} + \sum_{i=2}^{k}\frac{\alpha\omega_i(\epsilon_i R_{i,j})^2}{2}\frac{r_i^{-\alpha}}{r_i}$$

$$+\frac{\alpha\omega_1(\epsilon_1 R_{1,j})^2}{2}(r_1+\epsilon_1 R_{1,j})^{-\alpha-1} > \frac{\alpha\omega_1(\epsilon_1 R_{1,j})^2}{2}r_{\min}^{-\alpha-1} \overset{(F)}{\geq}$$

$$\frac{\alpha\omega_1(\epsilon_1 R_{1,j})^2}{2}\Big(\frac{N}{MR_{\max}}\Big)^{\alpha+1} \geq \frac{\alpha\omega_{\min}\epsilon^2 R_{\min}^2}{2\times R_{\max}^{\alpha+1}}\Big(\frac{N}{M}\Big)^{\alpha+1} = \Delta$$

In (F) we use the property that the minimum throughput (r_i) that a client can get is $\leq \frac{MR_{\max}}{N}$. This is because $Nr_{\min} \leq \sum r_i \leq MR_{\max}$. The upper bound on convergence time follows by dividing the difference between maximum and minimum values of f by Δ. To simplify the presentation of results, we assume that for $\alpha>1$ the upper bound on f is zero (note that for these α values, f is always a negative function). Similarly, for $0<\alpha<1$ we assume that the lower bound on f is zero (for these α values, f is always positive). We can then derive the bounds in Theorem 4 by making these simplifying assumptions and leveraging Δ, and Lemma 2. ■

D. Algorithm Performance for $\alpha=0$, 1, and ∞

Case 1: $\alpha=0$. This value of α corresponds to throughput maximization without any fairness constraint. Here, $U(r_i)=r_i$ and problem P_1 can be **optimally** solved in a **single step** if each BS j gives all of its time resources to the associated client with maximum $\omega_i R_{i,j}$. However, the optimal solution is **not necessarily unique**. For example, consider a single BS with two associated clients, each with rate R to the BS and unit weight. There are infinitely many optimal allocations, *e.g.*, ($r_1=r_2=\frac{R}{2}$), ($r_1=R$, $r_2=0$), and ($r_1=0$, $r_2=R$).

Case 2: $\alpha \to 1$**.** This value of α corresponds to proportional fairness. In particular, for $\alpha = 1$ the utility function definition in $\mathcal{P}_1$ is commonly changed to $U(r_i) = \log(r_i)$. We prove the convergence and optimality of WFRA (with $\alpha = 1$) in [8].

Case 3: $\alpha \to \infty$**.** Here, the objective function in $\mathcal{P}_1$ is commonly changed to a max-min function. In [17], we studied the max-min fair resource allocation problem and presented a slightly different water filling algorithm that converges to an equilibrium. However, we proved that the resulting equilibrium is not necessarily optimal. We can apply the same methodology to prove the convergence of WFRA (as $\alpha \to \infty$), however, the outcome is still not necessarily optimal. For instance, consider the example given in [17]: Assume 2 clients with $\omega_i = 1$, and 2 BSs. Let $R_{1,1} = 1$, $R_{1,2} = 2$, $R_{2,1} = 4$, and $R_{2,2} = 3$. Then, $\lambda_{1,1} = 1$, $\lambda_{1,2} = 0.4$, $\lambda_{2,1} = 0$, and $\lambda_{2,2} = 0.6$ is an equilibrium, which results in $r_1 = r_2 = 1.8$. However, the optimal outcome is $r_1 = r_2 = 2.4$, which can be achieved by $\lambda_{1,1} = 0.4$, $\lambda_{1,2} = 1$, $\lambda_{2,1} = 0.6$, and $\lambda_{2,2} = 0$.

IV. Performance Evaluation

In this section, we present the results of extensive simulations to characterize the performance of WFRA. Discussion and comparison against DComp (an alternative distributed algorithm based on dual decomposition) are presented in [9].

Simulation Setup. We used a comprehensive multi-RAT system level simulator to conduct simulations with three access technologies: (i) LTE, (ii) WiFi, and (iii) 5G New Radio (NR). Each RAT is calibrated with industrial benchmarks [18] to ensure its conformance to the 3GPP standard specifications. Fig. 4 summarizes some of the key simulator parameters.

Name and description of common parts across all three RATs	
Simulation duration	100 seconds
Channel model	Urban street canyon
Traffic model	Full buffer
Number of clients	5/10/15 per sector (i.e., total number of 105/210/315 clients)
Link adaptation	Ideal MCS feedback every 10 ms
Doppler	3 km/hr environmental mobility
MAC level packet size	500 bytes
LTE parameters: 2 GHz band, 20 MHz bandwidth; DL FDD mode; 44 dBm Tx power; up to 64 QAM modulation	
NR parameters: 6 GHz band, 20 MHz bandwidth; DL FDD mode; 30 dBm Tx power; up to 256 QAM modulation	
WiFi parameters: 802.11n, 5 GHz band, 20 MHz bandwidth; 11 channels, RTS/CTS disabled, 4 ms TxOP, DL mode; 20 dBm Tx power; up to 64 QAM modulation	

Fig. 4. **Some of the key simulation paramaters.**

Fig. 5(a) shows our simulation topology, which consists of 7 LTE macrocells each with three sectors. We randomly drop a single WiFi, a single 5G NR small cell, and a varying number of clients in each sector. Each of our clients has a weight (ω_i) of 1 and has three RATs (WiFi, LTE, and 5G NR). We associate each client's RAT to the corresponding BS that provides the highest downlink SNR. We generate channels between each client and all the BSs according to the 3GPP specifications [18]. Channel variations in our simulator are captured through a Doppler mobility parameter, which creates slow variation (fading) around an average $R_{i,j}^{\text{avg}}$ between every client and BS. Each of our data points is an average of 100 simulation trials. We implement the following algorithms:

Conv. This scheme combines traffic across multiple RATs without changing the Conventional scheduling algorithm on each BS. In particular, we use a round-robin (RR) queuing mechanism at the WiFi BSs and a proportional fair scheduler [8] at the LTE/NR BSs. In the RR scheduler, the BS maintains a different queue for each client and sequentially serves a single packet from each queue at every round.

OMMA. Opportunistic Multi-MAC aggregation [19] is a delay-equalizing resource allocation algorithm, which splits traffic across multiple RATs so as to equalize the packet delay across the RATs, and hence minimize the maximum delay per client. OMMA does not account for fairness among the clients, but similar to WFRA only uses local information at each BS.

WFRA. We use our algorithm to determine the number of packets that should be served from each queue in WiFi and the time resources that should be dedicated to each queue (client) in LTE/NR. We let each BS change its parameters every 10 ms with 0.5 probability (*i.e.*, $\forall j \;\; p_j = 0.5$). We also set the discretization factor (ϵ) equal to 0.05. For the initial allocation, each BS randomly allocates its time to its clients. Note that WFRA as presented in Section II-D does not account for various types of overhead (*e.g.*, WiFi contention). To address the issue, we introduce the notion of effective rate (R^{eff}) and replace all $R_{i,j}$s in WFRA with $R_{i,j}^{\text{eff}}$s. To calculate $R_{i,j}^{\text{eff}}$, each BS keeps track of the *total time* spent in successfully transmitting the last 5 packets of each client. It then divides the total number of data bytes across these packets by the total time spent in sending them, to derive $R_{i,j}^{\text{eff}}$. We average over 5 packets to take into account the impact of channel fluctuations.

Throughput-Fairness Characteristics. Figs. 5(b) and 5(c) depict the median and cell-edge clients' throughput across all the schemes. Cell-edge client is defined as the client that is at the 5th percentile of the throughput CDF (Cumulative Distribution Function) across all the clients. For WFRA, we conduct simulations for three different α values: 0.5, 0.9, and 2. Note that Conv and OMMA do not depend on α. We observe that WFRA provides different throughput-fairness tradeoffs that depend on α. For example, WFRA's median and cell-edge throughput are 48 and 6 Mbps for $\alpha = 0.5$, 39 and 15 Mbps for $\alpha = 0.9$, and 34 and 18 Mbps for $\alpha = 2$, respectively. For larger values of α, WFRA approaches the max-min fairness, which allocates resources so that clients achieve a more equal throughput. In contrast, for smaller values of α, WFRA approaches a throughput maximization objective, with less emphasis on fairness.

We next quantify the closeness of each algorithm's outcome to an **ideal α-fair allocation**. We take the following steps to identify the ideal throughput (r_i^{ideal}) for each client i and for a desired value of α. We set the Doppler to 0 and let WFRA converge to the optimal ideal allocation. With Doppler set

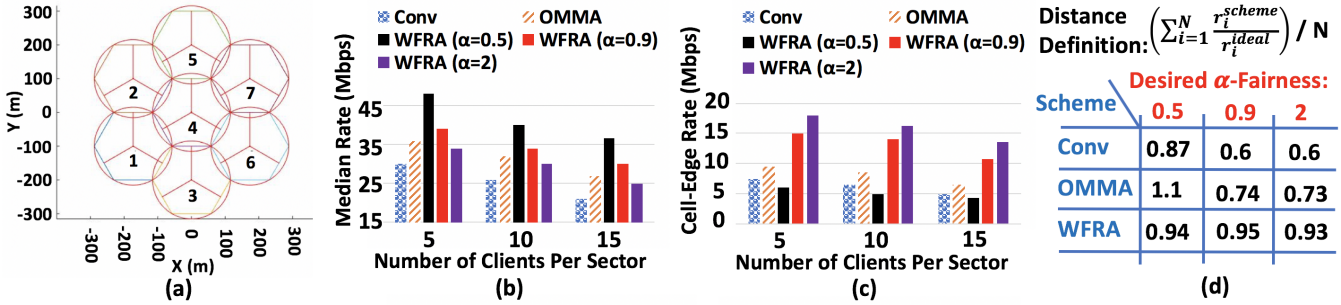

Fig. 5. **(a): Simulation topology. A wrap around implementation ensures that the distribution of BS distances to clients remains the same across all the clients; (b): Median (middle) client total throughput; (c): Cell-edge (5th percentile) client total throughput; (d): Distance to the ideal fairness.**

to 0, there would be no channel variations or packet losses (*i.e.*, ideal conditions). We next set the Doppler to 3 km/hr and measure the vector of throughput values ($\mathbf{r}$) across all the clients for each scheme. Finally, we calculate $\sum_{i=1}^{N} \frac{r_i^{\text{scheme}}}{r_i^{\text{ideal}}}$ and divide it by N (total number of clients) to quantify the distance of each scheme to the ideal α-fair allocation. This metric allows us to compare two n-dimensional vectors. A value close to 1 means an outcome is close to the ideal outcome. Fig. 5(d) depicts the distance of each scheme to the ideal α-fair allocations with 15 clients/sector (we observed similar results with 5/10 clients). We observe that even with channel dynamics, WFRA gets very close to the ideal allocations.

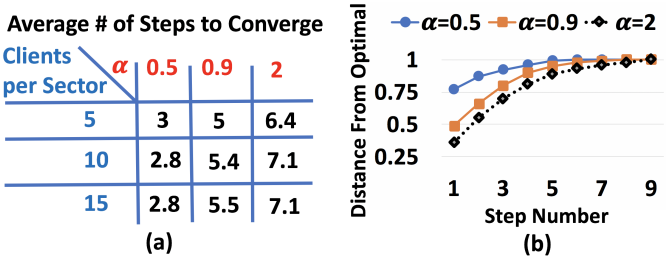

Fig. 6. **WFRA's convergence time (a) and evolution (b) properties.**

Convergence Time. Figs. 6(a) and 6(b) depict WFRA's convergence time properties. Here, we set the Doppler to 0 to avoid channel dynamics. Note that if channel conditions are continuously changing, WFRA would not converge as it would continuously adapt to system dynamics. Fig. 6(a) shows the average number of required steps for WFRA to converge. There are two observations: (i): WFRA converges in a small number of steps, and (ii) there is a small change in the convergence time for different client numbers. This is due to WFRA's water filling operation, in which a BS finds the $\lambda_{i,j}$s to all of its clients in one shot and also because irrespective of the number of clients, half of the BSs adjust their $\lambda_{i,j}$ at each scheduling interval ($p_j = 0.5$).

Fig. 6(b) shows sample evolutions of the previously defined distance metric ($(\sum_i \frac{r_i^t}{r_i^{\text{opt}}})/N$) as function of step number t (*i.e.*, scheduling interval) in WFRA. The samples correspond to simulation realizations with 15 clients per sector. We observe that irrespective of the value of α, the distance metric is more than 0.8 within 4 steps, which shows that WFRA can get very close to the optimal outcome within a few iterations.

V. Conclusion

We studied the theoretical aspects of generic α-fair throughput aggregation in HetNets. We showed that if a BS wants to unilaterally maximize the system objective, it should allocate its resources according to a water filling operation. We used this observation to design a distributed algorithm that achieves α-fairness. We analyzed several theoretical aspects of the algorithm and showed that our algorithm can quickly adapt to system dynamics and realize the desired fairness outcome.

VI. Acknowledgements

This research was supported in part by NSF grants CNS-1910517 and CNS-1942305.

References

[1] "5G Spectrum," https://www.gsma.com/spectrum/5g-spectrum-guide/
[2] J. Zhu, S. Seo, S. Kanugovi, and S. Peng, "User-plane protocols for multiple access management service," in *IETF Internet Area Working Group (INTRAE)*, 2019.
[3] B. Bojovic, L. Giupponi, Z. Ali, and M. Miozzo, "Evaluating unlicensed LTE technologies: LAA vs LTE-U," in *IEEE Access*, 2019.
[4] P. Nuggehalli, "LTE-WLAN aggregation [industry perspectives]," in *IEEE Wireless Communications Magazine*, 2016.
[5] 3GPP, "Study on NR-based access to unlicensed spectrum," in *3GPP Technical Report, 38.889 (Release 16)*, 2019.
[6] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," in *IEEE/ACM Transactions on Networking*, 2000.
[7] D. Ibarra, N. Desai, and I. Demirkol, "Software-based implementation of LTE/Wi-Fi aggregation and its impact on higher layer protocols," in *Proceedings of IEEE ICC*, 2018.
[8] E. Aryafar, A. Keshavarz-Haddad, and C. Joe-Wong, "Proportional fair RAT aggregation in hetnets," in *Proceedings of ITC 31*, 2019.
[9] http://web.cecs.pdx.edu/~aryafare/2020-TR-WiOpt.pdf
[10] W. Wang, X. Liu, J. Vicente, and P. Mohapatra, "Integration gain of heterogeneous WiFi/WiMAX networks," in *IEEE Transactions on Mobile Computing*, 2011.
[11] S. Shakkottai, E. Altman, and A. Kumar, "Multihoming of users to access points in WLANs: a population game perspective," in *IEEE Journal on Selected Areas in Communication*, 2009.
[12] "CVX for convex programming," http://cvxr.com/cvx/
[13] FCC, "2016 broadband progress report," January 2016.
[14] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," in *IEEE Journal on Selected Areas in Communications*, 2006.
[15] Jensen Inequality, https://en.wikipedia.org/wiki/Jensen%27s_inequality
[16] "Taylor Theorem," https://en.wikipedia.org/wiki/Taylor%27s_theorem
[17] E. Aryafar, A. Keshavarz-Haddad, C. Joe-Wong, and M. Chiang, "Max-min fair resource allocation in hetnets: Distributed algorithms and hybrid architecture," in *Proceedings of IEEE ICDCS*, 2017.
[18] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," in *3GPP Technical Report, 38.901 (Release 15)*, 2018.
[19] S. Goyal, T. B. Le, A. Chincholi, T. Elkourdi, and A. Demir, "On the packet allocation of multi-band aggregation wireless networks," in *Springer Wireless Networks Journal*, 2018.