On the Capacity of HetNets

S.V. Hanly and Philip A. Whiting
Department of Engineering, Macquarie University
North Ryde, NSW 2109, Australia
Email: stephen.hanly,philip.whiting@mq.edu.au

***Abstract*—Heterogeneous wireless networks provide a means to increase network capacity by introducing small cells and adopting a layered architecture. These networks allocate resources flexibly through a combination of time sharing and cell range expansion/contraction allowing a wide range of possible schedulers. In this paper we define capacity of such networks in terms of the maximum number of downloads per second which can be achieved for a given offered traffic density. In a simplified model we show that this capacity is determined via the solution to a continuous linear program (LP). If the solution is smaller than 1 then we show that there is a scheduler such that the number of mobiles in the network has ergodic properties with finite limiting moments. On the other hand if the solution is greater than 1 then we show that no such scheduler exists.**

I. Introduction

HetNets have been proposed as a means to increase the capacity of wireless networks by introducing short range pico cells into the existing coverage area of the large cells - the macro cells. Since the pico cells can operate at high rates and several such cells can operate in the same macro cell - significant increase in user data throughput is possible. This is particulary the case if traffic is spatially concentrated into small parts of the coverage area known as traffic "hot-spots".

A brief description of HetNet operation is as follows. First time is split using small duration frames known as Almost Blanking Subframes (ABS), [10]. These are used to time share between macro cell transmissions and pico cell transmissions. In this way interference between the macro cell(s) and the pico cells is avoided. An additional level of flexibility is provided by allowing the pico cells to adjust their areas of coverage, that is to expand and serve more mobiles or to contract to serve fewer mobiles but at potentially higher rates. This procedure is known as Cell Range Expansion (CRE). Such networks therefore have considerable flexibility in matching the traffic demands made on them against the radio resources available.

A central question for HetNet deployment is its capacity measured in user arrivals per second, for a given infrastructure and traffic profile (probability density of mobile locations). Determination of this capacity is somewhat fraught as arbitrary scheduling can be performed in response to a given offered traffic load. A number of papers have appeared which provide some bearing on this question, see for example [2], [7], [5], [4] and [3]. None of these papers address performance in the dynamic setting we treat here.

In this paper we consider the case where mobiles arrive to a fixed location, have a single file to down load and then depart once the file has been down loaded. The intensity of the arrivals varies depending on the location. This traffic is offered to a simplified network as described in [3]. Within this set up we show that the capacity emerges as the solution to a continuous linear program (LP), with integration over a mean measure (traffic density) but otherwise along the lines described in [3]. The solution to the above capacity problem allows network operators to make some inferences as to the number and location of resources needed to meet the traffic loads which they are encountering or anticipate. The simplified model we present is not sufficient for this task but clearly the results can be generalised to incorporate far more realistic scenarios and these can be expected to yield real insight into operating networks. HetNet planning has also been addressed in [8].

The remainder of the paper can be summarised as follows. Section II presents our model for the HetNet, along the lines of [3]. As we mentioned, we leave aside for the moment the generalisation to more complex instances of such networks so that the main arguments are given on a transparent footing. Section III shows that if the (stationary) offered traffic is

below capacity, then there exists a schedule which has ergodic properties and user download delays have uniformly bounded expectation. Section IV shows that if the offered traffic is above capacity, then the number of mobiles within the network builds at a linear rate almost surely, under any schedule which serves all mobiles within finite time. Finally in Section V we present some brief conclusions and suggestions for further work.

II. Model and Preliminaries

A. Model

We consider a HetNet consisting of a single macro cell and $L \geq 1$ pico cells. S denotes the (bounded) coverage area of the macro cell whilst $C_\ell \subset S, \ell = 1, \cdots, L$ denote the respective coverage areas of the pico cell. The pico cell coverage areas are disjoint so that mobile transmissions can be supported from the macro cell and whichever pico cell coverage area they are in, if any.

We suppose that users arrive as a Poisson stream with net arrival rate λ_S and wish to receive a single file of length D bits to be downloaded by the network. The locations of the mobiles are chosen independently at random according to a continuous density $\eta\,(d\xi)$ with support on S and bounded uniformly away from 0. Mobiles remain fixed at their initial location until they obtain their file. Hence in unit time, the expected number of arrivals at the vicinity of a point ξ in the macro cell coverage area S is given by $\lambda\,(d\xi) = \lambda_S\ \eta\,(d\xi)$.

The raw transmission rates of the users are determined by their fixed location ξ. In case they are out of pico cell coverage this is taken to be $S_0(\xi)$. If there is coverage from exactly one of the L pico cells, i.e. $\xi \in C_\ell$ then the rate pair $(R_\ell(\xi), S_\ell(\xi))$ is obtainable. If a mobile is out of pico cell coverage, we assign a pico cell index $\ell = 0$. We assume that $0 < R_{\min} \leq R_\ell(\xi) \leq R_{\max}, \ell = 1, \cdots, L$ for the pico cell rates and that there are corresponding bounds for the macro cell rates, $S_{\min}, S_{\max}$ for $S_\ell(\xi), \ell = 0, \cdots, L$, the macro cell rates. Define $\underline{R} \doteq \min\{R_{\min}, S_{\min}\} > 0$ and $\overline{R}$ similarly. Finally define $\rho_\ell(\xi) \doteq R_\ell(\xi)/S_\ell(\xi) > 0, \ell = 1, \cdots, L$ for pico cell users, which is bounded away from 0. Finally we will suppose that the macro cell rates depend continuously on location as do the pico cell rates restricted to their coverage area.

We now define the possible scheduling policies that we will consider. Time is slotted, and scheduling decisions are made at each timeslot. The first time that a file can possibly be scheduled is in the first full timeslot after its arrival. At each timeslot, n, the scheduler chooses a fraction of the timeslot, f_n, that it will devote to the pico base stations, and the remaining fraction, $1 - f_n$, is allocated to the macro base station (BS). Backlogged users can receive part of their file from the macro BS, and part of their file from the nearby pico BS. Each base station transmits to one user at a time. A scheduler, π, chooses f_n, and also how much time each backlogged user receives, in each timeslot n. We will only consider scheduling policies which clear users - those schedules which transmit all user files within finite time. We term such schedules clearing schedules. Schedules which serve users according to order of arrival during macro time and during pico time are clearing schedules.

For a user which arrives at point ξ let $0 \leq x_\ell^\pi(\xi) \leq D$ be the number of bits sent by pico ℓ and the remainder of the file $y_\ell^\pi(\xi) = D - x_\ell^\pi(\xi)$ from the macro, for users with locations $\xi \in C_\ell$.

B. A Continuous Linear Program

Our aim is to determine τ^* the least possible transmission time (time macro is transmitting plus time some pico is transmitting) needed to clear users from the network per unit time. Obviously τ^* depends on $\eta(d\xi)$ and determines the maximum possible arrival rate λ_S which the network can support.

Consider the case in which a set of users are in a HetNet with their files ready for transmission. The question arises what is the shortest time for all user files to be transmitted? The solution to this problem can be obtained via a LP as shown in [3], and which we now present.

Let the time allocated to the pico cells (which can serve their users concurrently) be denoted by f seconds. Also, let $x_{\ell,n}$ and $y_{\ell,n}$ represent the amount of data (in bits) received by user $n \in \mathcal{C}_\ell = \{1, \ldots, N_\ell\}$ in pico cell ℓ from that pico cell and the macro cell, respectively. Finally let $R_{\ell,n}, S_{\ell,n}$ be the correspond-

ing rates. Then the solution can be obtained from:

$$\min \quad f + \sum_{\ell=1}^{L} \sum_{n \in \mathcal{C}_\ell} \frac{y_{\ell,n}}{S_{\ell,n}} \tag{1}$$

$$\text{sub} \quad \sum_{n \in \mathcal{C}_\ell} \frac{x_{\ell,n}}{R_{\ell,n}} \leq f \quad \forall \ell \tag{2}$$

$$x_{\ell,n} + y_{\ell,n} \geq D \quad \forall \ell, \ \forall n \in \mathcal{C}_\ell \tag{3}$$

$$f \geq 0, x_{\ell,n} \geq 0, y_{\ell,n} \geq 0 \quad \forall \ell, \ \forall n \in \mathcal{C}_\ell$$

It was shown in [3] that there is a set of constants $\rho_\ell > 0, \ell = 1, \cdots, L$, which determine the optimal solution. For users in pico cell ℓ such that

$$\frac{R_{\ell,n}}{S_{\ell,n}} > \rho_\ell$$

we set $x_{\ell,n} := D$ and if the reverse inequality holds then we set $y_{\ell,n} := D$. Only where there is equality can it occur that both $x_{\ell,n}, y_{\ell,n}$ are positive.

The above LP has a continuous analogue in which $x_{\ell,n}, y_{\ell,n}$ are replaced by $x_\ell(\xi), y_\ell(\xi) = D - x_\ell(\xi)$ where $x_\ell(\xi)$ are (Lebesgue) integrable functions. The LP becomes,

$$\min \quad f + \sum_{\ell=1}^{L} \int \frac{y_\ell(\xi)}{S_\ell(\xi)} \lambda(d\xi) \tag{4}$$

$$\text{sub} \quad \int \frac{x_\ell(\xi)}{R_\ell(\xi)} \lambda(d\xi) \leq f \quad \ell = 1, \cdots, L \tag{5}$$

By standard results for the calculus of variations, or otherwise, it can be shown that there exists $\rho_\ell, \ell = 1, \cdots, L$ such that $\sum_\ell \rho_\ell = 1$ and the optimal solution is $x_\ell(\xi) = D, \rho_\ell(\xi) > \rho_\ell$ and 0 if the reverse equality holds. In the case of equality, the choice is arbitrary since this is a measure 0 set anyway.

A natural candidate for τ^* is the optimum solution to the continuous LP (4). In what follows it will be shown that if $\tau^* < 1$ then there is a policy that stabilises the network such that the expected delay of any customer is finite.

On the other hand if the solution to the continuous LP satisfies $\tau^* > 1$, we will show that there is no scheduling policy (amongst clearing schedules) which can stabilise the network. Indeed, for all such policies, the total amount of time needed to clear the outstanding files present in the network builds up at a linear rate over time with probability 1.

Note that τ^* is an increasing, linear function of the arrival rate λ_S. Define λ_S^* to be the unique solution to the equation $\tau^*(\lambda_s) = 1$:

$$\lambda_S^* \doteq \sup \{\lambda_S > 0 : \tau^*(\lambda_S) < 1\}.$$

Given the above properties, it is natural to regard λ_S^* as the traffic capacity of the HetNet network, given the offered traffic density $\eta(d\xi)$ and rate functions $R_\ell(\xi), S_\ell(\xi), \ell = 0, \cdots, L$.

It should be noted that λ_S^* can be obtained from the continuous LP. In fact iterative algorithms can be used to determine the optimal solution similar to those described in [3] for the discrete LP (1). Hence the continuous LP can be used for network planning, taking as input estimates for $\eta(d\xi)$ and the physical rates.

III. Achievability of the LP Solution for $\tau^* < 1$

For achieveability, we will propose a scheduling policy that serves customers in bundles; we will term it a *bundling* policy. We do not claim that this policy is optimal in any sense, other than we can show that it stabilizes the system with finite expected delay per customer.

A bundling policy with parameter T marks all arrivals over a period of T slots, which we shall call a bundle. Transmission of a bundle cannot occur until the entire bundle has arrived. Bundles queue for transmission in the order in which they have arrived. The network transmits the files of the earliest stored bundle, and when it has finished, it immediately begins service of the next one queued. However, a bundle must be complete before it can be transmitted. Thus, if the buffer of bundles is empty when a bundle completes, then the base stations will be idle until the next bundle has finished being collected. In other words, if the buffer of bundles is empty when a bundle begins service, then the network will start on the next bundle of arrivals when it is finished with the former, or after T slots, whichever is the later. Bundles can accumulate over time, and are serviced in order of their arrival. In any period of servicing a bundle, the files being served all arrived over an interval of length T slots.

A. Preliminary Results

Under our bundling policies the amount of pico time allocated will be a function of location only. For future use we note the following results.

First, given $T > 0$, the empirical point measure determined by the arrivals during some interval of length T converges to the mean measure almost surely,

$$\lambda_T(\omega, d\xi) \Rightarrow \lambda(d\xi) \tag{6}$$

see for example [11] or [1].

Next let $v : S \to \mathbb{R}_+$ be non-negative, bounded and measurable. The following theorem is an immediate consequence of the SLLN,

Lemma 1. *Let $n = 1, \cdots, N_T$ be the random number of arrivals in $[0, T]$, with the nth arrival being at location $\xi_n(\omega)$ then*

$$\frac{1}{T}\sum_{n=1}^{N_T} v(\xi_n(\omega)) \to \int_S v(\xi)\lambda(d\xi) \tag{7}$$

almost surely and in $\mathcal{L}_1$.

Let $v_T^{(\ell)}, \ell = 1, \cdots, L$ be L sequences of non-negative random variables such that $v_T^{(\ell)} \to v^{(\ell)}$ (a constant) almost surely and in $\mathcal{L}_1$. Define,

$$h_T = \max_{\ell=1,\cdots,L} v_T^{(\ell)} \tag{8}$$

then $h_T \to \max_{\ell=1,\cdots,L} v^{(\ell)}$ almost surely and in $\mathcal{L}_1$.

Proof

The first part is an immediate consequence of the Strong Law for Large Numbers (SLLN), [13].

For the second part, it follows from [13] Theorem 13.7 that the sequences $v_T^{(\ell)}, \ell = 1, \cdots, L$ are uniformly integrable (UI). Hence given $\varepsilon > 0$ there is $K_\varepsilon > 0$ such that

$$\mathbb{E}\left(v_T^{(\ell)}; v_T^{(\ell)} \geq K_\varepsilon\right) < \frac{\varepsilon}{L}$$

Clearly,

$$\mathbb{E}\left(h_T; h_T \geq LK_\varepsilon\right) \leq \sum_{\ell=1}^{L} \mathbb{E}\left(v_T^{(\ell)}; v_T^{(\ell)} \geq K_\varepsilon\right) < \varepsilon \tag{9}$$

and so h_T is UI. But $h_T \to \max v^{(\ell)}$ almost surely by continuity of $\max$. Convergence in $\mathcal{L}_1$ follows from [13] Theorem 13.7. □

B. Rate Ratio Bundling Policies

Bundling policies were outlined earlier. A rate ratio bundling policy is defined by a vector $\mathbf{a} \in \mathbb{R}^L, a_\ell \in [\underline{\rho}_\ell, \overline{\rho}_\ell]$. It works by assigning pico time according to $\mathbf{a}$, that is mobiles in pico cell ℓ at location ξ are transmitted only using pico time if

$$a_\ell < \frac{R_\ell(\xi)}{S_\ell(\xi)}$$

all other mobiles are transmitted using only macro cell time. Given this rule for BS association, one can then compute the time required by each mobile in the bundle (since the position-dependent bit rate is fixed) and the number of bits sent by the pico BS ℓ and macro BS, $(x_\ell^{(\mathbf{a})}(\xi), y_\ell^{(\mathbf{a})}(\xi))$, respectively, are given by:

$$(x_\ell^{(\mathbf{a})}(\xi), y_\ell^{(\mathbf{a})}(\xi)) = \begin{cases} (D, 0) & a_\ell < \frac{R_\ell(\xi)}{S_\ell(\xi)} \\ (0, D) & a_\ell > \frac{R_\ell(\xi)}{S_\ell(\xi)} \end{cases}$$

The total pico cell time is computed for each pico cell ℓ and the policy allocates pico cell time according to the maximum.

Note that we haven't specified how time is allocated within each slot. This is actually immaterial to the result, but for concreteness, we can assume that the macro-associated mobiles are served first, in an arbitrary order, and then the pico-associated users are served. Again, the order of service within each picocell is arbitrary. So initial timeslots will be devoted entirely to the macro BS, and later timeslots will be allocated to the pico BSs, which run concurrently. An alternative approach would have been to split each timeslot between the macro and picos; however, the precise schedule within the bundle does not alter the total time taken.

It follows that the time to clear bundle b is given as $T\tau_{b,T}^{(\mathbf{a})}$ slots.

$$T\tau_{b,T}^{(\mathbf{a})} \doteq T\left(f_{b,T}^{(\mathbf{a})} + \sum_{\ell=0}^{L}\int_S \frac{y_\ell^{(\mathbf{a})}(\xi)}{S_\ell(\xi)}\lambda_T^{(b)}(d\xi)\right) \tag{10}$$

where $f_{b,T}^{(\mathbf{a})}$ is the minimal pico time per arrival slot used and must satisfy,

$$f_{b,T}^{(\mathbf{a})} \geq \int_S \frac{x_\ell^{(\mathbf{a})}(\xi)}{R_\ell(\xi)}\lambda_T^{(b)}(d\xi), \quad \ell = 1, \cdots, L$$

and we neglect wasted time due to rounding up the total transmission time to a whole number of slots.

$\lambda_T^{(b)}$ is the empirical point measure on S induced by the realisation of bundle b.

Define

$$\tau^{(\mathbf{a})} \doteq \sum_{\ell=0}^{L} \int_S \frac{y_\ell^{(\mathbf{a})}(\xi)}{S_\ell(\xi)} \lambda\left(d\xi\right) + f^{(\mathbf{a})} \tag{11}$$

where $f^{(\mathbf{a})} \doteq \max_\ell \int_S \frac{x_\ell^{(\mathbf{a})}(\xi)}{R_\ell(\xi)} \lambda\left(d\xi\right)$.

An important case is when $\mathbf{a}$ is chosen according to the solution of the continuous LP so that $\tau^{(\mathbf{a})} = \tau^*$.

Since $\frac{y_\ell^{(\mathbf{a})}(\xi)}{S_\ell(\xi)}, \ell = 0, \cdots, L$, $\frac{x_\ell^{(\mathbf{a})}(\xi)}{R_\ell(\xi)}, \ell = 1, \cdots, L$ are bounded, measurable and non-negative, we may apply the first part of Lemma 1 to obtain that for an arbitrary bundle b, and for an arbitrary vector of rate ratios, a,

$$\lim_T \int_S \frac{y_\ell^{(\mathbf{a})}(\xi)}{S_\ell(\xi)} \lambda_T^{(b)}(d\xi) \to \int_S \frac{y_\ell^{(\mathbf{a})}(\xi)}{S_\ell(\xi)} \lambda(d\xi) =: y_\ell^{(a)} \tag{12}$$

almost surely and in $\mathcal{L}_1$. Clearly the same holds for $x_\ell^{(\mathbf{a})}(\xi)/R_\ell(\xi), \ell = 1, \cdots, L$ with corresponding limit $x_\ell^{(a)}$.

The second part of Lemma 1 shows that for arbitrary b,

$$f_{b,T}^{(\mathbf{a})} \to f^{(\mathbf{a})}$$

almost surely and in $\mathcal{L}_1$. We summarise the above as a lemma,

Lemma 2. *For any bundle b, $\tau_{b,T}^{(\mathbf{a})} \to \tau^{(\mathbf{a})}$ almost surely and in $\mathcal{L}_1$ as $T \to \infty$.*

This lemma holds no matter what the value of $\tau^{(\mathbf{a})}$ is, including the case $\tau^{(\mathbf{a})} > 1$. Lemma 2 will also be used in demonstrating the converse result.

Now consider the case when we fix a to be the vector of rate ratios from the solution of the continuous LP. Let us denote the time $\tau_{b,T}^{(\mathbf{a})}$, in this case, by $\tau_{b,T}^{(LP)}$. Note that this time is not the optimal time for the random locations of the mobiles in the bundle, but, nevertheless, we have convergence in $\mathcal{L}_1$ to $\tau^{(\mathbf{a})}$, and in this case, $\tau^{(\mathbf{a})} = \tau^*$, the optimal value of the continuous LP. Since the convergence is in $\mathcal{L}_1$ we obtain

$$\lim_T \mathbb{E}\left(\tau_{b,T}^{(LP)}\right) = f^* + \sum_{\ell=0}^{L} y_\ell^* = \tau^*$$

Hence if $\tau^* < 1$ for T sufficiently large, the expected time per slot needed to clear an arbitrary bundle is smaller than 1. (Note that we can always choose T sufficiently large to deal with the maximum one slot rounding up needed for any bundle so that the above inequality still holds.)

We now show that this implies the network itself is stable under the LP bundling policy, in the sense that the expected waiting time for a user to be transmitted is uniformly bounded. The conclusion follows from standard arguments for the $D/G/1$ queue.

Consider the delays before transmitting the files bundled at epochs $nT_B, n \in \mathbb{N}$ where T_B is taken sufficiently large so that, $\mathbb{E}\left(\tau_{b,T}^{(LP)}\right) < 1$. At each epoch, let W_n be the total number of slots needed to clear the remaining files, apart from the current bundle. We assume each bundle is rounded up to a whole number of slots. It follows that the sequence of waiting times $W_n, n \in \mathbb{N}$ corresponds to an instance of a $D/G/1$ queue with arrivals at times nT_B. G corresponds to the random number of slots H_n^B needed to transmit bundle n. Let ρ_B be the average number of slots per bundle needed for the bundling interval with T_B chosen sufficiently large that $\rho_B < 1$.

Define $S_n \doteq \sum_{k=1}^n H_k^B - T_B n$, and $H_k^B, k = 1, \cdots, \ell$ are the iid number of slots to clear bundle k. Since $H_1^B \leq CX$ for some constant C with $X \sim \mathcal{P}(\lambda_S T_B)$ a Poisson variate, it follows that $\mathbb{E}\left(\left(H_1^B\right)^N\right) < \infty, \forall N \in \mathbb{N}$.

We now show that the limiting distribution for W_n exists and has finite expectation. By time reversal, see [6], W_n has the same distribution as $\max\{0, S_1, \cdots, S_n\}$. The limit distribution exists and has finite expectation determined by

$$\mathbb{E}\left(W_\infty\right) = \sum_{n=1}^{\infty} \frac{1}{n} \mathbb{E}\left(S_n^+\right) \tag{13}$$

in case that the sum on the RHS of (13) is finite, see [9]. It is enough to show that $\mathbb{E}\left(S_n^+\right) \leq K/n$ for some constant $K > 0$ as this implies $\sum_{n=1}^{\infty} \frac{1}{n} \mathbb{E}\left(S_n^+\right) < \infty$ and so the limit distribution W_∞ together with its expectation exists as a consequence of the submartingale convergence theorem, [13].

The following argument is standard. Let G_k^B be the centered versions of H_k^B and let G_n be the survivor

function for S_n, given any $x > 0$,

$$\begin{aligned} G_n(x) &\doteq \mathbb{P}\Big(S_n \geq x\Big) \\ &= \mathbb{P}\Big(\sum_{k=1}^{n} G_k^B \geq n\eta_B + x\Big) \end{aligned} \tag{14}$$

where $\eta_B = 1 - \rho_B > 0$. Of course the integrals are actually sums since S_n is defined on the integers. Since G_k^B are iid with finite fourth moments, Markov's inequality shows that $G_n(x) \leq (An^2)/(n\eta_B + x)^4$ for some constant $A > 0$. Since

$$\begin{aligned} \mathbb{E}\left(S_n^+\right) &= \mathbb{E}\left(S_n; S_n > 0\right) \\ &= \int_0^\infty G_n(x)dx \end{aligned}$$

we may set $K = A/(3\eta_B^3)$.

From the above we may deduce the following theorem,

Theorem 1. *If the solution to the continuous LP, satisfies $\tau^* < 1$ determined via (4) then there is a number of slots $T_B \in \mathbb{N}$ and a constant $U > 0$ such that the expected number N_n of mobiles still in the network at slot n satisfies*

$$\mathbb{E}_{\pi_B}\left(N_n\right) < U, \ \forall n \in \mathbb{N} \tag{15}$$

and π_B schedules according to LP bundling at intervals T_B.

IV. Converse

In the converse part we will show that if the $\inf$ of the solutions to (4) is bounded to be strictly greater than 1 then for any clearing schedule π the total residual transmission time increases to infinity. To be more specific, given a scheduling policy π let V_T^π be the (random) total time that transmission of any of the arrivals during $[0, T]$ is taking place under schedule π. In other words, the total time the network is actively transmitting the file of at least one of the users which arrived in $[0, T]$. This time is not assumed to be accrued continuously: There can be interspersed services of later arrivals, but we do not count the service of the later arrivals in V_T^π.

For fixed T and a given set of user arrivals V_T^π is upper bounded by the time which would be taken if user files are transmitted one after the other at rate $\underline{R}$. It is lower bounded by the time taken if the files are transmitted sequentially at rate $(L+1)\overline{R}$.

We now state our main theorem.

Theorem 2. *There is a constant $\eta > 0$ such that for any clearing schedule π it holds that*

$$\liminf_T \frac{V_T^\pi(\omega)}{T} > 1 + \eta \tag{16}$$

almost surely.

Hence after time T the residual work (time to clear the remaining users) is at least ηT for all T sufficiently large. The implication is that with probability 1 and for all T sufficiently large there must be at least

$$N_T^\pi(\omega) \geq \lfloor \frac{\eta T \underline{R}}{D} \rfloor \tag{17}$$

users present at time T. This is because no clearing schedule transmits files at a rate smaller than $\underline{R}$ at any time. Thus any clearing schedule must have at least $\eta T \underline{R}$ bits to be transmitted if the remaining time left to transmit is ηT. Since there are D bits per file there must be (17) users still in the network.

It is important to remark that we are not focusing on bundling strategies in this section, and T does not correspond to a bundling interval for any particular bundling strategy. A clearing policy, π, *can* be a bundling policy, but it does not have to be. If π is a bundling policy, then T does not refer to the bundling interval of π.

The idea of the proof is that first, for any realisation over T slots we can never use less time than the corresponding LP solution, $V_T^{(LP)}$. Second it will be shown that if the optimal solution to (4) is

$$\tau^* = f^* + \sum_{\ell=1}^{L} \int_S \frac{y_\ell^*}{S_\xi} \lambda(d\xi) > 1$$

then there is a constant $\eta > 0$ such that

$$\liminf_T \frac{V_T^{(LP)}(\omega)}{T} > 1 + \frac{\eta}{2} > 1$$

almost surely. This implies that the residual amount of work grows linearly over time.

In the following, we obtain a bound on V_T^π by considering a different system, in which all the arrivals during $[0, T]$ have instead arrived together at time 0. We are then concerned with the time taken to service

these arrivals (now all arrived at time 0) and we want to relate this time back to V_T^π in the actual system of interest. Since we are only interested in the time spent servicing arrivals during $[0, T]$, we do not consider any other arrivals in the start-at-zero system; all arrivals in that system are already there by time 0.

The performance of the start-at-zero system is identical to the performance of a bundle of duration T, as described in the first part of this paper. In particular, we can consider threshold strategies, where cell association (and hence service time) is determined by a fixed vector a of rate ratio thresholds. Let $V_T^{(a)}$ be the time taken under threshold policy a in the start-at-zero system. Then $\frac{V_T^{(a)}}{T}$ is the solution $\tau_{b,T}^{(\mathbf{a})}$ in (10) applied to start-at-zero, so by Lemma 2, we have that

$$\lim_T \frac{V_T^{(a)}}{T} = \tau^{(a)}$$

where $\tau^{(a)}$ is given in (11). But $\tau^{(a)} \geq \tau^* = 1 + \eta$, for some $\eta > 0$, independent of the choice of a. Thus, we have the following lemma:

Lemma 3. *There exists $\eta > 0$ such that for all threshold vectors a,*

$$\lim_T \frac{V_T^{(a)}}{T} \geq 1 + \eta \quad a.s.$$

The smallest amount of time for servicing the users in the start-at-zero system is $V_T^{(LP)}$, the time used if we solve the discrete LP for the given users that all start at time 0.

Lemma 4. *For all sample paths ω and for any clearing schedule π,*

$$\liminf_T \frac{V_T^{(LP)}(\omega)}{T} \leq \liminf_T \frac{V_T^\pi(\omega)}{T} \tag{18}$$

Proof

Let N_T be the number of mobiles which arrive in the interval $[0, T]$ and index the users in order of arrival, $n = 1, \cdots, N_T$. Let x_n, y_n denote the number of bits sent in the nth users pico cell (if any) and the bits sent in the macro cell. Since π is a clearing schedule, it must be the case that,

$$x_n + y_n \geq D, \forall n = 1, \cdots, N_T \tag{19}$$

Let $\xi_n, n = 1, \cdots N_T$ denote the locations of the mobiles. Also let f_T^π be the total amount of time during which at least one of the N_T mobiles was being transmitted by a pico cell. It follows that,

$$f_T^\pi \geq \sum_{n:\xi_n \in C_\ell} \frac{x_n}{R_n} \tag{20}$$

for each $\ell = 1, \cdots, L$. Finally by definition,

$$f_T^\pi + \sum_{n=1}^{N_T} \frac{y_n}{S_n} = V_T^\pi(\omega) \tag{21}$$

From (19),(20) it can be seen that the LP constraints are satisfied. Therefore $V_T^{(LP)}(\omega) \leq V_T^\pi(\omega)$, for any sample path. We thus obtain (18) which completes the proof. □.

Proof of Theorem 2

Lemma 4 shows that it is only necessary to show our result for the policy $V_T^{(LP)}$ as (18) shows that the result then holds for arbitrary scheduling rules.

To show the result for the $V_T^{(LP)}$ policy, we now construct a finite set $\mathcal{A}$ of scheduling policies. These will be constructed in such a way that the LP policy can be taken "close" to some member of the $\mathcal{A}, |\mathcal{A}| < \infty$, in a way to be described next.

For each pico ℓ, a rate ratio ρ_ℓ must satisfy,

$$0 < \underline{\rho}_\ell \leq \rho_\ell \leq \overline{\rho}_\ell < \infty$$

as determined by the max/min macro and pico rates. It has been shown in [3] that the optimum time to clear the files, in this case $V_T^{(LP)}$, is determined by some choice of $\rho_\ell, \ell = 1, \cdots, L$, which we will denote by $\upsilon_\ell, \ell = 1, \cdots, L$ and for such a choice, $x_\ell(\xi) = D$ if $\rho_\ell(\xi) > \upsilon_\ell$ and is 0 if the opposite inequality holds. If there is equality the user may be served in both the macro and the pico cell. Under a continuous rate model with probability one no pair of users agree on $\rho_\ell(\xi)$ and hence there is at most one split user.

Now we define the set $\mathcal{A}$ to consist of N_P^L assignment rules by choosing N_P equally spaced points on the interval $[\underline{\rho}_\ell, \overline{\rho}_\ell]$, including the end points. Let $H_n(\ell) \doteq [\rho_n(\ell), \rho_{n+1}(\ell)]$ be the interval determining two adjacent points. An assignment, $a \in \mathcal{A}$, is a vector obtained by choosing $(n_1, n_2, \ldots n_L)$, and then setting $a := (\rho_{n_1}(1), \cdots, \rho_{n_L}(L))$. A user in picocell ℓ is assigned to the macrocell if its rate ratio is less than a_l, and the pico BS otherwise. We need not be concerned

with user splitting between macro and pico because the probability of falling on a boundary is 0.

We now consider the optimum assignment which is determined via the rate ratios $\boldsymbol{\upsilon} \doteq (\upsilon_\ell)_{\ell=1}^L$. This is the random assignment that solves the linear program, achieveing $V_T^{(LP)}$. From this assignment, we construct an approximating assignment lying in the set $\mathcal{A}$: Denote by $\mathbf{a}_{\boldsymbol{\upsilon},T} \in \mathcal{A}$ the assignment which contains the optimal rate ratio in the corresponding interval. Also let $\mathbf{n}_{\boldsymbol{\upsilon},T} := (n_\ell, \ell = 1, \cdots, L)$ determine the intervals in which the optimum ratio appears. Note that this is a random assignment, taking different values on different realizations of the start-at-zero system.

The following large deviations result is proven in the Appendix:

Lemma 5. *Let $N_H^{n,l}$ be the number of arrivals whos rate ratios fall into the nth interval in pico-cell ℓ. Then for any $\epsilon > 0$ there is a fixed set of intervals, with N_P sufficiently large, and a corresponding random variable $T_E < \infty$, such that for $T \geq T_E$, the following holds, for all n and l:*

$$\frac{N_H^{n,l}}{T} < \epsilon \quad a.s.$$

Consider the random assignment rule $\mathbf{a}_{\boldsymbol{\upsilon},T} \in \mathcal{A}$. The user assignment under $\mathbf{a}_{\boldsymbol{\upsilon},T}$ is the same as under the optimal allocation, except for users in the interval in which the optimal threshold lies. Lemma 5 shows that for sufficiently large T, the number of users in each of these intervals is upper bounded by $L\epsilon T$. Since the minimum rate is $\underline{R}$, the time to service these users is upper bounded by $\frac{L\epsilon T D}{\underline{R}}$. It follows that for $T > T_E$,

$$\frac{V_T^{(LP)}}{T} \geq \frac{V^{(\mathbf{a}_{\boldsymbol{\upsilon},T})}}{T} - \frac{L\epsilon D}{\underline{R}} \tag{22}$$

$$\geq \inf_{a \in \mathcal{A}} \frac{V^a}{T} - \frac{L\epsilon D}{\underline{R}} \tag{23}$$

Thus,

$$\liminf_T \frac{V_T^{(LP)}}{T} \geq \liminf_T \inf_{a \in \mathcal{A}} \frac{V^a}{T} - \frac{L\epsilon D}{\underline{R}} \tag{24}$$

$$= \inf_{a \in \mathcal{A}} \liminf_T \frac{V^a}{T} - \frac{L\epsilon D}{\underline{R}} \tag{25}$$

$$\geq (1+\eta) - \frac{L\epsilon D}{\underline{R}} \tag{26}$$

the last inequality from Lemma 3. Thus, if we take $\epsilon = \frac{\eta \underline{R}}{2LD}$, we obtain

$$\liminf_T \frac{V_T^{(LP)}}{T} \geq 1 + \frac{\eta}{2}. \tag{27}$$

The proof of the Theorem concludes by noting that

$$\liminf_T \frac{V_T^{\pi}}{T} \geq 1 + \frac{\eta}{2}. \tag{28}$$

holds for *any* clearing schedule π, due to Lemma 4, and hence the workload must build linearly over time for any clearing schedule. □

V. Conclusion

In this paper it has been demonstrated that HetNet capacity can be defined in terms of the solution to a continuous LP. This LP is defined in terms of the point process mean measure of traffic arrivals, the physical rates at user locations and the probability of the files to be downloaded. This set up can be generalised considerably to include a number of aspects not treated here. Amongst these are between pico cell interference and more general traffic models for example with location dependent demands. Of course the underlying optimization problem will be more complex but the principle that it is capacity determining will remain. Note that users moving within the service area present some difficulty for this approach as either a) mobiles remain within the network indefinitely which is unrealistic or b) that users depart the coverage area before their files are delivered so that the treatment of the problem has to change.

Appendix

Proof of Lemma 5

Define

$$m_\ell(n) \doteq \mathbb{E}\left(N_H^{n,\ell}\right)$$

to be the expected number of pico ℓ users arriving per unit time with rate ratios falling into $H_n(\ell)$. Given $\varepsilon > 0$ we may take N_P sufficiently large so that $m_\ell(n) < \varepsilon$ for all ℓ, n. For simplicity we will suppose that $m_\ell(n) > 0$ for all ℓ, n.

Let $J_{n,l}^{(\delta)}$ be the event that

$$\frac{1}{T} N_H^{(\ell,n)} \notin [(1-\delta)m_\ell(n), (1+\delta)m_\ell(n)]$$

By standard large deviation arguments applied to Poisson variates we obtain

$$\mathbb{P}\Big(\frac{1}{T}N_H^{(\ell,n)} \notin [(1-\delta)m_\ell(n),(1+\delta)m_\ell(n)]\Big) \leq e^{-TI_{n,\ell}^{(\delta)}}.$$

Let $J_{\delta,T} \doteq \cup_{n,\ell} J_{n,\ell}^{(\delta)}$. By the union bound we obtain

$$\mathbb{P}\Big(J_{\delta,T}\Big) \leq \sum_{\ell=1}^{L}\sum_{n=1}^{N_P} e^{-TI_{n,\ell}^{(\delta)}} \leq N_P L e^{-I^{(\delta)}T}$$

for some $I^{(\delta)} > 0$. It follows from the first Borel-Cantelli lemma that for any given set of intervals the event $J_{\delta,T}$ will occur finitely many times with probability 1, there being a last slot T_E. Thus, given $\epsilon > 0$ in Lemma 5, choose ε to be any number smaller than $\frac{\epsilon}{(1+\delta)}$ for any $\delta \in \big(0,\frac{1}{2}\big)$, and use the corresponding T_E. □

ACKNOWLEDGMENT

S.V Hanly was supported by the CSIRO Macquarie University Chair in Wireless Communications. This chair has been established with funding provided by the Science and Industry Endowment Fund. P.A. Whiting was supported by the Macquarie University Vice-Chancellor's Innovation Fellowship.

REFERENCES

[1] Patrick Billingsley, *Convergence of Probability Measures* John Wiley, New York, 1968.

[2] H.S. Dillon, R.K.Ganti, F. Baccelli, J.G. Andrews *Modeling and Analysis of K-Tier Downlink Heterogeneous Cellular Networks*, IEEE Jour. on Selected Areas of Communications, Vol. 30, No. 3, 2012.

[3] Sem Borst, Stephen Hanly and Phil Whiting, *Optimal Resource Allocation in HetNets* in Proceedings ICC 2013, Budapest, Hungary.

[4] Sem Borst, Stephen Hanly and Phil Whiting, *Throughput Utility Optimization in HetNets* VTC Spring 2013, Dresden, Germany

[5] S. Corroy, L. Falconetti, R. Mathar, *Cell Association in Small Heterogeneous Networks: Downlink Sum Rate and Min Rate Maximization* IEEE Wireless Communications and Networking Conference: PHY and Fundamentals, Paris, April 1-4, 2012.

[6] William Feller, *An Introduction to Probability Theory and Its Applications* John Wiley, New York 1970.

[7] Ching-Hao Huang and Chen-Yin Liao, *An interference management scheme for heterogeneous network with cell range extension* Network Operations and Management Symposium (APNOMS), 2011 13th Asia-Pacific.

[8] S. Kaneko, T. Matsunaka, Y. Kishi *A Cell-Planning Model for HetNet with CRE and TDM-ICIC in LTE-Advanced* in Proceedings VTC, 6-9 May,2012.

[9] Frank Spitzer *A Combinatorial Lemma and Its Application to Probability Theory* Trans. AMS Vol. 82 No. 2 July 1956.

[10] R1-094225 *DL Performance with HotZone Cells* 3GPP TSG-RAN WG1 Meeting # 58bis, Qualcomm Europe, 2009.

[11] S. Varahadan *Weak Convergence of Measures on Separable Metric Spaces* Sankhyā: The Indian Journal of Statistics (1933-1960), Vol. 19 No. 1/2, pp 15–22, 1958.

[12] H. Wang, X. Zhou, M.C. Reed, *Coverage and Throughput Analysis with a Non-Uniform Femtocell Deployment* arXiv:1305.3694v1

[13] David Williams, *Probability with Martingales* Cambridge University Press, 1989.