Cooperation and decision-making in a wireless multi-provider setting

A. Zemlianov and G. de Veciana
Department of Electrical and Computer Engineering
The University of Texas at Austin
{zemliano,gustavo}@ece.utexas.edu

***Abstract*— In this paper we investigate network design for a wireless service provider using two orthogonal technologies: a WAN technology with uniform spatial coverage and set of LAN access points each with limited coverage. We assume that the system is designed so that users (or their agents) independently and greedily select among the two options based on maximizing a specified utility function which may be a function of the quality of the wireless link, distance to the access points, and/or congestion on system resources. We focus on two complementary aspects of this problem. On the one hand we study system performance under such decision-making strategies. We show convergence of decision-making process to an equilibrium, and that a *congestion- sensitive* utility can provide substantial (300%) performance improvements over natural *proximity-based* criterion. On the other hand, we consider various problems associated with dimensioning typically expensive backhaul links, for the WAN and set of LAN hotspots. Our results show how to best jointly exploit technologies with different coverage scales so as to statistically multiplex spatial load fluctuations in order to reduce backhaul costs.**

I. INTRODUCTION

It is increasingly the case that users can access wireline networks through diverse service providers and technologies. In this complex networking landscape, moving decision-making from access points to devices is a path to achieving system *scalability* [1]. Thus, wireless end-nodes increasingly have the capability to choose among several communication interfaces they might use to access providers and/or transfer data among themselves. For example, a cell phone may be able to choose among two interfaces so as to realize a call through a wide area cellular network or an 802.11 LAN access point, see e.g., [2].

Users' connection strategy could be based on proximity to an access point, amount of interference, quality of service or, more abstractly, based on a utility function capturing a user's valuation of available services and their current costs. Most likely such decision-making would be carried out by software "agents" and driven by users' preferences or engineering design goals. In turn, the strategy that agents implement to choose among available providers will have a substantial impact on the capacity and performance of wireless systems.

In this paper we investigate the interplay between decision-making mechanisms and network design for such a multi-provider scenario. Specifically our focus is on a setting where users may choose among two wireless data access providers (see Figure 1): a wireless wide area network (WAN) service provider engineered to achieve uniform spatial coverage; and a hotspot provider, i.e., an aggregator of LAN access points (hotspots) each with limited coverage, and realizing only limited overall coverage. To capture the spatial interplay among these and spatially distributed users, in Section II we introduce a stochastic geometric model akin to those introduced in [3]. We will model decision-making mechanism of agents using utilities, and assume agents make greedy decisions, i.e., they choose the provider offering the highest utility. In Section III we show the convergence of the process of agents' choices to an equilibrium under assumption that utilities of agents connected to the WAN APs depend not only on congestion level, but also potentially on agent's position relatively to the WAN AP. Moreover, in contrast to our previous work [4], we shift the focus from analysis of competitiveness to the analysis of the benefits that WLAN and WAN providers might get from cooperation. Specifically, in Section IV, we show that on the one hand, *congestion- sensitive* decision-making strategies can provide substantial (300-600%) performance improvements over natural *proximity-based* strategies. On the other hand, we study the complementary role that such heterogeneous wireless data provider scenarios may play by allowing spatial multiplexing (smoothing) of load fluctuations across resources with different coverage scales. In particular, in Section V we show that under congestion dependent decision making strategies, one might potentially significantly reduce the overall backhaul costs – backhaul links from LAN access points to the wired network represent a significant

This is an extended version of the paper that is to appear under same name in *INFOCOM-2005*. It includes proofs that were skipped in the original paper due to space considerations.

fraction of the cost of operating such infrastructure [5].

II. Spatial Model and Notation

To capture the geometry of the network we use the stochastic-geometric framework introduced in [3]. The basic idea is to represent the locations of subscribers and access points (APs) as realizations of spatial point processes (e.g. Poisson) and the service zones associated with the access points as functionals of the realizations of these processes. The main advantage of such models is that they allow one to analytically capture the effect of spatial and load variations in the system based on a reduced set of salient parameters.

We will use three *simple*[1] point processes Π^a, Π^h and Π^w, to represent the locations of subscribers, hotspots and WAN APs respectively. We will refer to the "service zone" of a WAN or a hotspot AP as the set of locations on the plane, that the AP can serve. Agents which fall within the service zones of several APs are able to choose which AP to connect to. In the next few paragraphs we describe our models for the service zones associated with each AP and discuss the criteria the agents use to choose which AP to connect to.

Geometry of overlagged multiprovider scenario. Figure 1 exhibits a realization for a stochastic geometric model for two competing wireless access providers: a WAN service provider and a provider (aggregator) of LAN access points/hotspots. WAN base stations are shown as boxes, with associated coverage areas modeled by cells of a Voronoi[2] tessellation, i.e., each access point is responsible for locations which are closest to it[3]. Thus, the WAN provider's service is available at all spatial locations. By contrast, the second provider's LAN access points, shown as triangles, have limited coverage areas which are modeled by discs centered at each access point. This captures a technology with a highly constrained transmit power, e.g., 802.11 access points sharing unlicensed spectrum.

We formally define the service zones as follows. With each hotspot $h_k \in \pi^h$ we associate a disc $B(h_k,d)$ of radius $d>0$ and centered at h_k. We assume that service from h_k is available only within the disc (see Figure 1). In addition, we assume that agents desiring to connect to a hotspot will connect only to the *closest* feasible hotspot. This yields a service zone S_k^h for hotspot AP h_k given by:

$$S_k^h \triangleq V_k^h \cap B(h_k,d).$$

For each WAN AP $w_m \in \pi^w$ we define its service zone, S_m^w, to be its Voronoi cell, V_m^w, augmented by the service zones of the hotspots that have their APs within V_m^w:

$$S_m^w = V_m^w \bigcup \left(\bigcup_{k\in\mathcal{K}_m} S_k^h \right) \setminus \left(\bigcup_{l\in\cup_{n\neq m}\mathcal{K}_n} S_l^h \right),$$

where $\mathcal{K}_m$ denotes the set of indices of hotspots located within the Voronoi cell V_m^w (for notation summary, see Table (I)).

Note that this definition constrains each agent $a_i \in \pi^a$ to select between connecting to the closest hotspot AP h_k (if it is covered by its service zone) and the WAN AP w_m *which contains* h_k in its service zone[4]. In the sequel

TABLE I
NOTATION SUMMARY

Π^a	Point process modeling agents' locations
Π^h	Point process modeling hotspots' locations
Π^w	Point process modeling WAN AP locations'
π^a, π^h, π^w	Realization of Π^a, Π^h, Π^w
$\pi(A)$	All points of a realization π that fall within the set A
$\lvert\pi(A)\rvert$	Number of points in $\pi(A)$
$\lvert x\rvert$	Length of vector $x\in\mathbb{R}^2$
$B(x,r)$	Disc of radius r centered at $x\in\mathbb{R}^2$
V_m^w	Voronoi cell of WAN AP $w_m \in \pi^w$
V_k^h	Voronoi cell of hotspot AP $h_k \in \pi^h$
$\mathcal{K}_m$	$\{k:\ h_k \in \pi^h(V_m^w)\}$, indices of hotspots located within the Voronoi cell V_m^w
S_k^h	Service zone of hotspot h_k
S_m^w	Service zone of WAN AP w_m
C_m	Subset of S_m^w where agents can make choices
$\bar{C}_m$	$S_m^w \setminus C_m$
M_m^w	Total number of agents in S_m^w
M_k^h	Total number of agents in S_k^h
M_{C_m}	Total number of agents in C_m
$M_{\bar{C}_m}$	Total number of agents in $\bar{C}_m$
$N_m^w(t)$	Total number of agents connected to WAN AP w_m at time t
$N^h(a_i,t)$ or $N_k^h(t)$	Number of agents connected to hotspot h_k at time t, where k is s.t. $a_i \in S_k^h$
$U_i^w\left(N_m^w(t)\right)$	Utility function of agent $a_i \in S_m^w$, connected to WAN AP w_m at time t
$U_j^h\left(N_k^h(t)\right)$	Utility function of agent $a_j \in S_k^h$ connected to hotspot h_k at time t

[1]The location of each WAN or hotspot AP is not shared by any other AP [6], i.e. points do not overlap.

[2]Voronoi cell of $w_m \in \pi^w$ is the set of all points on the plane that are closer to w_m than to any $w_n \in \pi^w$, $n \neq m$.

[3]In practice, there would be overlap among coverage areas associated with base stations, yet this is a reasonable approximation in the case where relatively high power levels are used, see e.g., [7].

[4]As will be seen later this requirement makes each agent's choice contingent on information available locally at WAN AP w_m.

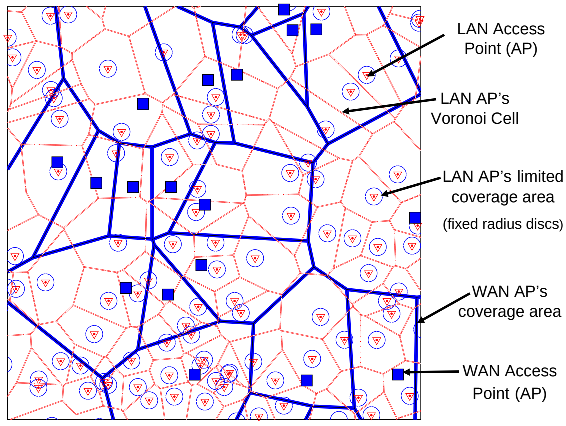

Fig. 1. Geometry of a multi-tier wireless network.

we will make the following assumption:

Assumption 1: For all $m \in \mathbb{N}$, the service zones S_m^w contain an almost surely finite number of agents and hotspots.

We let C_m be the subset of S_m^w that includes spatial locations where agents would have the option to choose among a hotspot and WAN AP w_m:

$$C_m \triangleq \bigcup_{k \in \mathcal{K}_m} S_k^h .$$

Users which fall in $\bar{C}_m \triangleq S_m^w \setminus C_m$ can not make a choice and will be assumed to automatically connect to WAN AP w_m. By contrast, an agent $a_i \in C_m$ is also covered by some hotspot h_k's service zone and can choose between connecting to *either* h_k or the WAN AP w_m.

Decision-making models. We will consider two basic mechanisms for decision making. We refer to the first mechanism in which agents in C_m simply connect to the closest hotspot as the *proximity based* (PX) mechanism. Under the second, *utility based* (UT) mechanism, an agent's decision is based on a utility function. We will consider two types of utility functions which we define below.

Definition 1: We say that the utility function U_j of an agent a_j connected to WAN (or hotspot AP), x, is *congestion* and *agent* dependent if U_j is a function of a total number of agents connected to x, and possibly is different for each j.

Definition 2: We say that a utility function U_j is *solely congestion* dependent if, for each $N \in \mathbb{N}$, $U_j(N) = U_i(N)$ whenever the agents a_j and a_i lie within the same service zone of x.

Consider an agent $a_i \in \pi^a(C_m)$ that is connected to WAN AP w_m at time t and assume that the total number of agents that are connected to w_m at that time is $N_m^w(t)$. We model the level of "satisfaction" of agent a_i with the service via a congestion and agent dependent utility function $U_i^w\left(N_m^w(t)\right)$ that depends on the current congestion level and possibly the agent's location[5] within S_m^w.

We assign a solely congestion dependent utility function $U_j^h\left(N_k^h(t)\right)$ to an agent $a_j \in \pi^a(S_k^h)$ connected to a hotspot at time t. Here $N_k^h(t)$ denotes the total number of agents that are connected at time t to the same hotspot as agent a_j. As opposed to the case with service from the WAN, we require that the perception of service from hotspots to be the same for agents connected to the *same* hotspot, i.e., if $a_i, a_j \in S_k^h$, then $U_j^h(N) = U_i^h(N)$, for any $N \in \mathbb{N}$. However, we do not impose this restriction for agents connected to different hotspots, thus we retain the flexibility of including potentially different hotspots' types in the model[6].

In the sequel we will use the following assumption for the utility functions:

Assumption 2: For all $i \in \mathbb{N}$, $U_i^w(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}$ and $U_i^h(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}$ are continuous, monotonically decreasing functions.

Once utility functions have been specified for each agent, we will assume agents make decisions consistent with maximizing their utility, i.e., connect to the provider offering the higher utility. However, we will account for a fixed cost of switching to another interface. We stress here that this is simply a model for decision-making, and need not involve any specific transaction of money among agents.

We assume that for each agent in C_m there is a sequence of times, at which the agent makes decisions. If t is a time when agent $a_i \in C_m$ is making a choice, then, we postulate that a_i switches to the WAN AP w_m from a hotspot h_k if and only if it was connected to h_k at time t^- and

$$U_i^w\left(N_m^w(t^-)+1\right) > U_i^h\left(N_k^h(t^-)\right) + c^w ,$$

where t^- refers to the time immediately prior to t and c^w represents a cost of switching to the WAN AP. Similarly, the agent $a_i \in S_k^h$ switches to a hotspot h_k at time t if and only if it was connected to a WAN AP w_m at t^- and

$$U_i^h\left(N_k^h(t^-)+1\right) \geq U_i^w\left(N_m^w(t^-)\right) + c^h ,$$

where c^h represents the cost of switching to a hotspot. Note that we break ties in favor of hotspots.

[5]Note that this allows to model a situation when the agents, that are farther from the WAN AP have potentially worse communication channels.

[6]For example, hotspots could support different bandwidths.

III. Equilibrium: convergence and structure.

Convergence to equilibrium. In this section we consider the dynamics of agents' decision making. In particular we investigate if the dynamics converge to a particular fixed point which we refer to as an "equilibrium".

Definition 3: Consider a service zone S_m^w of WAN AP w_m for a particular realization of agents, hotspots and WAN APs on the plane. We refer to a particular configuration of agents' choices within S_m^w as equilibrium, if, given this configuration, no agent desires to alter its choice.

Let us denote M_m^w the total number of agents that fall within S_m^w for a particular realization. We will make the following assumption.

Assumption 3: If $a_i, a_j \in \pi^a(S_k^h)$ where $k \in \mathcal{K}_m$ then $|U_i^w(N) - U_j^w(N)| < c^h + c^w$ for $1 \leq N \leq M_m^w$.

Assumption 3 requires that the utility associated with connections to the WAN does not vary too much for agents located within the service zone of the same hotspot. For example if the performance of WAN connections simply degrades with distance, then this assumption requires the coverage radius of a single hotspot to be small enough.

Theorem 1: Consider the service zone S_m^w for a particular fixed realization π^a, π^h and π^w. Assume that agents make decisions at times modelled by a Poisson process with rate μ, and with probability $p(a_i) > 0$ a decision time is associated with agent a_i. Then, under Assumptions 1-3, given any initial configuration of agents' choices, say, at time $t = 0$, the system converges a.s. to an equilibrium configuration as $t \longrightarrow \infty$.

The proof of this result is lengthy and is given in [8]. The argument, however, is straightforward, since one just needs to show that the dynamics of agents decisions represents a transient Markov chain.

"Shape" of equilibrium. Note that, in general, the specific character of the agents' choices equilibria depends on the utility functions, distances from access points, and resource allocation mechanisms at the access points. For a service zone of a particular WAN AP an equilibrium might not be unique. For solely congestion dependent utilities, however, as we show in [4], the set of all equilibria in each WAN service zone could be made quite "tight", by appropriately selecting the utility functions. In this case, the equilibrium condition roughly corresponds to determining the level K_m^* which partitions the users at any hotspot $h_k \in S_m^w$ into those that choose the WAN and those that choose the hotspot. Figure 2 exhibits the structure of an equilibrium for a simple solely congestion dependent utilities. The cylinders correspond to the loads, i.e., number of active users, within the coverage area of the hotspots. The intuition is that only the hotspots that are overloaded *relative* to the WAN load will have users that choose the WAN. In [4] we show for a broad class of solely congestion dependent utility functions, that, in fact, given any realization of agents', hotspots' and WAN APs' locations, the configuration of agents' equilibrium choices *maximizes the worst agent's utility* within a service zone of any WAN AP. This advocates the use of utility based choice mechanisms over any other.

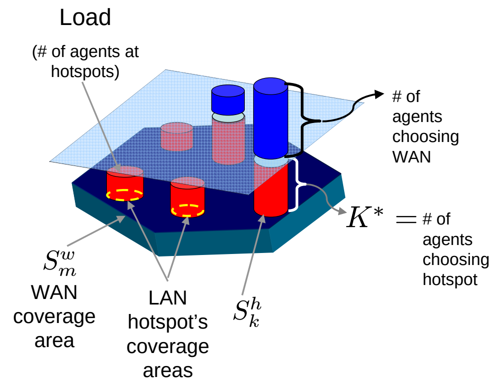

Fig. 2. "Shape" of equilibrium.

IV. System performance in equilibrium

Here we will construct several simulation examples that demonstrate the gains that could be achieved by employing the utility based choice mechanisms.

Simulation settings. The simulation examples that follow in the next paragraph are based on the same geometric and traffic assumptions. We let the locations of agents, hotspots and WAN APs be given by independent Poisson point processes with densities λ^a, λ^h and λ^w with values specified in Table II. Each agent generates a Poisson stream of download requests with rate γ, where each request is for a file of average size f. Furthermore we assume that no request is blocked from service, and thus restrict ourselves to scenarios with light average resource utilization[7].

We let the utility function of each agent be given by the negative of the average (over time) delay of file transfers for the agent[8]. Thus the quality of service that an agent experiences depends in part on the resource

[7] In practice if loads were excessive, users would either be blocked or would choose not to connect because the utility is too low. In this case the users' "unsatisfaction" would be reflected partially in the value of blocking probability, whereas we can estimate unsatifaction by measuring, e.g. the average utility of connected users.

[8] We assume that each agent can reliably estimate her average delay within a typical time between two sequential changes in configuration of agents' choices within a service zone of the WAN AP.

allocation strategy that is employed at the AP of the agent's choice. For our simulations we assume that the agents connected to the same resource are served according to a processor sharing service discipline. This assumption is also made in [9] where multi-class processor sharing model was used to analyze the "Qualcomm" HDR [10] downlink scheduling protocol used in 3G wireless systems.

For simplicity we let the rate for each agent from any AP be constant over time. This assumption greatly simplifies performance analysis of the system, but it neglects the boost in performance that can be achieved via serving the agents with "good" channels in opportunistic manner[9]. We note however, that when the channel fading is Rayleigh, the boost in effective rate associated with serving users opportunistically is roughly proportional to the logarithm of the number of connected users. Thus one could argue, that once the number of users connected to a resource is large enough, the gain in service rate does not vary much with the number of connected users and could be modelled as a constant "effective" factor that multiplies each agent's rate.

With these considerations, we arrive at the following simple expression for the the utility of agent a_i connected to a hotspot h_k at time t is:

$$U_i^h\left(N_k^h(t)\right) = -\frac{f}{B_k^h - \gamma f N_k^h(t)}\,, \qquad (1)$$

where B_k^h is the service rate of any agent connected to a hotspot h_k. Note that we used an expression for the average delay of a single class M/GI/1-PS queueing discipline [12]. Thus the utility of agents connected to hotspots are solely congestion dependent.

We allow agents connected to the WAN to be possibly served with different rates, that might depend on the location of agents relatively to their WAN APs. Let $B_m^w(a_j)$ denote the rate of agent a_j connected to the WAN AP w_m. Then the delay of agent a_j connected to the WAN AP w_m at time t is given by that of a multi-class $M/GI/1-PS$ queue, and thus:

$$U_j^w(t) = -\frac{f}{B_m^w(a_j) - \gamma f \sum_{a_i \in \mathcal{W}_m(t)} B_m^w(a_j)/B_m^w(a_i)}\,, \qquad (2)$$

where $\mathcal{W}_m(t)$ is the set of all agents connected to w_m at time t.

Remark 1: Note that, the utility function (1) does not obey our Assumption 2 on utilities, since it is defined only for a bounded sub-interval of $\mathbb{R}^+$ given by $N_k^h(t) < \frac{B_k^h}{\gamma f}$. In addition, the utility defined by (2) is not of a congestion and agent dependent type, meaning that Theorem 1 is not directly applicable to this case. We note, however, that for light loads, considered in our scenarios, the convergence of decision dynamics to equilibrium when agents' utilities are defined by (1-2) can still be established. Although, we will leave the details of the corresponding proof out of this paper.

[9]See e.g. [11] for the analysis of queueing models of opportunistic scheduling.

TABLE II
SIMULATION PARAMETERS

Parameter	Notation	Value
Agents' density	λ^a	100 km^{-2}
Hotspots' density	λ^h	30 km^{-2}
WAN APs' density	λ^w	0.81 km^{-2}
Hotspot coverage radius	d	100 m
Requests/user	γ	1 min^{-1}
Average file size	f	80 kB
Simulated area	$\lvert A\rvert$	100 km^2
WAN b/w (scenario 1)	$B^w(a_i)$	1 Mbps
WAN b/w (scenario 2)	$B^w(a_i)$	0.04 – 2.46 Mbps

System performance. Here we consider two scenarios, where the first models the WAN service as being uniformly available (i.e. the WAN service is the same for all agents irrespective of their locations), while the second allows variation in the WAN rate across agents according to some distribution. In both scenarios we vary the available bandwidth B^h at each hotspot from 0.1 to 1 Mbps. The choice of the range for B^h, apart from exposition convenience, was stipulated by the fact that the cost of the backhaul is a major bottleneck that affects the performance of the hotspots [5]. Thus although up to 11 Mbps wireless access rates are theoretically available at each hotspot, it is rarely the case that hotspot providers support backhauls with bandwidth exceeding $1-1.5$ Mbps – typically DSL connection speeds are more common.

We will consider two performance metrics. The first is the mean delay averaged across users:

$$\bar{D} \triangleq \frac{1}{|\pi^a(A)|} \sum_{a_i \in \pi^a(A)} D(a_i)\,,$$

where A is the simulated area and $D(a_i)$ is average file transfer delay seen by agent a_i. The second metric is the average worst case user's delay per WAN service zone, and is defined as

$$\bar{W} = \frac{1}{|\pi^w(A)|} \sum_{w_m \in \pi^w(A)} \max_{a_i \in S_m^w} D(a_i)\,,$$

for a simulated region A. Since we do not have blocking, all our results are conditioned on the event that the overall system is stable. However, the parameters of

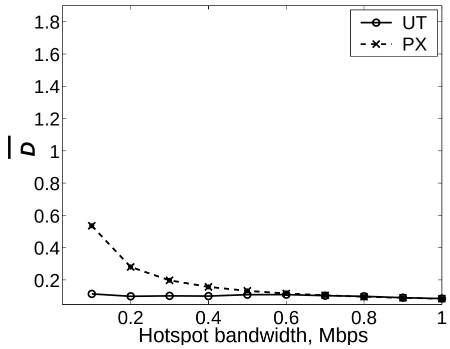

Fig. 3. Scenario 1: Average performance.

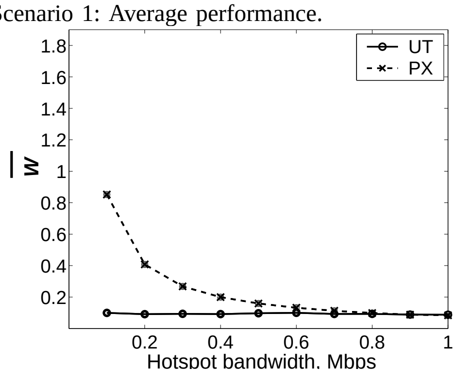

Fig. 4. Scenario 1: Average worst per cell performance.

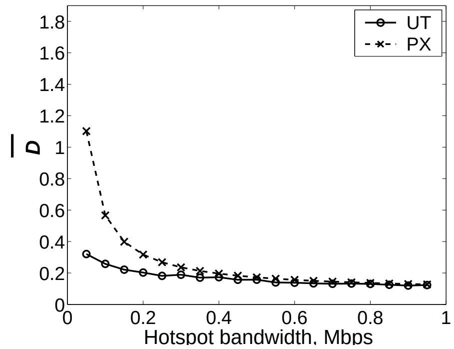

Fig. 5. Scenario 2: Average performance.

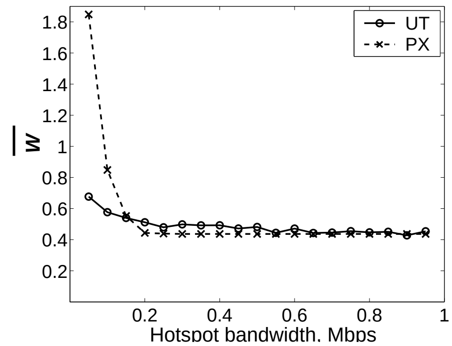

Fig. 6. Scenario 2: Average worst per cell performance.

simulation are chosen in such a way that the probability of instability is very small.

In the first scenario, all agents have the same WAN service rate $B_m^w = 1$ Mbps and thus the utility of agent a_j connected to a WAN AP w_m at time t is:

$$U_j^w\Big(N_m^w(t)\Big) = -\frac{f}{B_m^w - \gamma f N_m^w(t)}\,.$$

Therefore, in this case the utilities of agents are solely congestion dependent. Figures 3, 4 show $\overline{D}$ and $\overline{W}$ after convergence to equilibrium versus B^h for utility based and proximity based selection strategies. There are significant gains both in the average per user and worst case performance per cell if the available bandwidth at hotspots is less than 60% of that available at the WAN. Given the parameters in Table II one might deduce that PX based strategy needs at least five times more bandwidth at the backhaul than the UT strategy to achieve the same average per user performance and even more bandwidth to achieve the same value of $\overline{W}$.

For the second scenario we assume that any agent connected to the WAN has rate that is determined by WAN SNR values given in Table I of [9] and based on the CDMA 1xEV-DO system [11]. For agents in $\overline{C}_m$ we assign SNR values independently according to the cdf given in Figure 1 of [9]. Finally, we assign the WAN SNR values for agents in different hotspots independently according to the same cdf, and let the agents that are in the service zone of the same hotspot have the same SNR.

We obtain substantial performance gains in average performance (Figure 5) once UT choice strategy is employed instead of PX, but performance gains are less significant than in the first scenario. Moreover, the worst case performance per WAN service zone (Figure 6) might, although negligibly, become worse for UT than for PX, once the bandwidth used at each hotspot becomes large enough. This suggests that UT strategy is likely to be more effective once the service from the WAN is closer to being uniformly good within WAN service zones.

V. Analysis and Design of Cooperative Multiprovider Wireless Systems

Backhaul allocation problem. The results of previous section are not unexpected. The performance gains arise from ability of WAN APs to cover considerably larger service areas than hotspots and thus statistically multiplex spatial load fluctuations. Given the WAN service is uniform enough and utilization of the WAN APs is low, the WAN APs can serve as a pooled resource which absorbs load fluctuations within hotspots. Thus the cost of the hotspots' backhaul can be potentially reduced, if hotspots cooperate with WAN APs.

In this section we will further quantify the savings that can be achieved from such cooperation. We will assume that a service provider is operating a multi-tier network where WAN APs coexist with hotspots. The provider wishes to design a network so as to minimize the backhaul expense, but desires to keep the performance of any user within the network at an appropriate level.

We fix a particular realization for the WAN APs and hotspots and consider a service zone S_m^w of a single WAN AP w_m. Assume for simplicity that S_m^w includes H_m hotspots that have *non-overlapping* service zones of area πd^2. We assume that the number, M_m^w, of agents within S_m^w is random, whereas agents distributed on the plane according to Poisson point process with density λ^a. Let each of the agents generate a stream of download requests with arrival rate γ and average file size f. Finally, assume that the service rates associated with the WAN and hotspots are the same within their respective service zones. Thus if an agent $a_i \in S_m^w$ is connected to the WAN AP, then the expected (over time) delay $D(a_i)$ experienced by this agent is given as in the M/GI/1-PS discipline:

$$D(a_i) = \frac{f}{B_m^w - \gamma f N_m^w}\,, \tag{3}$$

where B_m^w is the bandwidth available at WAN AP w_m. If a_j is connected to hotspot h_k, then the average over time delay, $D(a_j)$ of this agent is given by:

$$D(a_j) = \frac{f}{B_k^h - \gamma f N_k^h}\,, \tag{4}$$

where B_k^h is the bandwidth available hotspot h_k.

The provider seeks a solution to the following optimization problem:

Problem 1: (*Minimizing Backhaul Costs*):

$$\min_{B_m^w,\ \{B_k^h\},\ k\in\mathcal{K}_m} \left[B_m^w + \sum_{k\in\mathcal{K}_m} B_k^h \right], \tag{5}$$

under constraints:

$$\mathbb{P}\left(\max_{a_i \in S_m^w} D(a_i) > \theta \right) \le \delta \tag{6}$$

$$0 \le B_m^w \le \hat{B}^w\,,\ B_k^h \ge 0\,. \tag{7}$$

Here (6) ensures that the delay of the agent with worst performance in S_m^w is smaller than target θ with some pre-specified probability. The constraints (7) assume that the wireless access bandwidth at hotspots is unlimited and thus only constrained by the backhaul. By contrast, only $\hat{B}^w$ is available at each WAN AP[10].

[10] Usually the wireless access bandwidth at hotspots exceeds by far the available backhaul resources, while for 3G service the wireless access bandwidth is likely to be bottlenecked by the available spectrum.

Let $\hat{M}_m^w$ be the largest integer such that:

$$\mathbb{P}(M_m^w > \hat{M}_m^w) \le \delta\,,$$

and

$$\hat{N}_m^w = \left\lfloor \frac{\hat{B}^w}{\gamma f} - \frac{1}{\theta\gamma} \right\rfloor. \tag{8}$$

Here $\hat{N}_m^w$ is the the largest number of agents that the WAN AP could serve with delay not exceeding θ, if $\hat{B}^w$ was allocated for this AP. We have the following proposition, which we prove in Appendix:

Proposition 1: There are three regimes to consider in solving Problem 1:
(i) If

$$\hat{N}_m^w \ge \hat{M}_m^w\,, \tag{9}$$

then a policy that allocates $B_m^w = \gamma f \hat{M}_m^w + \frac{f}{\theta}$ to the WAN AP w_m and no bandwidth to any of the hotspots is optimal for Problem 1 when H_m is large enough.
(ii) If

$$\mathbb{P}(M_{C_m}^w > \hat{N}_m^w) > \delta\,, \tag{10}$$

then no solution to Problem 1 exists.
(iii) If both (9) and (10) are violated, then a policy that is optimal for large enough H_m allocates $B_m^w = f\gamma\hat{N}_m^w + f\theta$ units of bandwidth to the WAN AP and $B^h(\theta, K) = f\gamma K + f\theta$ units of bandwidth to each of the hotspots in S_m^w, where K is the smallest integer such that:

$$\mathbb{P}\left(M_{C_m}^w + \sum_{k\in\mathcal{K}_m} (M_k^h - K)\mathbf{1}_{\{M_k^h > K\}} > \hat{N}_m^w \right) \le \delta\,. \tag{11}$$

Regime (iii) identified by Proposition 1 can be viewed as the regime when both WAN and hotspots benefit from cooperation. Indeed, the WAN AP is unable to handle all the traffic due to the limit on the wireless access bandwidth. In the same time the backhaul allocated to each hotspot enables a hotspot to serve at most K agents, where K is given by (11). It is optimal to shift the "overload" in each hotspot to the WAN AP. The next proposition shows that to realize such load shifting distributively, one just has to implement utility based connection strategy for the agents.

Proposition 2: Assume that Problem 1 has a solution and that the bandwidth in S_m^w has been split between hotspots and WAN AP according to Proposition 1. Let the agents' selection criterion be based on utility, whereas the utility function for each agent is given by the negative of the agent delay, given by either (3) or (4). Then, the probabilistic requirement (6) is met when agents connect according to their equilibrium choices.

Proof: (Outline.) One could use Proposition 3.3 in [4] to verify that the choice of utilities guarantees that the performance of an agent with worst utility when agents are connected according to their equilibrium choices is at least as good as for any other connection strategy. ■

Optimal bandwidth allocation vs. minimum allocation for PX strategy. In what follows we compare the optimal total bandwidth in the sense of Proposition 1 with the total bandwidth that would be required to meet the probabilistic constraint (6) if the system is designed for PX based agents' choice mechanism. To simplify exposition, we will make this comparison under the assumption of having no upper constraint on the bandwidth that is used by the WAN AP, i.e. let $\hat{B}^w = \infty$.

As seen earlier, the optimal strategy for Problem 1 allocates sufficient resources on the WAN and allows all agents to connect to the WAN APs. Recall that under PX strategy the agents falling within the service zones of the hotspots must connect to the hotspots. Under both optimal and PX resource allocation the agents that do not fall within the service zones of any hotspot must be served by the WAN, thus there is a comparable cost for both PX and optimal strategy that is associated with provisioning at the WAN for this type of agents. At the same time, we expect to see overprovisioning cost associated with agents at hotspots to be quite large for PX strategy in comparison to the optimal.

We find the lower bound on the savings in overprovisionning by considering a suboptimal strategy that allocates two separate channels for agents that are within and outside C_m. Under this strategy, bandwidth cost associated with meeting a delay requirement for the users in $\bar{C}_m$ is exactly the same as for PX strategy. We are left to compare only the savings in bandwidth associated with serving the agents in C_m by either hotspots or the WAN AP.

We will find the minimal bandwidths B_{PX} and B_o that is required to meet the delay requirement (6) of agents within C_m, once the PX or optimal connection strategy respectively is deployed. We define the access bandwidths:

$$\Delta B_{PX} = B_{PX} - \bar{B}, \text{ and } \Delta B_o = B_o - \bar{B}.$$

Here $\bar{B}$ is the minimal bandwidth that has to be used to serve the agents in C_m when there are exactly average number of them, $H_m \lambda^a \pi d^2$, residing in C_m, thus

$$\bar{B} = \gamma f H_m \lambda^a \pi d^2 + \frac{f}{\theta}.$$

Proposition 3: For large H_m,

$$\frac{\Delta B_o}{\Delta B_{PX}} = O\left(\frac{1}{\sqrt{H_m}}\right).$$

Proof: We first find the minimum bandwidth that has to be allocated to hotspots to meet the delay requirement (6). Denote $\bar{M} = \mathbb{E}\left[M_k^h\right] = \lambda^a \pi d^2$ and note that $\mathbf{var}[M_k^h] = \bar{M}$. Let $\kappa^h(\delta, H_m)$ be the smallest positive number such that:

$$\left[\mathbb{P}(M_k^h \leq \bar{M} + \kappa^h(\delta, H_m)\sqrt{\bar{M}})\right]^{H_m} \geq 1 - \delta. \qquad (12)$$

Clearly, $\kappa^h(\delta, H_m)$ is a nondecreasing function of H_m for any fixed δ. From (3), we obtain:

$$\mathbb{P}\left(\max_{a_i \in S_m^w} D(a_i) > \theta\right) \leq \delta,$$

if and only if

$$B_{PX} = B_{PX}(\delta, H_m, \theta) = \gamma f(\bar{M} + \kappa^h(\delta, H_m)\sqrt{\bar{M}}) + \frac{f}{\theta}.$$

Thus $H_m B_{PX}(\delta, H_m, \theta)$ is the minimum total bandwidth that has to be allocated for hotspots when PX strategy is deployed, which gives the excess bandwidth:

$$\Delta B_{PX} = (H_m - 1)\frac{f}{\theta} + H_m \kappa^h(\delta, H_m)\sqrt{\bar{M}}.$$

Now we find the total bandwidth that would be needed by the WAN AP to serve the agents within the hotspots and meet the delay requirement. Following the same logic as above, we find:

$$B_0 = B_0(\delta, H_m, \theta) = \gamma f(H_m \bar{M} + \kappa^w(\delta)\sqrt{H_m \bar{M}}) + \frac{f}{\theta},$$

where we defined $\kappa^w(\delta)$ as:

$$\mathbb{P}(\sum_{k \in \mathcal{K}_m} N_k^h \leq H_m \bar{M} + \kappa^w(\delta)\sqrt{H_m \bar{M}}) = 1 - \delta.$$

Note that, assuming the Central Limit Theorem holds, we have that $\sum_{k \in \mathcal{K}_m} M_k^h$ is distributed normally with variance $H_m \bar{M}$. Since the mean and the variance uniquely define any normal distribution, we have that $\kappa^w(\delta)$ does not depend on H_m. Therefore, for sufficiently large[11] H_m and any fixed $\delta > 0$ we have $\kappa^w(\delta) \leq \kappa^h(\delta, H_m)$. The excess bandwidth when all agents in C_m are served by the WAN is given by:

$$\Delta B_o = B_o - \bar{B} = \kappa^w(\delta)\sqrt{H_m \bar{M}}.$$

Comparing ΔB_m^w and ΔB^h we find:

$$\frac{\Delta B_o}{\Delta B_{PX}} = O\left(\frac{1}{\sqrt{H_m}}\right), \qquad (13)$$

where we used that $\kappa^h(\delta, H_m) \leq \kappa^w(\delta)$ for sufficiently large H_m. ■

Remark 2: Note that when the delay requirement is very stringent, in particular when $\theta \ll (\gamma \lambda \pi d^2)^{-1}$, then

[11] In fact we have checked numerically that for $1 \leq \lambda \leq 100$ and $0 < \delta < 1$ we have $\kappa^w(\delta) \leq \kappa^h(\delta, H_m)$ already when $H_m \geq 3$.

the excess bandwidth $\Delta B_{PX} \gg \bar{B}$. Since B_o is of the same order as $\bar{B}$ we may find that the scaling of Proposition 3 holds also when the excess bandwidths ΔB_o and ΔB_{PX} are replaced by total bandwidths B_o and B_{PX} respectively.

Remark 3: Note that we estimated the cost of bandwidth overprovisionning at hotspots for very mild spatial load fluctuations. We expect even more profound savings in backhaul costs once the traffic is more bursty, e.g. when there are hourly traffic fluctuations associated with users' migration.

Optimal number of hotspots for a given backhaul size. Proposition 1 could also be used to compute the optimal number of hotspots that should be placed within S_m^w. In particular a service provider might wonder if putting more hotspots in the area but having them able to support less users is better then doing otherwise. Indeed, we have a tradeoff between the risk associated with having users uncovered by any hotspot and the risk of having too often users that hotspots can not support.

By Proposition 1, part (iii) we know that B_m^w should be made as large as the available spectrum can support, and thus we have to decide only on how many hotspots would give best performance within S_m^w when there is a constraint on the total used by hotspots backhaul bandwidth, B_{tot}^h. More formally, we need to find H_m, such that

$$\mathbb{P}\left(M_{C_m}^{\underline{w}} + \sum_{k \in \mathcal{K}_m} (M_k^h - K)\mathbf{1}_{\{M_k^h > K\}} > \hat{N}_m^w\right) \qquad (14)$$

is minimized under constraint:

$$H_m(f\gamma K + f\theta) \leq B_{tot}^h \,. \qquad (15)$$

Note that behavior of the probability given by (14) as a function of H_m might be quite complex. Figure (7) exhibits this behavior in a typical scenario. Observe that P at the graph (7) goes down steadily with H_m, until $H_m = 10$. At this point, the number of agents that each hotspot can support without violating the delay requirement is just above average of the number of agents falling in each hotspot (see Figure (8)). The graph is much less regular once each hotspot is able to support less than this number of agents. Still, the lowest point at the graph is for $H_m = 17$, whence the number of agents that each hotspot can support without violating the delay requirement is below the average number of agents falling within a single hotspot. Hence, in this scenario, when the optimal number of hotspots is used, each hotspot would need to cooperate with the WAN by "shifting" its frequently occurring overloads.

Backhaul allocation when WAN service is not uniformly available. Let us now assume that the WAN AP does not provide uniform quality of service to all points

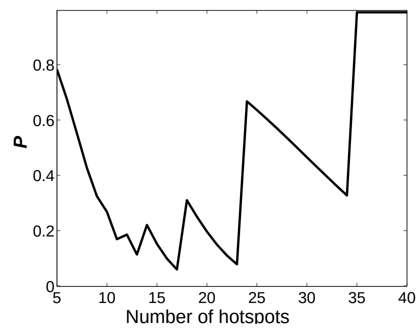

Fig. 7. Probability of exceeding target delay vs. H_m.

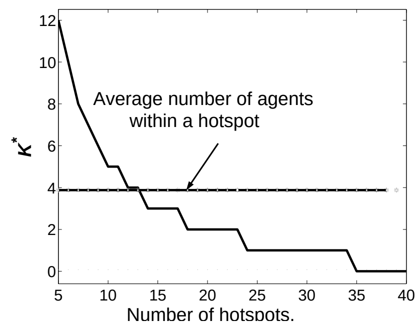

Fig. 8. The largest number of agents a hotspot could serve without violating target delay vs. H_m.

within its service zone. This situation would occur if, for example, the signal from the WAN AP at a particular location is shadowed by an obstruction. In this case, the performance of the system will depend not only on the number of employed hotspots and size of their backhaul, but also on their positions.

To model such scenario we will let the service zone S_m^w to be comprised of H_m *non-overlapping, equal-sized* subzones (sites) $\{S_k^h\}_{k=1}^{H_m}$, where S_k^h could be completely covered by a single hotspot. Suppose that the rate $B_m^w(a_i) = B_m^w(h_k)$ of an agent located within the site S_k^h, for $k \in \mathcal{K}_m$ belongs to a discrete set $\mathcal{B} \triangleq \{b_r^w\}_{k=1}^{N_B}$ that consists of N_B different rates. We assume that of $H_m \gg N_B$, thus there are likely to be many sites in S_m^w with the same WAN rate.

We assume that no hotspots have been installed, but S_k^h, for $k \in \mathcal{K}_m$ represents a site of possible hotspot installation. Our goal is to consider optimal choices for hotspots' sites and dimensioning of backhaul bandwidth. Since the analog of Problem 1 is much more complex in this setting, we will make several simplifying assumptions and reformulate the problem as to retain the key aspects of it.

(i) We assume that the spectrum at the WAN is fully utilized. Since the WAN service could be arbitrarily poor in some locations it might happen that a single hotspot installed at such locations would exploit the backhaul bandwidth much more effectively. Thus the solution to an analog of Problem 1 posed in this scenario might lead in some cases to the conclusion that the available at the WAN spectrum for communication should not be fully utilized. Since the cost of purchased spectrum probably exceeds by far the backhaul associated costs within a cell, such solution would indicate that the placement of the WAN AP or overall WAN design is poor.

(ii) We assume that the operation regime is such that each hotspot takes the largest number of agents within its service zone that it can serve with average delay of at most θ and shifts the remaining agents to the WAN AP.

(iii) Note that the assumption (i) allows us to consider optimization of backhaul allocation associated only with hotspots, and assumption (ii) permits us to account only for the performance of agents connected to the WAN AP. Our last modification to Problem 1 is that in place of $\max_{a_i \in S_m^w} D(a_i)$ we will concentrate on the delay averaged across agents that are connected to the WAN AP w_m:

$$\bar{D}_m^w = \frac{1}{N_m^w} \sum_{a_j \in \mathcal{W}_m} D(a_j),$$

where $\mathcal{W}_m$ dentes the set of agents that are connected to the WAN via rule described in (ii). Using the expression for average delay in the multi-class M/GI/1-PS queue we have:

$$\bar{D}_m^w = \frac{1}{N_m^w} \left(\frac{1}{\sum_{a_j \in \mathcal{W}_m} \frac{f}{B^w(a_j)}} - \gamma \right)^{-1}. \tag{16}$$

With this set of assumptions we arrive at:

Problem 2:

$$\min_{\{B_k^h\},\ k \in \mathcal{K}_m} \left[\sum_{k \in \mathcal{K}_m} B_k^h \right], \tag{17}$$

under constraints:

$$\mathbb{P}(\bar{D}_m^w > \theta) \leq \delta, \tag{18}$$

$$B_k^h \geq 0,\ \forall k \in \mathcal{K}_m. \tag{19}$$

Let us assume that each hotspot h_k, $k \in \mathcal{K}_m$ is provided enough bandwidth to serve up to K_k agents with delay not exceeding θ, i.e.:

$$B_k^h = \gamma f K_k + \frac{f}{\theta}.$$

Then, solving Problem 2 reduces to finding the optimal set of values $\{K_k\}_{k \in \mathcal{K}_m}$, where $K_k \geq 0$ for all $k \in \mathcal{K}_m$. In [8] we show how to find approximate values for K_k. In summary, we approximate the distribution of the number of agents within S_k^h via a Gaussian random variable η such that $\mathbb{E}\eta = \mathbf{var}\eta = E[M_k^h]$. Then

$$g(K_k) \triangleq \mathbb{E}\left[(\eta - K_k)\mathbf{1}_{\{\eta > K_k\}}\right],$$

approximately gives the average number of agents within a hotspot that connect to the WAN and

$$L\left(\{K_k\}_{k \in \mathcal{K}_m}\right) \triangleq \sum_{k \in \mathcal{K}_m} g(K_k), \tag{20}$$

approximately gives the average of the total number of agents within S_m^w that connect to the WAN. We replace N_m^w in (16) via its average approximated by (20) and treat K_k for each $k \in \mathcal{K}_m$ as taking continuum values. This allows us to reduce Problem 2 to a nonlinear programming one, at which point we use Kuhn-Tucker conditions to arrive at the approximate solution for the set $\{K_k\}_{k \in \mathcal{K}_m}$.

Proposition 4: We have $K_k = K_l$ if $B_m^w(h_k) = B_m^w(h_l)$. Furthermore, if $K_k \neq 0$ then

$$\frac{\mathbb{P}(\eta > K_k)}{B_m^w(h_k)} = \nu,$$

for some constant ν, such that the set $\{K_k\}_{k \in \mathcal{K}_m}$ obeys:

$$\sum_{k \in \mathcal{K}_m} \frac{f}{B_m^w(h_k)} g(K_k) = \left(\frac{1}{\theta L\left(\{K_k\}_{k \in \mathcal{K}_m}\right)} + \gamma \right)^{-1}.$$

Simulation results. We implemented the approximate solution of Problem 2 given by Proposition 4 when the traffic parameters and the geometry of WAN network is as we had for Scenario 2 in Section IV. Within a single service zone that has size of an average typical WAN service zone, we simulated 50 sites with different WAN rate, where rates were generated randomly and independently for each site as described in the setup of Scenario 2. The results of the optimization for a particular realization of WAN rates, are shown in Figure 9. Note that the bandwidth is allocated only to hotspots at the sites that experience the worst WAN rate. We find that on average the total bandwidth required for hotspots in the WAN service zone is less than 2.5Mbps, once appropriate backhaul is allocated to optimally selected sites. For comparison, to guarantee the same average performance in the setup of the Scenario 2, where each hotspot was allocated the same bandwidth, one needs about 7.5 Mbps total to be allocated to hotspots on average per service zone of a WAN AP.

VI. Conclusion

In this paper we have taken a first step towards analyzing a possible future wireless network landscape which incorporates heterogenous technologies, e.g., WAN, LAN, Bluetooth etc. In order to allow end nodes to leverage available resources, end nodes will be equipped with a multiple (or flexible) interfaces enabling them to access among various services. Given the complexity of such systems, and to achieve a degree

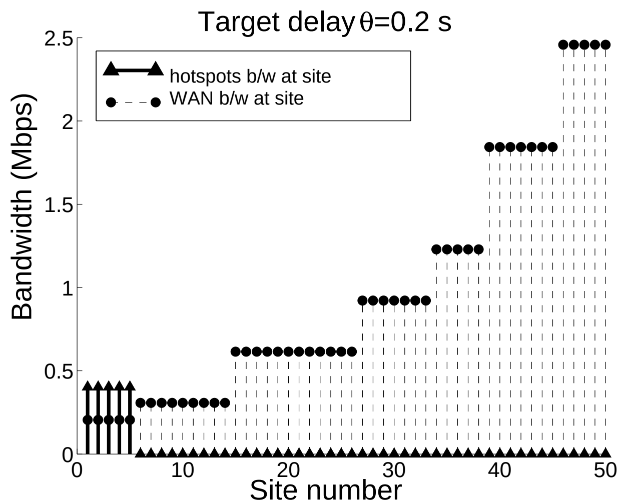

Fig. 9. Distribution of hotspots' backhaul across sites: the dashed stems with circular endings show (ordered) WAN rate at each of the 50 sites, the solid stems with triangular endings – the (optimized) bandwidth for hotspots installed at these sites.

of scalability, it makes sense to allow end nodes to decide which services are preferable, at a given point in time. However, in this context the criterion used to make such decisions becomes an important part of the overall system design. It will not only impact the performance that the user population will see, but also, the resources (e.g., backhaul links density of access points) the providers need to put into place to handle the traffic loads.

To our knowledge this is the first effort to attempt to model and evaluate such heterogeneous systems. In this paper we have shown that even a complex spatial and congestion dependent decision-making process will likely have nice convergence properties to sets of equilibria. From a performance perspective we showed that congestion dependent decision making is likely to provide on average much better performance to users than simple proximity based strategies, but only when hotspot bandwidth are limited. Since backhaul links corresponds to high recurring costs, it is at this point not unreasonable to expect these to be dimensioned conservatively and thus enabling congestion-dependent decision making by end nodes when presented with WAN and hotspot service options to be worthwhile.

At the same time, in this paper we address the complementary problem of joint network design for a system incorporating WAN and set of hotspots to support a spatially distributed set of users. The key insight, is that WAN capacity is particularly valuable, because it permits statistical multiplexing of spatial fluctuations in user loads over a wide area. By contrast hotspots have the potential to substantially and inexpensively enhance capacity in a restricted area. Thus under the uniform loads investigated in this paper, it is the case that WAN resources are typically used as much as possible with only the necessary bandwidth allocated across hotspots to alleviate overloads on the WAN. However, if there are spatial inhomogeneities in the capacity the WAN can provide to users, or in the characteristics of the load, the synergies between these technologies take a different form. Indeed one may conclude that hotspots and the associated backhaul bandwidth is truly worthwhile at spatial locations with a a high steady (i.e., low variance) offered load and where the WAN is not able to provide reasonable service. Our work shows that a joint system design is likely to exploit such variations in order to reduce overall system cost significantly.

APPENDIX I
PROOF OF THEOREM 1.

Assumption 3 yields the following technical Lemma, that is used to prove Theorem 1.

Lemma 1: Consider $a_i, a_j \in \pi^a(S_k^h)$ where $h_k \in S_m^w$ and suppose Assumptions 2 and 3 hold. Furthermore let a_j be connected to h_k and suppose a_i switches from WAN AP w_m to h_k at time s. Then a_j can not switch from hotspot h_k to WAN AP at time $t > s$ if no other agent has switched in S_m^w during the time interval (s,t).

Proof: We prove the lemma by contradiction. Assume that an agent a_j has switched from h_k to WAN AP w_m at time t, then the following condition must have been satisfied:

$$U_j^w\Big(N_m^w(t^-)+1\Big) > U_j^h\Big(N_k^h(t^-)\Big) + c^w. \tag{21}$$

Furthermore, suppose agent a_i switched from the WAN AP w_m to hotspot h_k at time $s < t$, and thus:

$$U_i^w\Big(N_m^w(s^-)\Big) + c^h \leq U^h\Big(N_k^h(s^-)+1\Big). \tag{22}$$

Since we have assume that no other agent within S_m^w has switched in time interval (s,t), we have $N_m^w)(s^-) = N_m^w(t^-)+1$ and $N_k^h(s^-) = N_k^h(t^-)-1$. Now combining (21) and (22), we have:

$$U_j^w\Big(N_m^w(t^-)+1\Big) - U_j^w\Big(N_m^w(t^-)+1\Big) > c^h + c^w,$$

that is in contradiction to Assumption 3. ■

Now we give the proof of Theorem 1.

Proof: Note that under the assumptions of Theorem 1, the dynamics for the configuration of agents' decisions in S_m^w follow a continuous time Markov chain with state $\mathcal{X}(t) := \{X(a_i,t) |\ a_i \in \pi^a(C_m)\}$, where $X(a_i,t) \in \{0,1\}$ – denotes the "connection state" of the agent a_i at time t and takes the value 0 if the agent is connected to a hotspot and 1 if it is connected to a WAN AP. (Note that we need only to consider the states of agents located within C_m.) We will classify transitions for this chain as "up", "down" and "stay", corresponding to agents switching from hotspots to the WAN AP, vice versa, or staying with their current choice. For simplicity we can uniformize the continuous-time chain to focus on a discrete time Markov chain capturing times where decisions are made. We shall denote these decision times by $s = 1, 2, \ldots$. The transition probabilities for the discrete Markov chain are determined by two factors: the probability that a particular agent reconsiders her decision at that time, and whether the current configuration cause the agent to change providers.

By Assumption 1, each service zone contains an a.s. finite number of agents, thus there is an a.s. finite number of different configurations for agents' choices so the set of possible configurations is finite a.s.. It follows that some of the states must be revisited by the chain infinitely often. To show the convergence of a system to an equilibrium, it is sufficient to construct a feasible path for the chain evolution which hits an equilibrium state with positive probability, *starting from any initial configuration*.

Below we present the steps of an algorithm to construct a path $\mathcal{P}$ consisting of a sequence of transitions for the state $\mathcal{X}(s)$, which, starting from any arbitrary configuration of agents' choices $\mathcal{X}(0)$, ends up in an equilibrium configuration after a finite number of steps. Let $A^u(s)$ denote the set of agents that, given the configuration at time s, could make "up" transitions and $A^d(s)$ the set of agents that can make "down" transitions. Let us also define a nondecreasing composite function,

$$J_i(N) \triangleq (U_i^w)^{-1} \circ (U_i^h(N) + c^w) \,,$$

where $(U_i^w)^{-1}$ denotes a unique and decreasing, due to Assumption 2 inverse of U_i^w. We describe our algorithm in terms of pseudo-code shown in Table III, where for convenience we denote $N^h(a_i,t) = N_k^h$, where k is such that $a_i \in S_k^h$. Note that our notational convention is that an agent making her decision at time slot $s \geq 1$ is basing this decision by observing the state of the system prior to that time, i.e. time $s-1$.

After initialization, the algorithm (see Table III) alternates between the Up- and Down- transition phases. During the Up-transition phase only the "up"-switchings occur, where the agents performing these transitions are selected to be those which are the most "unsatisfied". This phase ends once the set of agents that are able to perform the "up"-transitions depletes. At that time the algorithm switches to the "down"-transition phase, where at most one agent performs a "down"-transition. We introduce an auxiliary integer sequence $\{Z(t)\}_{t=1}^{\infty}$ with values that depend on the state of the system prior to a transition at times $t = 1, 2, \ldots$, and show that this sequence is nonincreasing. This allows us to argue that $Z(t)$ converges to a limit Z^* after an a.s. finite number of transitions. Then, we demonstrate that the equilibrium must be reached in a.s. finite time once $Z(t)$ has reached the level Z^*. ■

Initialization:
$s = 1$ and $\mathcal{X}(s) = \mathcal{X}(0)$
$Z(s) := 0$
go to Up-transition phase

Up-transition phase:
if $A^u(s) \neq \emptyset$
{ $j := \arg\max_{i:\ a_i \in \pi^a(A^u(s))} \left\lfloor J_i\left(N^h(a_i,s)\right)\right\rfloor$
$Z(s) := \left\lfloor J_j\left(N^h(a_j,s)\right)\right\rfloor$
let a_j make an "up" transition
update the state $\mathcal{X}(s)$
$s := s+1$ }
otherwise: go to Down-transition phase

Down-transition phase:
if $A^d(s) \neq \emptyset$:
{ choose any $a_j \in A^d(s)$
let a_j make a "down" transition
update the state $\mathcal{X}(s)$
$Z(s) := Z(s-1)$
$s := s+1$
go to Up-transition phase }
otherwise: done

TABLE III
PSEUDO-CODE FOR CONSTRUCTING THE PATH $\mathcal{P}$ CONVERGING TO EQUILIBRIUM.

Note that if the algorithm does not enter an Up-transition phase then there can only be "down" transitions in the system. Since the number of agents that are connected to each WAN AP is finite, the system will inevitably converge to an equilibrium which has no agents connected to the WAN AP w_m. Instead, assume that the system enters the Up-transition phase at time t_0. We will show that the sequence $Z(s)$, $s = t_0, t_0+1, \ldots$, defined in Table III, is a non-increasing sequence.

We start by relating the function $J_j(\cdot)$ to agent a_j's eligibility for an "up" transition at time t. We must have:

$$U_j^w\left(N_m^w(t-1)+1\right) > U_j^h\left(N^h(a_j,t-1)\right) + c^w \,,$$

for an agent to be eligible to switch "up" at time $t \geq 1$. This is equivalent to:

$$N_m^w(t-1) < J_j\left(N^h(a_j,t-1)\right) - 1 \,, \tag{23}$$

which in turn can be strengthened to:

$$N_m^w(t-1) \leq \left\lfloor J_j\left(N^h(a_j,t-1)\right)\right\rfloor - 1 \tag{24}$$

with a strict inequality in (24) if $J_j\Big(N^h(a_j,t-1)\Big) \in \mathbb{N}$.

Now consider any Up-transition phase. Note that $\Big\lfloor J_i\Big(N^h(a_i,s)\Big)\Big\rfloor$ can only decrease for each agent $a_i \in S_m^w$. Indeed, for each i, the function $J_i(\cdot)$ is nondecreasing and $N^h(a_i,s)$ could only be reduced during an Up-transition phase. Now, since the number of agents connected to the WAN AP w_m could only increase and by Assumption 2, $U_i^w(\cdot)$ is a decreasing function, the value $U_i^w(N_m^w(s))$ can only decrease during an Up-transition phase. Clearly, by the general eligibility requirement (24), we have that the set of agents eligible for "up" transitions can only diminish within the Up-transition phase. Hence $A^u(s_1) \subset A^u(s_2)$, when $s_1 < s_2$ are both restricted to the period of the same Up-transition phase. Thus, for such s_1 and s_2:

$$\begin{aligned} Z(s_1) &= \max_{i: a_i \in A^u(s_1)} \Big\lfloor J_i(N^h(a_i,s_1))\Big\rfloor \\ &\geq \max_{i: a_i \in A^u(s_2)} \Big\lfloor J_i(N^h(a_i,s_2))\Big\rfloor = Z(s_2)\,, \end{aligned}$$

and hence $Z(s)$ is a nonincreasing sequence whenever s is within a single Up-transition phase.

We now show that $Z(s)$ is in fact nonincreasing for all $s \geq t_0$. Suppose that an Up-transition Phase finished at time $\tau+1$, and a_j was the agent that switched "up" at time τ, hence $Z(\tau) = \Big\lfloor J_j\Big(N^h(a_j,\tau-1)\Big)\Big\rfloor$. We will consider two scenarios. In the first scenario there is only one "down" transition at time $\tau+1$ and $A^u(\tau+2)$ becomes nonempty. We will show that in this scenario $Z(\tau+2) \leq Z(\tau)$. In the second scenario there is a sequence of $n>1$ "down" transitions, before the set $A^u(\tau+n+1)$ becomes nonempty for the first time. In this scenario we will show once again that $Z(\tau+n+1) \leq Z(\tau)$.

Scenario 1: $A^u(\tau+2) \neq \emptyset$**.** Observe that once an agent a_i has performed a "down" transition at time $\tau+1$, we have:

$$J_k\Big(N^h(a_k,\tau+1)\Big) = J_k\Big(N^h(a_k,\tau)\Big)\,,$$

for all agents $a_k \in S_m^w$ that do not fall within the service zone of the same hotspot as a_i. For such agents we also have that:

$$\Big\lfloor J_k\Big(N^h(a_k,\tau)\Big)\Big\rfloor \leq \Big\lfloor J_j\Big(N^h(a_j,\tau)\Big)\Big\rfloor = Z(\tau)\,,$$

since a_j was chosen to make an "up" transition at time τ. Hence we have that for each agent that does not fall within service zone of the same hotspot as a_i:

$$\Big\lfloor J_k\Big(N^h(a_k,\tau+1)\Big)\Big\rfloor \leq \Big\lfloor J_j\Big(N^h(a_j,\tau)\Big)\Big\rfloor = Z(\tau)\,. \tag{25}$$

Now, by Lemma 1, no agent a_k that falls in the service zone of the same hotspot as a_i could switch "up" immediately after a_i has switched "down", and thus $a_k \notin A^u(\tau+2)$. But then, in view of (25) and the definition for $Z(t)$, we conclude that:

$$Z(\tau+2) \leq Z(\tau).$$

Scenario 2: $A^u(\tau+l) = \emptyset$ **for** $l = 1,\ldots n$ **and** $A^u(\tau+n+1) \neq \emptyset$**.** We will show that

$$Z(\tau+n+1) \leq Z(\tau)\,, \tag{26}$$

by contradiction. Assume that the inequality (26) is not satisfied. Then, we must have that:

$$\Big\lfloor J_k\Big(N^h(a_k,\tau+n-1)\Big)\Big\rfloor > \Big\lfloor J_j\Big(N^h(a_j,\tau)\Big)\Big\rfloor\,, \tag{27}$$

for some agent a_k within C_m. Indeed, consider an agent a_i that switches "down" at time $\tau+n$. Since a_i's switching down does not affect the number of agents connected to hotspots that do not contain a_i in their service zone, we have:

$$J_r\Big(N^h(a_r,\tau+n-1)\Big) = J_r\Big(N^h(a_r,\tau+n)\Big)\,,$$

for all agents $a_r \in S_m^w$ that do not fall within the service zone of the same hotspot as a_i. Moreover, by Lemma 1, no agent a_p that belongs to the service zone of the same hotspot as a_j can be eligible for an "up" transition at time $\tau+n+1$, i.e. $a_p \notin A^u(\tau+n+1)$. Hence, if

$$\begin{aligned} &\max_{l: a_l \in A^u(\tau+n+1)} \Big\lfloor J_l\Big(N^h(a_l,\tau+n)\Big)\Big\rfloor \\ &\qquad = Z(\tau+n+1) > Z(\tau)\,, \end{aligned}$$

then

$$\max_{l: a_l \in S_m^w} \Big\lfloor J_l\Big(N^h(a_l,\tau+n-1)\Big)\Big\rfloor > Z(\tau)\,,$$

which translates into (27).

Next we show that the agent a_k, where k satisfies (27), was eligible to switch "up" at time $\tau+n$. Consider again an agent a_j that switched "up" at time τ. To be eligible for making an "up" switch at time τ, according to (24) we must have:

$$N_m^w(\tau-1) \leq \Big\lfloor J_j\Big(N^h(a_j,\tau-1)\Big)\Big\rfloor - 1 \tag{28}$$

with a strict inequality in (28) if $J_j\Big(N^h(a_j,\tau-1)\Big) \in \mathbb{N}$. Now consider the agent a_k, and note that

$$N_m^w(\tau+n-1) \leq N_m^w(\tau-1)\,, \tag{29}$$

since one agent has switched "up" at time τ and at least one agent has switched "down" at time interval $(\tau,\tau+n-1]$. Considering the agent a_k at time $\tau+n-1$ in view of (27), (28) and (29) we obtain

$$N_m^w(\tau+n-1) < \Big\lfloor J_k\Big(N^h(a_k,\tau+n-1)\Big)\Big\rfloor - 1\,.$$

This leads to:

$$U_k^w\left(N_m^w(a_k,\tau+n-1)+1\right) > U^h\Big(N^h(a_k,\tau+n-1)\Big),$$

and hence the agent a_k was eligible for an "up" transition at time $\tau+n$. We thus have a contradiction with the assumption that no agents were eligible for "up" transitions in the interval $(\tau,\tau+n]$. This proves that the inequality (26) holds.

To summarize we have proved that:

1) $Z(s)$ is nondecreasing if s is restricted to the period of a single Up-transition phase
2) If $\tau+1$ is the time when an Up-transition phase has finished and $\tau+n+1$, for $n\geq 1$ is the time when the next Up-transition phase has started, then: $Z(\tau)\geq Z(\tau+n+1)$.

Therefore $Z(s)$ is a nonincreasing sequence and since $Z(s)$ is integer valued it must have an integer-valued limit Z^* which is reached by the sequence in a.s. finite time t_1. We are left to show that the algorithm needs a.s. finite number of steps before an equilibrium is in fact reached.

For $t\geq t_1$, we have $Z(t)=Z^*$ and there are three scenarios for system evolution. The first scenario corresponds to the case where only "down" transitions take place in the system, and in the second – only "up" transitions are possible. In both of these scenarios the system reaches an equilibrium once all agents in C_m have switched to either the WAN or their respective hotspots. The third scenario is when the system undergoes both "up" and "down" transitions that are intermingled and we consider this scenario below.

From (24) we obtain that the agent is eligible for an "up" transition at time $t>t_1$ if and only if:

$$J_j(N^h(a_j,t-1)) = Z^*+\eta, \tag{30}$$

where $\eta\in[0,1)$, and

$$N_m^w(t-1)\leq Z^*-1, \tag{31}$$

with strict inequality if $J_j(N^h(a_j,t-1))\in\mathbb{N}$. Now, assume that an Up-transition phase has ended at time $\tilde{\tau}+1>t_1$. Then this phase could finish either of the conditions (30) or (31) or both were violated at time $t=\tilde{\tau}+1$ for all agents $a_j\in C_m$.

Suppose that at time $t=\tau+1$ inequality (30) is violated for all $a_j\in C_m$, but inequality (31) is not. It is sufficient to show that no agent can become eligible for an "up" transition ever again at times $t\geq\tilde{\tau}+1$, since then the algorithm exits once all agents in C_m have connected to their respective hotspots. To show this, we note first that if for any agent $a_j\in C_m$

$$J_j(N^h(a_j,\tau))\geq Z^*+1 \tag{32}$$

then, from (31)

$$N_m^w(\tau) < J_j(N^h(a_j,\tau))-1,$$

which indicates, via (23) that a_j is eligible to switch "up" at time $\tilde{\tau}$. This is in contradiction to what we assumed in the beginning of the paragraph and hence we can only have:

$$J_j(N^h(a_j,\tau))\leq Z^*-1, \tag{33}$$

for any agent $a_j\in C_m$. Now observe that, by (30) an agent $a_j\in C_m$ may become eligible for an "up" transition at time $t>\tilde{\tau}+1$ only if $J_j(N^h(a_j,t))$ has increased to or above level Z^*. Since $J_j(\cdot)$ is nondecreasing, there must be a "down" transition that would occur within the hotspot that contains a_j in its service zone. But if such "down" transition happens at time $\tilde{\tau}+1$, no agent within the service zone of a hotspot containing a_j can become eligible for an "up" transition at time $\tilde{\tau}+2$ as we proved in Lemma 1. Clearly we have that (31) is still satisfied for $t=\tilde{\tau}+1$, but then we must have

$$J_j(N^h(a_j,\tilde{\tau}+1)) < Z^*, \tag{34}$$

for any $a_j\in C_m$ since no agent is still available for an "up" transition. By induction we can prove that no agent is eligible for an "up" transition at any time $t\geq\tilde{\tau}+2$.

Now suppose that (31) is violated at time $t=\tilde{\tau}+1$. Note that the condition (31) was met at time $t=\tilde{\tau}$, since an "up" transition occurred at time $\tilde{\tau}$. Due to this "up" transition, we also have $N_m^w(\tilde{\tau})=N_m^w(\tilde{\tau}-1)+1$ which yields $N_m^w(\tilde{\tau})=Z^*$. Now let $\mathcal{L}(\tilde{\tau}+1)$ denote the set of agents for which:

$$\left\lfloor J_k(N^h(a_k,\tilde{\tau}))\right\rfloor = Z^*. \tag{35}$$

If no "down" transition occurs at time $\tilde{\tau}+1$ then the algorithm has exited and thus an equilibrium has been reached. Otherwise, assume that the "down" transition at time $\tilde{\tau}+1$ was in the service zone S_l^h of a hotspot $h_l\in S_m^w$. Note, that by Lemma 1, no agent that falls within S_l^h can become eligible for an "up" transition at time $\tilde{\tau}+2$, thus, since $N_m^w(\tilde{\tau}+1)=Z^*-1$, the condition (30) must be violated for those agents at time $\tilde{\tau}+2$. Hence these agents could not be within the set $\mathcal{L}(\tilde{\tau}+1)$ since otherwise they would be eligible for an "up" transition at time $\tau+2$. Furthermore, transition to S_l^h does not change the number of agents connected to hotspots $h_n\neq h_l$, and thus we have:

$$J_i(N^h(a_i,\tilde{\tau}+1)) = J_i(N^h(a_i,\tilde{\tau}))$$

for $a_i\notin S_l^h$. We thus conclude that $\mathcal{L}(\tilde{\tau}+1)=\mathcal{L}(\tilde{\tau}+2)$.

Now if $\mathcal{L}(\tilde{\tau}+1)=\emptyset$ then, similarly to as we argued above, no agent can ever become eligible for an "up" transition and the algorithm exits in finite time. Otherwise, if $\mathcal{L}(\tilde{\tau}+1) \neq \emptyset$ then at least one agent $a_j \in \mathcal{L}(\tilde{\tau}+2)$ is eligible for an "up" transition at time $\tilde{\tau}+2$ by sufficient conditions (30)-(31), since $N_m^w(\tilde{\tau}+1) = Z^* - 1$ and $\lfloor J_j(N^h(a_j, \tilde{\tau}+1)) \rfloor = Z^*$. Thus a_j can perform an "up" transition, which can only diminish the set $\mathcal{L}$ at a subsequent time $\tilde{\tau}+3$. By induction we thus can show that the set $\mathcal{L}(t)$ necessarily depletes in finite time, whence the algorithm exits.

In summary we have shown that from any starting configuration there exists a path, that with positive probability reaches an equilibrium state. Since the state space is finite, there must be a state which is visited infinitely often. Whence the Markov chain will necessarily eventually hit an equilibrium state.

APPENDIX II
PROOF OF PROPOSITION 1

Proof: Observe, that distribution of the agents is homogeneous, thus by symmetry we must allocate the same amount of bandwidth to each of the hotspots. To show the optimal allocation for the first regime, consider an allocation strategy $\mathcal{A}_1$, where each of the hotspots is given B^h units of bandwidth. Then a hotspot could serve at most $N^h(\theta, B^h)$ agents, where

$$N^h(\theta, B^h) = \left\lfloor \frac{B^h}{\gamma f} - \frac{1}{\theta\gamma} \right\rfloor .$$

The total number of agents which hotspots could serve is:

$$N_1 = H_m \left\lfloor \frac{B^h}{\gamma f} - \frac{1}{\theta\gamma} \right\rfloor \leq \frac{H_m B^h}{\gamma f} - \frac{H_m}{\theta\gamma} . \qquad (36)$$

Now consider an allocation strategy $\mathcal{A}_2$ that shifts ΔB units of bandwidth from each hotspot to the WAN AP, where $\Delta B < B^h$ is such that

$$\frac{B^h - \Delta B}{\gamma f} - \frac{1}{\theta\gamma} \in \mathbb{N} .$$

We will assume that the WAN AP uses the shifted bandwidth to serve the agents within the hotspots, Then, from (5) and (6) the total number of agents in C_m that could be served by such system, without violating the delay requirement is:

$$\begin{aligned} N_2 &= H_m \left\lfloor \frac{B^h - \Delta B}{\gamma f} - \frac{1}{\theta\gamma} \right\rfloor + \left\lfloor \frac{H_m \Delta B}{\gamma f} - \frac{1}{\theta\gamma} \right\rfloor \\ &\geq H_m \frac{B^h - \Delta B}{\gamma f} - \frac{H_m}{\theta\gamma} + \frac{H_m \Delta B}{\gamma f} - \frac{1}{\theta\gamma} - 1 \\ &= \frac{H_m B^h}{\gamma f} - \frac{H_m + 1}{\theta\gamma} - 1 \geq N_1 - 1 - \frac{1}{\theta\gamma} . \end{aligned} \qquad (37)$$

Now note that under allocation strategy $\mathcal{A}_1$, N_1 agents in total could be served in the hotspots only if each of them in fact had $N^h(\theta, B^h)$ agents to serve. However, if the system gets large enough ($H_m \gg 1$) with probability arbitrary close to 1 there is at least $\left\lceil 1 + \frac{1}{\theta\gamma} \right\rceil$ hotspots containing at most $N^h(\theta, B^h) - 1$ agents at their service zones. We thus obtain that the performance under allocation strategy $\mathcal{A}_2$ is at least as good as the performance under $\mathcal{A}_1$ once H_m is large enough. This shows that if (9) holds then a policy that allocates $B_m^w = \gamma f \hat{M}_m^w + \frac{f}{\theta}$ to the WAN AP w_m is optimal for sufficiently large H_m.

Consider the regime where the inequality (9) is not satisfied. Then, a policy which is optimal for large enough H_m allocates $B_m^w = f\gamma \hat{N}_m^w + f\theta$ units of bandwidth to the WAN AP, where $\hat{N}_m^w$ is given by (8). For a particular realization of agents let the number of agents that do not fall within a service zone of any of the hotspots be denoted as $M_{\bar{C}_m}^w$ and let

$$\Delta N_m^w = \hat{N}_m^w - M_{\bar{C}_m}^w . \qquad (38)$$

If $\Delta N_m^w > 0$ then ΔN_m^w agents can be served by WAN AP in any of the hotspots without violating the delay constraint at the WAN AP. Clearly, the optimal way to use the extra bandwidth is to serve agents from the most congested hotspots. In particular, assuming that agents in C_m are initially connected to their hotspots, an algorithm that selects which agents the WAN would serve, at each step takes the most congested hotspot and switches an agent connected to this hotspot to the WAN AP. As a result, the number of agents connected to hotspots and the WAN will be represented by the Figure 2 with a "slicing" plane at some level.

If, however, for some realization of agents $\Delta N_m^w < 0$ then no agents inside the hotspots' service zones can be served by the WAN. In this case, the agents in $\bar{C}_m$ connected to the WAN do not meet their delay requirement. Denote F^w the event:

$$F^w = \{\hat{N}_m^w < M_{\bar{C}_m}^w\} . \qquad (39)$$

Clearly if $\mathbb{P}(F^w) > \delta$, then the optimization problem (1) does not have a solution that meets the probabilistic requirement (6) (statement (ii) of the proposition).

Thus, in the remaining case, we assume that $\mathbb{P}(\underline{F^w}) < \delta$, and hence the delay requirement of the agents in $\bar{C}_m$ is always met. Below we find the minimum bandwidth that has to be allocated to hotspots so that the agents within C_m meet their delay requirement too. We fix $K > 0$ and let:

$$B^h(\theta, K) = f\gamma K + f\theta .$$

Thus $B^h(\theta, K)$ is the amount of bandwidth that each hotspot has to be supplied to serve up to K agents within its service area. Assume that $B^h(\theta,K)$ is indeed provided to each of the hotspots, then the event that any hotspot h_k has more then K agents connected to it is equivalent to the event F^h which has:

$$\sum_{k\in\mathcal{K}_m}(M_k^h-K)\mathbf{1}_{\{M_k^h>K\}}>\Delta N_m^w\,, \tag{40}$$

Clearly the event F^w implies the event F^h, and thus, guaranteeing that F^h does not occur is enough to guarantee that F^w does not occur either. Thus, via plugging (38) into (40) we have that the value of K is as given in part (iii) of the proposition. ■

APPENDIX III
PROOF OF PROPOSITION 4

We will derive the approximate solution to Problem 2 under the assumption[12] that $K_l=K_k$ when $b_l^w=b_k^w$, and $l,k\in\mathcal{K}_m$. First, we will elaborate on the expression (16) for $\bar{D}_m^w$. Note, that we can express the number of agents N_m^w connected to WAN AP w_m via the set $\{K_k\}_{k\in\mathcal{K}_m}$ as follows:

$$\begin{aligned}N_m^w &= \sum_{k\in\mathcal{K}_m}(M_k^h-K_k)\mathbf{1}_{\{M_k^h>K_k\}}\\ &=\sum_{r=1}^{N_B}\sum_{\{k\in\mathcal{K}_m|B_m^w(h_k)=b_r^w\}}(M_k^h-K_k)\mathbf{1}_{\{M_k^h>K_k\}}\,. \end{aligned}\tag{41}$$

Let $n_m(r)$ denote the number of sites in S_m^w with WAN rate equal to b_r^w. In the limit $n_m(r)\gg 1$ we can apply the Central Limit Theorem in (41), to obtain that:

$$N_m^w=\sum_{r=1}^{N_B}n_m(r)\xi_r\,,$$

where ξ_r is a normally distributed random variable with expectation and variance equal to

$$g(K_r)\triangleq\mathbb{E}\left[(M_r^h-K_r)\mathbf{1}_{\{M_r^h>K_r\}}\right]\,.$$

(Note, that since by our assumption S_k^h have the same sizes for all $k\in\mathcal{K}_m$, we have that M_k^h has the same distribution for all $k\in\mathcal{K}_m$. Hence, $g(K_r)$ depends only on K_r.) Thus, in the limit when $n_m(r)\gg 1$ for $r=1,\ldots,N_B$, we have that N_m^w is normally distributed with mean and variance equal to:

$$L\left(\{K_k\}_{k\in\mathcal{K}_m}\right)\triangleq\sum_{r=1}^{N_B}n_m(r)g(K_r)\,.$$

[12]Note that the optimal solution to Problem 2 might not have this property. However, one can show that for the solution to the original optimization problem 1 such property holds.

Similarly, elaborating on the sum Σ that appears in (16), we have:

$$\begin{aligned}\Sigma &\triangleq \sum_{a_j\in\mathcal{W}_m}\frac{f}{B_m^w(a_j)}=\sum_{k\in\mathcal{K}_m}\frac{f}{B_m^w(h_k)}(M_k^h-K_k)\mathbf{1}_{\{M_k^h>K_k\}}\\ &=\sum_{r=1}^{N_B}\frac{f}{b_r^w}\sum_{\{k\in\mathcal{K}_m|B_m^w(h_k)=b_r^w\}}(M_k^h-K_k)\mathbf{1}_{\{M_k^h>K_k\}}\\ &=\sum_{r=1}^{N_B}\frac{f\,n_m(r)}{b_r^w}\xi_r\,.\end{aligned}$$

Thus, in the limit when $n_m(r)\gg 1$ for $r=1,\ldots,N_B$, we have that Σ is normally distributed with mean and variance equal to:

$$\sum_{r=1}^{N_B}\frac{f\,n_m(r)}{b_r^w}g(K_r)$$

Now, it is simple to see that the constraint (18) reduces to requiring that:

$$\sum_{r=1}^{N_B}\frac{f\,n_m(r)}{b_r^w}g(K_r)\le\left(\frac{1}{\theta L\left(\{K_k\}_{k\in\mathcal{K}_m}\right)}+\gamma\right)^{-1}+\varepsilon\,, \tag{42}$$

where the variable ε depends on δ and is proportional to the variances of N_m^w and Σ. Note, that the variances of N_m^w and Σ scale as the square root of their respective averages. Thus, when both N_m^w and Σ are large on average, and δ is "moderately small", we can neglect ε in (42). We thus arrive into the following optimization problem:

$$\min\sum_{r=1}^{N_B}n_m(r)\left(\gamma fK_r+\frac{f}{\theta}\mathbf{1}_{\{K_r>0\}}\right)\,.$$

under constraint:

$$\sum_{r=1}^{N_B}\frac{f\,n_m(r)}{b_r^w}g(K_r)\le\left(\frac{1}{\theta L\left(\{K_k\}_{k\in\mathcal{K}_m}\right)}+\gamma\right)^{-1}\,. \tag{43}$$

It is simpler to treat this problem assuming each K_r takes continuum of values instead of in discrete set. To that effect we might compute $g(K_r)$ replacing M_k^h, for all $k\in\mathcal{K}_m$, by normal random variables η_k, that have the same average and the variance as M_k^h. (This step is supported by the fact that when $\mathbb{E}M_k^h$ is large enough the cdf of a Poisson random variable does not differ much at integer points from the cdf of the corresponding normal random variable.) Also, when $\gamma\theta\gg 1$ (delay requirement is not very stringent) we could eliminate the term $\frac{f}{\theta}\mathbf{1}_{\{K_r>0\}}$ from the objective. Then, both the constraint and the objective will be convex functions. Using Kuhn-Tucker conditions we arrive into the requirement that if $\{K_r^*\}$ is an optimal set of values, then if $K_r^*\neq 0$, we have:

$$n_m(r)\gamma f-\mu\frac{f\,n_m(r)}{b_r^w}g'(K_r^*)$$

$$+\mu\frac{\partial}{\partial K_r}\left(\frac{1}{\theta L\left(\{K_k^*\}_{k\in\mathcal{K}_m}\right)}+\gamma\right)^{-1}=0\,, \quad (44)$$

and the constant μ is such that the set of $\{K_r^*\}$ obeys the constraint (43) with equality. Assuming that $n_m(r)$ for each r is large enough, we can neglect the derivative associated with the last term in (44). Then we arrive into a simple requirement for $K_r^* \neq 0$

$$\frac{g'(K_r^*)}{b_r^w}=-\nu\,, \quad (45)$$

where the constant ν is such that is such that the set of $\{K_r^*\}$ obeys the constraint (43) with equality. Elaborating on $g_r'(K)$ we get:

$$g'(K_r)=\left(\int_K^\infty (x-K_r)p(x)dx\right)'=-P(\eta_r>K)\,,$$

where $p(\cdot)$ denotes the pdf of the normal random variable with expectation and variance equal to $\mathbb{E}\left[M_k^h\right]$. Combining this with (45) yields Proposition 4.

REFERENCES

[1] G. Rittenhouse, "Next generation wireless networks," in *Proc. INFORMS*, March 2004.

[2] GTRAN, "GTRAN dual-mode 802.11/CDMA wireless modem," http://www.gtranwireless.com.

[3] F. Baccelli, M. Klein, M. Lebourges, and S. Zuyev, "Stochastic geometry and architecture of communication networks," *J. Telecommunication Systems*, vol. 7, pp. 209–227, 1997.

[4] A. Zemlianov and G. de Veciana, "Modeling competition among wireless service providers," *Tech. Report, UT Austin, available at http://www.ece.utexas.edu/˜zemliano/informs_fin.pdf*, 2004.

[5] CDMA Development Group, "CDMA2000 and WiFi: Making a business case for interoperability," September 2003, http://www.cdg.org.

[6] D. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes*, Springer Verlag, 2 edition, 2002.

[7] F. Baccelli and B. Blaszczyszyn, "On a coverage process ranging from the boolean model to the poisson voronoi tessellation, with applications to wireless communications," *Adv. Appl. Prob.*, vol. 33, no. 2.

[8] A. Zemlianov and G. de Veciana, "Cooperation and decision-making in a wireless multi-provider setting (extended version)," *Tech. Report, UT Austin, available at http://www.ece.utexas.edu/˜zemliano/infocomext.pdf*, 2004.

[9] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *Proc. IEEE INFOCOM' 03*.

[10] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: a bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Commun. Mag.*, vol. 38, no. 7, pp. 70–77, 2000.

[11] S. Borst, K. Kumaran, K. Ramanan, and P. Whiting, "Queueing models for user-level performance of proportional fair scheduling in wireless data systems," 2002.

[12] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*, Prentice Hall, 1989.