Load Balancing of Best Effort Traffic in Wireless Systems Supporting End Nodes with Dual Mode Capabilities

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

Abstract — **We consider a wireless system consisting of a wireless wide area network (WAN) that cooperates with a set of wireless local network (WLAN) access points to serve a spatially distributed customer base. We assume that WAN and WLAN utilize orthogonal, i.e. non-interfering technologies, and that users' devices are equipped with dual-mode interfaces which enable them to communicate with both the WAN and the WLANs. We focus on the downlink and address the problem of optimal interface selection (decision-making) in order to minimize the total queueing backlog in the system. We start by considering an isolated WAN cell and describe algorithms for centralized and distributed implementation for optimal decision-making. We show via simulations that our proposed algorithms enable significant performance gains over more natural strategies (based on proximity) that we use as a benchmark for comparison. We then turn to a multi-cell scenario, and show how one can modify decision-making to factor in the other-cell-interference in the WAN. We construct simulation examples that demonstrate the value of such modifications for achieving better system performance.**

I. Introduction

In this paper we consider a wireless system that combines a wireless wide area network (WAN), engineered to provide uniform spatial coverage, and a set of wireless local area networks (WLANs), each with limited coverage and used to enhance throughput at local "hotspots". This is an example of a "hierarchical overlay", a "multi-tier", or an "umbrella" network, with macro- and micro-cell levels corresponding to the WAN and WLANs coverage areas respectively. Such systems are likely to become increasingly pervasive in the future, and each constituent network will be devised to meet its own engineering design goals [1, 2]. The problem of efficient design of such networks was previously addressed in the context of cellular voice applications [3]–[6]. Emerging integrated 3G and WiFi networks [7]–[11] and dual-mode wireless devices [12] have also stimulated research on how to design such systems to efficiently handle data [13]–[15].

We will consider networks where WAN and WLANs use non-overlapping portions of spectrum, users are spatially distributed and have dual-mode wireless devices capable of communicating with both WAN and WLANs. Users generate requests for data downloads, i.e. the traffic demand is asymmetric in that it assumes the down-link traffic greatly exceeds that of the up-link. We assume that requests from users not covered by any WLAN can only be routed to the WAN, while there is flexibility in routing requests from users covered by *both* WAN and WLAN.

In such networks the design of users' assignment strategies between the layers plays a crucial role [15]. The design depends in part on the application at hand (e.g. we focus on data) and relates to other system aspects, e.g. optimization objective, network capacity and allocation of resources. The purpose of this study is to formulate and evaluate centralized and distributed *load balancing* algorithms which enable such assignment decisions. The implementation of centralized algorithms could be accomplished via a "tightly coupled" internetworking solution [11], while the distributed version could be implemented via software "agents" within individual dual-mode devices.

We will derive load-balancing decision metrics that tie together the average physical channel rates available at users' locations, proximity to access points, utilization of resources and current demand for particular services. The most important feature that distinguishes this work from [14, 15] is how we incorporate interference at the WAN layer as a factor biasing the load-balancing decisions. The model that we develop here is applicable to systems where no power control takes place on the downlink at the WAN layer: a key example of such a system is Qualcomm's 1XEV-DO [16]. In this 1XEV-DO model users are served in a generalized processor-sharing fashion, and only one user is served per time slot with full power allocated at the WAN AP. Thus in such systems the larger the fraction of users routed to a particular WAN AP, the larger the fraction of time that the WAN AP has to be active on average. This in turn makes the WAN AP create more interference to its neighboring WAN APs, and forces degradation in service quality seen by users served by neighboring WAN APs. We show that load-balancing algorithms which factor such interference at the WAN layer lead to improved performance, especially in systems with spatially asymmetric loads.

Our methodology includes devising simple geometric models for service zones of the WAN and WLAN APs these are presented in Section II. Our objective will be to minimize the total queueing backlog in a multi-cell heterogeneous wireless system, thus our modeling approach involves some queueing analysis and approximations. In Section III under the assumption that the other-cell interference at the WAN layer is negligible we derive decision-metrics that can be used as a basis for load-balancing algorithms. In Section IV we describe how one can correct the decision-metrics so as to incorporate the "cost" of other-cell interference at the WAN. The corrected algorithms perform equally well in scenarios with both significant and small other-cell interference. We demonstrate this in Section V where we report our simulation results.

II. The model

Network geometry. In this section we introduce a simple geometric model that captures the key features of spatial interactions between the WAN and WLANs. In our setup, there are M WAN and K WLAN APs which are placed within a region D of a plane. We will assume that the locations of the WAN and WLAN APs are distinct and denoted by $\{w_m\}_{m=0}^{M-1}$, and $\{h_k\}_{k=0}^{K-1}$ respectively (see Figure 1). In what follows, with some abuse of notation we will refer to the AP located at point x as "AP x".

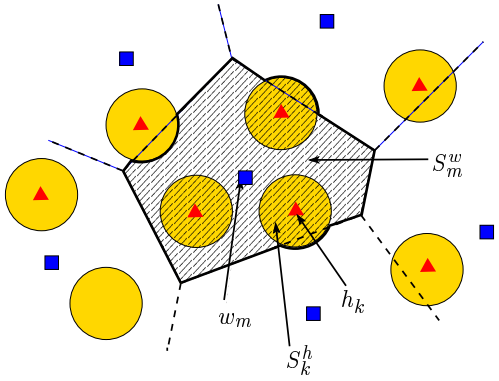

Fig. 1: Example of defining WAN and WLAN service zones to ensure Assumptions 1-3.

We denote by S_m^w (S_k^h) the service zone of the WAN AP w_m (WLAN AP h_k), where the service zone of an AP is a set consisting of spatial locations that the AP can serve. Note that in practice the geometry of these zones would depend on the underlying technology. Thus, for example, in an IS-95 system a mobile decides whether it belongs to the service zone of a particular AP by comparing the strength of pilot signals in its vicinity. However, to simplify the derivation of our load-balancing algorithms we will make the following assumptions:

Assumption 1. *The WAN has uniform coverage, i.e.* $\bigcup_{m=0}^{M-1} S_m^w = D$.

Assumption 2. *The service zones of distinct WAN (WLAN) APs are disjoint, i.e.* $S_{m_1}^w \cap S_{m_2}^w = \emptyset$ *and* $S_{k_1}^h \cap S_{k_2}^h = \emptyset$, *for* $m_1 \neq m_2$, *and* $k_1 \neq k_2$.

Assumption 3. *The service zone of any WLAN AP is fully contained within a service zone of some WAN AP, i.e. for all* $k = 0, 2, \ldots, K-1$, $S_k^h \subset S_m^w$, *for some* m, $0 \leq m \leq M-1$.

Assumption 1 is typically true when the WAN utilizes a portion of licensed spectrum and APs use large enough powers for their down-link transmissions. Assumption 2 is valid for scenarios where WLANs are sufficiently spaced or when they operate using orthogonal spectra. It also holds for some current WAN technologies, but would not be true for systems implementing a "soft handoff". For these systems a mobile "on the cell boundary" may be simultaneously served by several WAN APs, however Assumption 2 can still be accepted as a reasonable, first order approximation. Finally, Assumption 3 is motivated by the fact that WAN and WLAN technologies operate at significantly different spatial coverage scales.

Traffic assumptions. We will concentrate on the so-called semi-dynamic [17] scenario, where mobiles move slowly enough that one can neglect the effect of handovers. Mobiles generate requests for file transfers, that arrive at random spatial locations, and stay at these locations for the duration of the file transfer. The arrivals are modelled by a stationary, possibly nonhomogeneous, spatial Poisson process, i.e. the number of arrivals per unit time within disjoint spatial regions are independent, and the number of arrivals per unit time for a region ΔS is Poisson distributed with parameter given by $\int_{y \in \Delta S} \lambda(y) dy$. We also let the file sizes associated with the requests be independent and generally distributed with, possibly, spatially-dependent means, denoted by $f(y)$, for $y \in D$.

Assumptions for service type at access points. Motivated by the service mechanism used in current high data rate (HDR) [16] systems, we shall postulate that APs serve queued requests in a *processor-sharing* fashion and no request is blocked from service. In our simulation models the time is divided into slots of duration 1.67 ms and within each slot a small portion of the requested file transfer for a single user is realized. To derive our algorithms we will assume that users with active requests are served in a round-robin fashion which neglects the gains from multi-user diversity, exploited in HDR protocol. This assumption, will greatly simplify our analysis and will be relaxed when evaluating the proposed algorithms via simulation.

Data rates. Each time slot, the amount of data that is served to a user depends on the instantaneous data rate, available at user's location. Note that there are two factors that limit this rate: the received signal quality and the bandwidth of the backhaul that is provisioned for the corresponding APs. Our numerical experiments have been performed under the assumption that the backhaul at the WAN layer can carry as much traffic as necessary, but the one at WLAN layer is limited[1]. The instantaneous physical data rates at each location are obtained based on the instantaneous value of the SINR via SINR-rate correspondence table used in 1XEV-DO specification [16]. To find signal strengths from various APs we use standard attenuation models which combine large scale path loss, slow (log-normal) and fast (Rayleigh) fading components.

Decision-making and system objective. The central assumption of this paper is that the requests arriving in regions covered by both a WAN and a WLAN APs could be served by either. We will denote by π the decision-making control strategy, i.e. the rule that determines to which AP an incoming request is routed. In general, the strategy π may depend on many factors that describe the system state, i.e. current AP utilizations, physical data rates at various spatial locations, current number of requests served by APs, etc. In this paper we will not attempt to study all possible decision-making strategies, but confine ourselves to designing several which would be simple to implement in practice.

The efficiency of strategy π will be evaluated based on how well it optimizes a certain system objective, given a stationary spatial load of requests for file transfers. In this paper we will set the objective U_{system}^{π} to be the total mean queueing backlog within the system:

$$U_{system}^{\pi} = \sum_{m=0}^{M-1} \mathbb{E}^{\pi}[Q_m^w] + \sum_{k=0}^{K-1} \mathbb{E}^{\pi}[Q_k^h], \qquad (1)$$

with the convention that $U_{system}^{\pi} = \infty$ if the strategy π results in unstable queueing dynamics. Note that whenever the system is stable under π, the optimization also corresponds (via Little's law) to minimizing the average delay experienced by a typical user in the system.

III. Load Balancing in Interference-Free Scenarios

Let us start by formulating load-balancing algorithms for scenarios where the other-cell interference at the WAN layer is small enough that it can be neglected. By Assumption 3 two different requests that arrive within the service zone of a WLAN AP can only be routed to the WLAN AP or the *same* WAN AP. When no other-cell interference is present at both WAN and WLAN levels, the physical data rates are only affected by fading. It follows that π is decomposable into a collection $\{\pi_m\}_0^{M-1}$ of independent decision strategies operating independently for requests originating within S_m^w, $m = 0, \ldots, M-1$ respectively. We focus on designing each π_m separately, and without loss of generality concentrate on π_0.

[1]For example, although WiFi access points can operate at physical rates of over 10Mbps, it is rarely the case that the available backhaul bandwidth exceeds 1Mbps: bachhaul capacity usually incurs high recurring costs.

Let $\mathcal{K}_m$ denote the indices of WLAN APs that fall within the service zone S_m^w. Taking into account the above discussion, the optimization reduces to finding a decision strategy π_0 that minimizes the objective:

$$U_0^{\pi_0} = \mathbb{E}^{\pi_0}[N_0^w] + \sum_{k\in\mathcal{K}_0} \mathbb{E}^{\pi_0}[N_k^h].$$

III.A Centralized adaptive load balancing

We first consider a family of strategies where incoming requests are routed to APs in a probabilistic fashion. We will assume that the decision-making entity is able to differentiate between a finite number of disjoint service classes, and associates the incoming requests with a particular class based on the average physical data rates available at the requests' locations. In other words, each request is associated with a given class which in turn quantifies the mean rate that the request can achieve to the WAN and WLAN. Our goal is to optimize system performance by selecting appropriate per-class routing probabilities.

Let us denote by $\bar{B}^w(x)$ the time average physical data rate available at $x \in S_0^w$ from WAN AP w_0. Similarly, let $\bar{B}^h(x)$ denote the time average of the physical data rate available at $x \in S_k^h$, from WLAN AP $h_k \in S_0^w$. Based on this pair of rates the request at location x is assigned to one of the disjoint classes, which we will denote via $\mathcal{B}_i$, for $i = 1,\ldots,I$. We will assume that the probability of routing a request that belongs to class $\mathcal{B}_i$ to the WAN AP w_0 is given by P_i and let $\mathbf{P} \equiv \{P_i\}_{i=1}^I$.

Using the results in [18], we may express the system objective in terms of the vector $\mathbf{P}$ as:

$$U_0(\mathbf{P}) = \frac{\rho_0^w(\mathbf{P})}{1-\rho_0^w(\mathbf{P})} + \sum_{k\in\mathcal{K}_0} \frac{\rho_k^h(\mathbf{P})}{1-\rho_k^h(\mathbf{P})}, \quad (2)$$

where the utilization $\rho_0^w(\mathbf{P})$ of the WAN AP w_0 is given by:

$$\rho_0^w(\mathbf{P}) = \int_{y\in\bar{C}_0} \frac{\gamma(y)\mathbf{1}(y\in\mathcal{B}_i)}{B^w(y)}\,dy + \sum_i^I P_i \int_{y\in C_0} \frac{\gamma(y)\mathbf{1}(y\in\mathcal{B}_i)}{B^w(y)}\,dy, \quad (3)$$

and the utilization $\rho_k^h(\mathbf{P})$ of WLAN AP $h_k \in S_0^w$ is given by:

$$\rho_k^h(\mathbf{P}) = \sum_i^I (1-P_i) \int_{y\in S_k^h} \frac{\gamma(y)\mathbf{1}(y\in\mathcal{B}_i)}{\bar{B}^h(y)}\,dy. \quad (4)$$

In the above expressions $\gamma(y) \equiv f(y)\lambda(y)$, and we denote by C_0 ($\bar{C}_0$) the set of locations in S_0^w covered by some WLAN (not covered by any WLAN), i.e. $C_0 \equiv \bigcup_{k\in\mathcal{K}_0} S_k^h$ ($\bar{C}_0 \equiv S_0^w \setminus C_0$). Our optimization problem is then given by:

Problem 1.

$$\min\{U_0(\mathbf{P}) \mid 0 \le \mathbf{P} \le 1\}$$

The following proposition establishes a convexity property of our cost function, as is the case for Jackson networks [19].

Proposition 1. *Problem 1 is convex. The gradient elements of $U_0(\mathbf{P})$ are given by:*

$$G_i \equiv \frac{\partial U_0(\mathbf{P})}{\partial P_i} = G_i^w - G_i^h, \quad (5)$$

where

$$G_i^w = \frac{\tau_i^w}{(1-\rho_0^w)^2}, \qquad G_i^h = \sum_{k\in\mathcal{K}_0} \frac{\tau_i^h(k)}{(1-\rho_k^h)^2}, \quad (6)$$

and τ_i^w, $\tau_i^h(k)$ are defined as:

$$\tau_i^w = \int_{y\in C_0} \frac{\gamma(y)\mathbf{1}(y\in\mathcal{B}_i)}{B^w(y)}\,dy, \quad \tau_i^h(k) = \int_{y\in S_k^h} \frac{\gamma(y)\mathbf{1}(y\in\mathcal{B}_i)}{\bar{B}^h(y)}\,dy,$$

Since Problem 1 is convex, any version of a gradient descent method could be used to obtain globally optimal per-class routing probabilities $\mathbf{P}$. For example we can adapt the routing probabilities in a "greedy" fashion according to:

$$\mathbf{P}(t+1) = \mathbf{P} - \alpha\nabla U_0(t), \quad (7)$$

where $\alpha > 0$ is some appropriately chosen constant, and the gradient $\nabla U_0(t)$ is as given in Proposition 1.

We can now describe an implementation of the centralized adaptive load-balancing algorithm based on gradient estimation. In our implementation, WAN and WLAN APs are able to obtain initial estimates for the average physical data rates at locations of all incoming requests. Such estimates could in practice be obtained for each dual-mode device during initial session set-up and then maintained throughout a session's life-time. The incoming requests are first routed to the central controllers, residing at each WAN AP, and the controllers forward the request to either WAN or WLAN AP according to the current value of the routing probabilities. The controller at, e.g. WAN AP w_0 is able to communicate (via a wired connection) to all WLAN APs in its service region S_0^w and thus maintains a database that includes smoothed estimates for utilizations of the WAN and WLAN APs within S_0^w (equivalently, fraction of time each AP is busy), estimates for the average data rates (at both WAN and WLAN layers) for each active mobile in the system, and current values of the routing probabilities.

Whenever a request is forwarded to the controller, the information on the average data rates is used to associate request with a particular class of service. The size of file F and the corresponding estimates $\hat{B}^w$ and $\hat{B}^h$ of average physical data rates at WAN AP w_0 and WLAN AP $h_k \in S_0^w$, associated with a request of class $\mathcal{B}_i$, are then used to update the values of τ_i^w and $\tau_i^h(k)$ – we simply keep a finite history of the entries given by $F/\hat{B}^w$ and $F/\hat{B}^h$ and compute a running average of these entries. Given estimates for utilization of the WAN AP ρ_0^w, utilization of WLAN APs ρ_k^h, $k \in \mathcal{K}_0$, estimates of τ_i^w, $\tau_i^h(k)$, for $i = 1,\ldots,I$, $k \in \mathcal{K}_0$, the controller at w_0 updates the estimate of the gradient via (5,6), and the values of per-class routing probabilities, via (7).

III.B Distributed heuristic for load balancing

Implementation of the centralized controller requires tight coordination between the WAN and WLANs. In practice it is desirable not to require such coordination: WAN and WLAN networks could be operated by different providers and may not even be aware of each other's existence. To secure good performance for such networks, the intelligence has to be moved to the dual-mode devices themselves and implemented via software "agents", which select suitable access points with minimal feedback from WAN/WLAN APs. In this subsection we provide a particular design for such agents along with the design of the APs' feedback.

Unfortunately, the computation of gradients given via Proposition 1 is not amenable to distributed implementation. To overcome this problem, we reformulate our optimization problem by switching to more convenient variables. We partition S_0^w into a large number L of disjoint sets which we denote $\Delta S_1, \Delta S_2, \ldots, \Delta S_L$, containing a representative location $y_1, y_2, \ldots, y_L$ respectively. Note that when L is sufficiently large, one can approximate, for $y \in \Delta S_i$: $\gamma(y) = \gamma(y_i)$, $\bar{B}^w(y) = \bar{B}^w(y_i)$, and $\bar{B}^h(y) = \bar{B}^h(y_i)$.

Let us assume that requests emerging at location $y \in S_i$ are routed to the WAN with probability p_i and denote by $\mathbf{p}$ the vector $\{p_i\}_{i=1}^L$. Then we obtain the following representation of the system objective

as a function of the vector $\mathbf{p}$:

$$U_0(\mathbf{p}) = \frac{\rho_0^w(\mathbf{p})}{1-\rho_0^w(\mathbf{p})} + \sum_{k\in\mathcal{K}_0} \frac{\rho_k^h(\mathbf{p})}{1-\rho_k^h(\mathbf{p})},$$

where

$$\rho_0^w(\mathbf{p}) = \sum_{i=1}^{L} \frac{\gamma_i|\Delta S_i|p_i}{B^w(y_i)}$$

$$\rho_k^h(\mathbf{p}) = \sum_{i=1}^{L} \mathbf{1}(y_i \in S_k^h) \frac{\gamma_i|\Delta S_i|(1-p_i)}{B^h(y_i)},$$

and $|\Delta S_i|$ denotes the area of ΔS_i. The corresponding optimization problem is similar Problem 1, but with $U_0(\mathbf{p})$ in place of $U_0(\mathbf{P})$:

Problem 2.

$$\min\{U_0(\mathbf{p}) |\ 0 \le \mathbf{p} \le 1\}.$$

We also can state an analogue of Proposition 1 as follows:

Proposition 2. *Problem 2 is convex. The gradient elements of $U_0(\mathbf{p})$ are given as:*

$$\frac{\partial U_0(\mathbf{p})}{\partial p_i} = \gamma_i|\Delta S_i|\left(G_i^w - G_i^h\right), \tag{8}$$

where

$$G_i^w = \frac{1}{B^w(y_i)(1-\rho_0^w)^2},$$

$$G_i^h = \sum_{k\in\mathcal{K}_0} \frac{\mathbf{1}(y_i \in S_k^h)}{B^h(y_i)(1-\rho_k^h)^2}. \tag{9}$$

Based on Proposition 2 we suggest the following decision-making strategy. As before, prior to generating requests, each dual-mode device has to establish a connection with the APs it can access. During initial session setup and within the session lifetime, the devices maintain information about the average physical data rates that are available for communication with nearby APs. Prior to requesting a file download from one of the APs, a device at location $y_i \in S_0^w \cap S_k^h$ asks both WAN AP w_0 and WLAN AP h_k to signal their current measured utilizations ρ_0^w and ρ_k^h. The device then computes an estimate $\hat{G}_i^w$ and $\hat{G}_i^h$ via (9) using the estimates for the respective utilizations and average physical data rates. It forwards the download request to the WAN AP w_0 if $\hat{G}_i^w - \hat{G}_i^h < 0$ and addresses the request to WLAN AP h_k otherwise.

The proposed approach is based on the observation that the routing probability p_i has to be decreased if the derivative $\frac{\partial U_0(\mathbf{p})}{\partial p_i}$ is positive. By imposing "hard" decisions on the agents we diminish the per unit time average of the number of requests incoming to the WAN AP from the set of locations around $\Delta S(y_i)$. Note that, strictly speaking, decisions that agents make are no longer probabilistic and thus, the analysis based on the assumption that arrivals to APs follow Poisson processes no longer holds.

IV. Interference-aware load balancing

In this section we return to a more general case, in which the WAN service rate at each location depends on the current set of active WAN APs. The queueing dynamics can thus be represented by that of multi-class processor sharing queues, where the service rates at each queue vary over time as governed by the activity state of other queues. Rigorous analysis of such queueing systems is a hard task and the analysis is barely tractable even for two single class queues [20, 21]. Thus, we will adopt the approach introduced in [22] that relies heavily on approximations.

The following is a list of additional simplifying assumptions that we make in order to derive interference-aware load balancing algorithms. We will use notation $\mathcal{A}^\pi(t)$ to refer to the stochastic process which captures the set of WAN APs active at time t under control strategy π.

Assumption 4. *We consider a set of decision-making strategies Π such that for any $\pi \in \Pi$, the stochastic process $\mathcal{A}^\pi(t)$ is stationary and ergodic, and has same marginal distribution as $\mathcal{A}^\pi$.*

Assumption 5. *The system operates in a quasi-stationary regime, i.e. the queueing dynamics within the service zone of each WAN AP are much faster than the changes in the set of active WAN APs.*

Assumption 6. *The service rates at WAN AP w_m are affected only by the activity pattern of the set $\mathcal{N}_m$ of WAN APs that are immediate neighbors of w_m, i.e. the ones that share service zone boundaries with w_m.*

In what follows with some abuse of notation, we will use $\mathcal{A}_m^\pi(t) = \mathcal{A}^\pi \cap \mathcal{N}_m$ to refer to the set of WAN APs that are immediate neighbors to w_m and, for fixed control π are active at time t (correspondingly, we will also use $\mathcal{A}_m^\pi$ to refer to the random set that has the same stationary distribution as $\mathcal{A}_m^\pi(t)$).

Using Assumptions 4-6 we can readily obtain the mean number of transfers in progress at WAN AP w_m

$$\mathbb{E}^\pi[N_m^w] = \mathbb{E}_{\mathcal{A}_m^\pi}\left[\frac{\rho_m^w(\pi,\mathcal{A}_m^\pi)}{1-\rho_m^w(\pi,\mathcal{A}_m^\pi)}\right],$$

where $\rho_m^w(\pi,\mathcal{A}_m^\pi)$ is the utilization of WAN AP w_m under decision-making strategy π when the set of active neighboring WAN APs is given by $\mathcal{A}_m^\pi$ and we used $\mathbb{E}_{\mathcal{A}_m^\pi}[\cdot]$ to denote the expectation with respect to the distribution of $\mathcal{A}_m^\pi$. Using the notation introduced in Section B, a policy π corresponds to routing vector $\mathbf{p}$ and we can express $\rho_0^w(\mathbf{p},\mathcal{F})$ as:

$$\rho_0^w(\mathbf{p},\mathcal{F}) = \sum_{i=1}^{L} \frac{\gamma_i|\Delta S_i|p_i}{B_{\mathcal{F}}^w(y_i)}$$

where $\bar{B}_{\mathcal{F}}^w(y)$ denotes the average WAN physical data rate available at location y when the set of active WAN APs is $\mathcal{F}$.

Assumption 7. *Decision-making policies π_m are adjusted independently for different $m = 0,\dots,M-1$ so as to minimize "partial" system objectives $\widetilde{U}_m(\pi)$:*

$$\widetilde{U}_m(\pi) = \mathbb{E}^\pi[N_m^w] + \sum_{k\in\mathcal{K}_m} \mathbb{E}^\pi[N_k^h] + C_{IF}^w \sum_{n\in\mathcal{N}_m} \mathbb{E}^\pi[N_n^w],$$

where $C_{IF}^w > 0$.

The last preliminary step that we take is making assumptions that allow us to describe the actual changes that occur within the system due to small variations of a single component of vector π, say π_0. Note that for $m \in \mathcal{N}_0$ we can express $\mathbb{E}^\pi[N_m^w]$ as:

$$\begin{aligned}\mathbb{E}^\pi[N_m^w] \quad &= \Big(A_0^\pi \mathbb{E}_{\mathcal{A}_m^\pi}[N_m^w \mid w_0 \text{ is busy}] \\ &\quad + (1-A_0^\pi)\mathbb{E}_{\mathcal{A}_m^\pi}[N_m^w \mid w_0 \text{ is silent}]\Big),\end{aligned} \tag{10}$$

where A_0^π is the probability of WAN AP w_0 to be active under strategy π. We will make another key assumption, under which we will consider the effect on $\mathbb{E}^\pi[N_m^w]$ associated with the change of distribution

$\mathcal{A}_m^\pi$ on $\mathbb{E}^\pi[N_m^w]$, occurring due to a change in π_0, to be negligible, in comparison to the effect on the same quantity of the change in activity probabilities $A_0(\pi)$ induced by the change in π_0. Expressing this mathematically we have:

Assumption 8.

$$\begin{aligned}\delta_{\pi_0}\mathbb{E}^\pi[N_m^w] \quad &= \left(\mathbb{E}_{\mathcal{A}_m^\pi}[N_m^w \mid w_0 \textit{ is busy}]\right.\\ &\quad \left.-\mathbb{E}_{\mathcal{A}_m^\pi}[N_m^w \mid w_0 \textit{ is silent}]\right)\delta_{\pi_0}A_0^\pi\\ &\quad +o(\delta_{\pi_0}A_0^\pi),\end{aligned}$$

where $\delta_{\pi_0}F$ is a first variation of F due to the variation in π_0. In addition, assume that

$$\delta_{\pi_m}A_m^\pi \equiv \delta_{\pi_m}\left(\mathbb{E}_{\mathcal{A}_m^\pi}[\rho_m^w(\pi,\mathcal{A}_m^\pi)]\right) = \mathbb{E}_{\mathcal{A}_m^\pi}[\delta_{\pi_m}\rho_m^w(\pi,\mathcal{A}_m^\pi)]$$

Correction term in distributed estimation of the gradient. We will derive the correction to load-balancing algorithm for the case of distributed implementation (it can be done similarly for the centralized version of the algorithm). Under Assumptions 4-8 we can concentrate on decision-making within a particular service zone, and we choose to focus on S_0^w, as before.

Using Assumption 8 the convexity of the function $\widetilde{U}_0(\mathbf{p})$ with respect to the vector $\mathbf{p}$ could be established, however due to our other assumptions it might be expected to hold only "approximately" in reality. The elements of the gradient of the partial objective can now be expressed as:

$$\widetilde{G}_i \equiv \frac{\partial \widetilde{U}_0}{\partial p_i} = \gamma(y_i)\Delta S_i\left(\widetilde{G}_i^w - G_i^h\right),$$

where G_i^h is the same as in (9), and $\widetilde{G}_i^w$ includes a correction term, which involves "cost of interference" caused to neighboring cells:

$$\begin{aligned}\widetilde{G}_i^w &= \mathbb{E}_{\mathcal{A}_0}\left[\frac{1}{\overline{B}_{\mathcal{A}_0}^w(y_i)(1-\rho_0^w(\mathbf{p},\mathcal{A}_0))^2}\right]\\ &+C_{IF}\mathbb{E}_{\mathcal{A}_0}\left[\frac{1}{\overline{B}_{\mathcal{A}_0}^w(y_i)}\right]\sum_{m\in\mathcal{N}_0}\left(\mathbb{E}_{\mathcal{A}_m}[N_m^w \mid w_0 \text{ is busy}]\right.\\ &\quad \left.-\mathbb{E}_{\mathcal{A}_m}[N_m^w \mid w_0 \text{ is silent}]\right)\end{aligned}$$

The interference-aware decision-making algorithms are the same as we described in Section B, except that an agent at $y \in \Delta S_i$ selects to join the WAN or WLAN AP via comparing $\widetilde{G}_i^w$ with G_i^h. To enable the computation of $\widetilde{G}_i^w$ the WAN access points have to be involved in a more complex coordination between their neighboring APs: the WAN APs neighboring to w_0 have to maintain the estimates of the averages $\mathbb{E}_{\mathcal{A}_m}[N_m^w \mid w_0 \text{ is busy}]$ and $\mathbb{E}_{\mathcal{A}_m}[N_m^w \mid w_0 \text{ is silent}]$ and signal them to w_0 over some dedicated wireline or wireless channel. The estimation of expectations $\mathbb{E}_{\mathcal{A}_m}[\cdot]$ can be done in a straightforward way by maintaining a finite history of measurements for respective quantities.

V. Simulations

For most of our simulation experiments we found that distributed decision-making strategy performs either as well or even better than the centralized one. Figure 3 we illustrates this by comparing the performance of centralized and distributed load-balancing strategies for a scenario with homogeneous Poisson load and geometric setup shown on Figure 2. The mean delay seen by a typical request is plotted vs the backhaul bandwidth, available at each WLAN. To provide

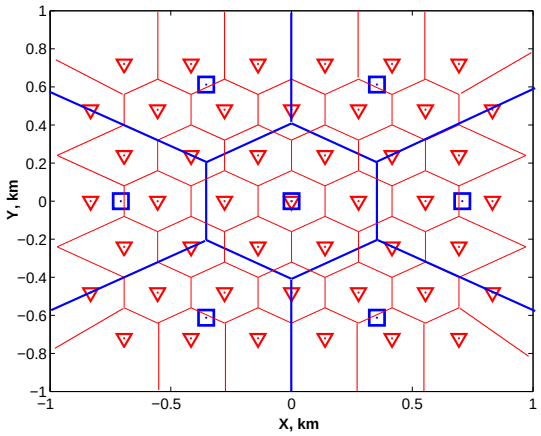

Fig. 2: Geometric setup: WAN APs shown as boxes and WLAN APs are shown as triangles. The power levels at WLANs are sufficient to cover their service zones represented by corresponding Voronoi cells.

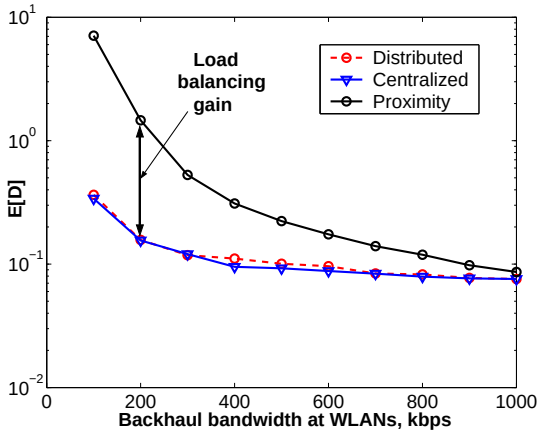

Fig. 3: Performance of interference-unaware load balancing algorithms under symmetric loads, compared to performance of proximity-based routing strategy.

a benchmark, we also use a simple and more natural (at least in current applications) proximity-based routing strategy, under which all requests that emerge within a service zone of a WLAN are simply routed to the corresponding WLAN AP.

The good performance of distributed algorithm can in part be attributed to the fact that it does not require estimation of current traffic demands at various locations, or on a per-class basis as has to be done for centralized implementation when estimating the quantities τ_i^w and $\tau_i^w(k)$. The imprecision of these quantities contributes to errors in estimating of gradient of the objective.

Our second experiment models an asymmetric traffic scenario, in which the aggregate load within S_0^w, (service zone of WAN AP at the center of Figure 2) greatly exceeds the load for the adjacent WAN service zones. To create even more load asymmetry, for this experiment we assume that WLANs falling within the WAN service zone S_0^w (at the center of Figure 2) are *shut off*, in which case the WAN AP w_0 has to serve all requests emerging within S_0^w. The power levels used at WAN APs are large enough to ensure "interference-limited" regime of operation.

Figure 4 shows simulation results for this scenario where we vary the file arrival rate within the service zone of the WAN AP in the center, keeping the load on adjacent WAN AP service zones fixed. The

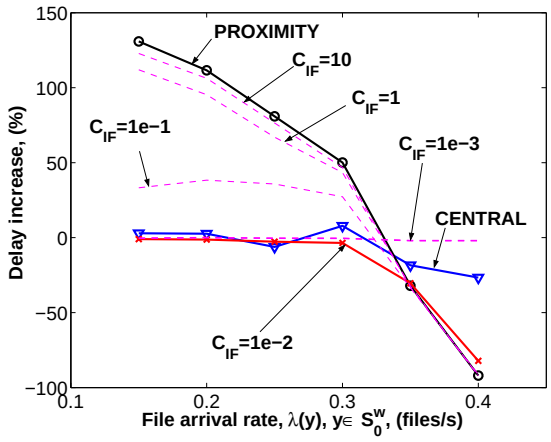

Fig. 4: Mean delay increase over interference-unaware distributed decision-making for other decision-making strategies, under asymmetric loads.

results are interesting in that they demonstrate how the proximity-based strategy, which is probably the worst possible decision-making strategy under low utilizations, becomes desirable with increased load at the central WAN cell. The key to understanding the seemingly unexpected outcome of the experiment is to note that the other-cell interference affects mostly the WAN AP at the center, especially when it approaches its service capacity when load on it increases. The interference unaware algorithms do not however, stimulate adjacent service zones to route more requests to WLANs in order to reduce utilizations of WAN APs and to reduce interference on the WAN service zone at the center.

The "partial" objectives used in our formulation of the interference-aware algorithms use the current queue lengths at all neighboring WAN APs to signal their congestion. Including these signals is equivalent to inducing a bias on the system that forces more requests to be routed towards WLANs when the whole system can benefit from it. Note that the value of this bias can be controlled via tuning the constant C_{IF}. We illustrate this tuning in Figure 4 by showing performance for different C_{IF}.

Note that with properly "tuned" interference-aware decision-making it is possible to achieve both the gains of distributed load-balancing under light loads and proximity-based routing, for heavy asymmetric loads. Good performance of interference-aware strategies has been verified also in other experiments which we do not describe due to space considerations. The gains from employing interference-aware policies, however, depend on environmental propagation characteristics, degree of load asymmetry and particular capacities available at various access points in the system.

VI. Conclusion

In this paper we presented some new results on achieving load-balancing among heterogeneous wireless systems that include a combination of WAN and a set of WLANs. Under the assumption that the access points serve the incoming download requests in a processor-sharing fashion, we formulate and evaluate centralized and distributed decision-making routing schemes that enable significant performance gains. The major contribution of this paper is explicit incorporation of other-cell interference as a factor affecting load-balancing decisions and exhibiting scenarios in which interference-aware algorithms achieve significant performance gains.

References

[1] U. Varshney and R. Jain, "Issues in emerging 4g wireless networks," *Computer*, pp. 94–96, June 2001.

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

[3] M. Benveniste, "Cell selection in two-tier microcellular/macrocellular systems," in *Proc. IEEE GLOBECOM*, 1995, pp. 1532–1536.

[4] K. L. Yeung and S. Nanda, "Channel management in micro/macrocellular radio systems," *IEEE Transactions on Vehicular Technology*, vol. 45, pp. 601–612, November 1996.

[5] W. Jolley and R. Warfield, "Modeling and analysis of layered cellular mobile networks," in *Teletraffic Datatraffic in a Period Change*, 1991, vol. ITC-13, pp. 161–166.

[6] X. Lagrange, "Multitier cell design," *IEEE Communications Magazine*, August 1997.

[7] M. Buddhikot, G. Chandranmenon, S. Han, Y. W. Lee, S. Miller, and L. Salgarelli, "Integration of 802.11 and third-generation wireless data networks," in *Proc. IEEE INFOCOM*, 2003.

[8] R. G. Sheng and L.-S. Tsao, "3G-based access control for 3GPP-WLAN internetworking," in *Proc. of the IEEE Vehicular Technology Conference*, May 2004.

[9] A. K. Salkintzis, C. Fors, and R. Pazhyannur, "WLAN-GPRS integration for next-generation mobile data networks," *IEEE Wireless Comm.*, vol. 9, pp. 112–124, October 2002.

[10] M. Jaseemuddin, "An architecture for integrating umts and 802.11 wlan networks," July 2003, vol. 2, pp. 716–723.

[11] N. Vulic and S. H. de Groot, "Architectural options for WLAN integration into the UMTS radio access level," in *Proc. of the IEEE Vehicular Technology Conference*, May 2004.

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

[13] S. Lincke-Salecker and C. S. Hood, "Integrated networks that overflow speech and data between component networks," *Int. J. Network Mgmt.*, vol. 12, pp. 235–257, 2002.

[14] A. Zemlianov and G. de Veciana, "Cooperation and decision-making in a wireless multi-provider setting," in *Proc. of IEEE INFOCOM*, March 2005, to appear.

[15] T. E. Klein and S.-J. Han, "Assignment strategies for mobile data users in hierarchical overlay networks: Performance of optimal and adaptive strategies," *IEEE Journal on Selected Areas in Communication*, vol. 22, no. 5, month =.

[16] 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.

[17] B. Blaszczyszyn, M. K. Karray, and F. Baccelli, "Blocking rates in large cdma networks via a spatial erlang formula," in *Proc. of IEEE INFOCOM*, March 2005, to appear.

[18] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *Proc. of IEEE INFOCOM*, 2003.

[19] D. P. Bertsekas and R. Gallager, *Data Networks*, Prentice Hall, 1991.

[20] J. W. Cohen and O. J. Boxima, *Boundary Value Problems in Queueing Systems Analysis*, North-Holland Publ. Cy., Amsterdam, 1983.

[21] G. Fayolle and R. Iasnogorodski, "Two coupled processors: the reduction to a riemann-hilbert problem," *Z. WAhr. verw. Geb*, vol. 47, pp. 325–352.

[22] T. Bonald, S. Borst, N. Hegde, and A. Proutière, "Wireless data performance in multicell scenarios," in *Proc. SIGMETRICS/Performance*, 2004.