Online Load Balancing and First-Hop Bandwidth Allocation in Public-Area Wireless Networks

Abstract— **Recent studies characterizing workloads in Public-Area Wireless Networks (PAWNs) have shown that: (i) user loads are often time varying and location-dependent; (ii) user load is often unevenly distributed across access points (APs); and (iii) the load on the APs at any given time is not well correlated with the number of users associated with those APs. Administrators in such networks thus have to address the challenge of unbalanced network utilization resulting from unbalanced user load, and also guarantee its users a minimum level of quality of service (e.g., sufficient wireless bandwidth).**

In this paper, we address the challenges of improving PAWN utilization and user bandwidth allocation through a common solution – dynamic, location-aware adaptation. We observe that by adaptively varying the bandwidth allocated to users in the wireless hop within certain bounds coupled with admission control at each AP, the network can accommodate more users as its capacity changes with time. Further, by adaptively selecting the AP that users associate with, the network can relieve sporadic user congestion at popular locations and increase the likelihood of admitting users at pre-negotiated service levels.

We describe how these algorithms enable the network to transparently adapt to user demands and balance load across its access points. We evaluate the effectiveness of these algorithms on improving user service rates and network utilization using simulations incorporating real workloads from campus, conference, and corporate environments. Our algorithms improve the degree of balance in the system by over 45% and allocate over 30% more bandwidth to users in comparison to existing schemes that offer little or no load balancing.

I. Introduction

The vision of pervasive ubiquitous computing where users have network access anytime, anywhere, is being enabled by deployments of high-speed wireless networks in common places of congregation such as airports, malls, hotels, parks, arenas, and so on [1], [2]. Two key challenges to the host organization deploying these Public-Area Wireless Networks (PAWNs) are: (i) capacity planning, making the best use of the available network resources to derive the best return on its investment; and (ii) guaranteeing at least a minimum amount of bandwidth to users. As the use of PAWNs spreads beyond simple data transfer to data- and performance intensive multimedia applications, the need to address quality of service issues, such as enhanced service provisioning and first-hop bandwidth management will become increasingly important. Further, as PAWNs scale to larger organizations and support a greater number of users, it will be crucial to consider techniques that adequately provide capacity to handle dynamically varying, location-dependent user load.

We envision that in the future PAWNs will support a wide range of quality of service (QoS) models to provide sustained levels of the wireless bandwidth to contending users. These models would range from free access without guarantees (for best-effort traffic) to paid connectivity for applications requiring fixed QoS (e.g., IP telephony) to adaptive differentiated service for real-time multimedia applications (e.g., streaming, audio/video conferencing). Such service policies provide a natural separation between different classes of users, allowing the creation of a tiered service model that benefits paid users. Therefore, as PAWNs allow users to *shop* for a desired level of QoS, it is important that the network have adequate capacity and that the first-hop bandwidth be managed scalably and efficiently.

Most PAWN deployments have hitherto addressed the problem of capacity planning through static overprovisioning of network capacity – installing enough wireless access points (APs) to handle an overall estimated network load. Unfortunately, there are limitations to this approach. First, installation and operation of more access points translates to a larger infrastructure and maintenance cost. Second, an increased number of APs in the network would limit the number of APs that can be operated on non-interfering channels due to the inherent limits of channel reuse in 802.11 networks.

Recent workload characterization studies of PAWNs [3], [4], [5], [6] have shown that user service demands are highly dynamic in terms of both time of day and location, and that user load is often distributed quite unevenly among the APs. Furthermore, it has been shown that the load on the APs at any given time is not well correlated with the number of users associated with those APs. A key consequence of this behavior is sporadic user congestion at certain popular spaces within the network resulting in (i) under-utilized network resources due to unbalanced load, and (ii) unsatisfied user service requests.

In this paper, we address the challenges of capacity planning and user bandwidth allocation through a common solution – dynamic, location-aware adaptation. In order to balance the network load, we propose that the network adaptively select the AP that the user associates with by incorporating the user's workload and geographic location within the network. In order to satisfy the user's service request, we propose that the initial process of association with an AP be performed in conjunction with explicit admission control at each candidate AP that can admit the user's request. Therefore, both the network and its users explicitly and cooperatively adapt themselves to changing load conditions. By admitting the user's traffic at an AP other than the one that would provide the service by default (as in association based on highest signal strength), the network load automatically gets distributed across its APs. As users' traffic is dynamically and adaptively directed from a heavily loaded AP to a lightly loaded AP, it increases their likelihood of receiving a pre-negotiated

bandwidth guarantee from the network.

This paper makes the following contributions:

1. We present the problem of first-hop wireless bandwidth allocation as a special case of the well-known online load balancing problem and present three online heuristics for first-hop bandwidth allocation. These heuristics improve the degree of balance in the system by over 45% and allocate over 30% more bandwidth to users than current approaches;

2. We prove that the general *offline* problem (i.e., where we have global knowledge of user arrivals and requests) of finding an optimal assignment of users to APs in an arbitrary network with arbitrarily sized user bandwidth requests, is NP-complete;

3. We propose three different heuristics for allocating users to APs based on their bandwidth requirements and evaluate their performance via trace driven simulations. Our simulations model three different PAWN settings using real workload characterization traces: (i) a conference WLAN workload [3], (ii) a university campus WLAN workload [4], and (iii) a corporate WLAN workload [5]. To the best of our knowledge, ours is the first study of wireless LAN bandwidth provisioning incorporating real WLAN workloads in the simulation.

The rest of this paper is organized as follows. In Section 2, we overview related work in QoS provisioning in wireless LANs. In Section 3, we introduce the problem of online bandwidth allocation using mathematical notation. In Section 4, we derive the NP-completeness result of the optimal offline resource allocation problem. In Section 5, we describe three heuristics for online bandwidth allocation. In Section 6, we discuss mechanisms for online load balancing in PAWNs. In Section 7, we evaluate the online bandwidth allocation heuristics and finally conclude in Section 8.

II. Related Work

The IEEE 802.11 standard does not provide any specifications for capacity planning. Further, the 802.11 CSMA/CA protocol with the *Distributed Coordination Function* (DCF) for media access itself does not provide any guarantees on the wireless bandwidth [1].

The 802.11 Working Group is still considering proposals for introducing QoS enhancements into the standard. One of these proposals calls for the use of per-flow resource-based admission control combined with prioritized data transmission for real-time traffic [7]. However, this scheme does not take into account the dynamically varying nature of the wireless medium. In [8], the authors discuss various bandwidth allocation techniques for managing bandwidth in the wireless hop. While this work addresses first-hop QoS by taking into account the end-to-end path properties of individual flows, it does not deal with reallocation of flows among APs.

There have been a number of other proposals to enhance or modify the MAC protocol in wireless LANs to provide long-term fairness to flows using centralized and distributed schemes [9], [10], [11], [12]. Again, all of these schemes have focused on enhancing the fairness properties of the wireless MAC in order to provide differentiation among contending flows, thus improving user QoS within a single cell in the network. They do not focus on the dynamics of the wireless network as a whole.

Recently, various vendors of wireless LAN products have incorporated load-balancing features in the latest release of network drivers and firmware for APs and wireless PC cards [13], [14]. APs supporting this feature maintain a count of the number of users associated with APs in each cell and broadcast beacons containing this information to users in the cell. New users receive beacons from multiple access points and use this load information to determine and associate with the least-loaded AP. However, these techniques do not take into account user workloads and QoS requirements and are local in scope, distributing users evenly across available overlapping cells.

In [15], the authors present load-balancing algorithms for efficient routing in multi-hop wireless access networks. However, their algorithms pertain to multi-hop wireless access networks where each node has to find a QoS-aware route to the egress node that connects to the backbone of the network. In contrast, we focus on networks where every mobile node is only one wireless hop away from the backbone, and hence wireless routing is not an issue. Further, they do not consider how network load changes with arriving and departing users; this cannot be neglected in PAWNs.

Hanly [16] has addressed the problem of maximizing spread spectrum capacity in a cellular network by finding an optimal allocation of users to base stations and an optimal set of transmitter power levels. Although it may appear that such approaches are also applicable to wireless LANs, several important differences exist. First, wireless LANs use distributed, contention-based MAC protocols where only one user accesses the channel at any given time. Second, in wireless LANs cell capacity is related to the individual workloads of users rather than the transmit power levels [3], [5].

Lastly, Azar [17] and Phillips [18] have extensively studied the complexity of the network load balancing problem from a theoretical standpoint, and have proven bounds for several heuristics for the online problem. In this paper, we adapt their theoretical model to wireless LANs and explicitly evaluate the performance of our heuristics using real workload measurements.

The algorithms presented in this paper jointly address the problems of increasing wireless network utilization and maintaining pre-negotiated user bandwidth agreements with the network.

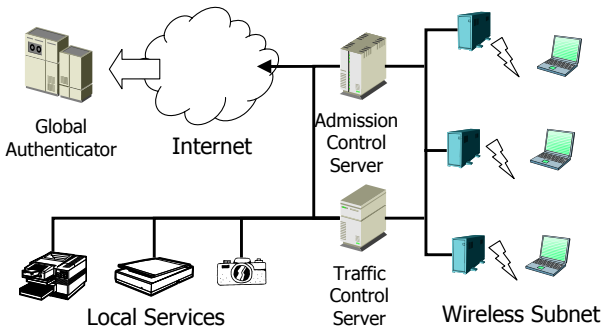


Fig. 1. The architecture of a PAWN

III. Online Bandwidth Allocation – Theoretical Formulation

In this section, we present a theoretical formulation of the problem of first-hop bandwidth allocation and show that it maps to a special case of the general *online load balancing* problem. We begin by describing our system and quality of service models, and then formally present the problem definition.

A. System Model

We consider a PAWN model with a high-speed wired backbone and first-hop wireless links configured according to the IEEE 802.11 standard in infrastructure mode [1]. The high-level architecture of our PAWN model is shown in Figure 1. A PAWN is serviced by APs providing overlapping coverage in the geographic area. Neighboring APs operate on different radio frequency (RF) channels to avoid interference. Our design assumes the existence of an *admission control server* (ACS) that maintains and manages all per-AP and per-user state in the network. The ACS makes repeated admission control decisions as users arrive and move within the network and thus helps determine their point of attachment to the network. The goal of the network is to alleviate user congestion at the hot-spots and thus improve network utilization. We now describe the quality of service model that specifies how users negotiate their QoS requirements with the network.

B. Quality of Service Model

Since the first-hop bandwidth in a wireless network is a scarce, shared resource, an equitable distribution of the available bandwidth among contending users necessitates QoS negotiation. While wired networks provide users with fixed levels of deterministic or statistical guarantees, through bandwidth reservation, many aspects of wireless networks preclude exact control over the first-hop bandwidth. First, wireless networks are characterized by time-varying and location dependent errors in the channel [12]. Second, users in a wireless network tend to be mobile and the QoS that has been negotiated at one location may not be honored as the users change their point of attachment to the network. Therefore, the network will have to adaptively vary the level of QoS provided to the user as the channel quality and capacity change with time due to the dynamics of the wireless environment.

We use the notion of bandwidth bounds introduced in [19] to characterize user QoS specification. Each user's rate requirement is specified by a $[b_{min}, b_{max}]$ bound. Once a user is admitted at an AP, the network attempts to guarantee the user a data rate of at least b_{min} with possible provisioning up to b_{max}. Typically, the lower bound b_{min} is determined by minimum rate requirement for the user's application, while the upper bound b_{max} is determined by its peak rate. If the user does not specify bandwidth bounds, the network assumes a best-effort request. Each AP in the network has a certain fraction of capacity reserved for best-effort users to allow for backward compatibility with existing schemes.

The notion of bandwidth bounds for QoS negotiation in the first-hop has several advantages. First, it offers the PAWN provider a way to adaptively plan its capacity and achieve load balancing. Second, it allows the user a way to negotiate a pipe to the backbone, with a guaranteed minimum bandwidth and excess capacity provisioning beyond b_{min}, as available. Third, QoS bounds can be used to characterize user workloads for both real-time multimedia and bursty data traffic. Finally, as mentioned in the introduction, supporting different levels of b_{min} provides a natural separation between user connections, allowing the creation of a tiered service model that benefits paid users. These QoS policies would either be advertised to the user by the PAWN provider. Alternatively, the QoS policy could be driven by some pre-negotiated policy between the users and wireless ISPs at the PAWN host organization. For example, Wayport has entered strategic relationships with other PAWN providers such as iPass, such that, users accessing these two networks (e.g., at a hotel and an airport), would receive a pre-negotiated level of service at a pre-determined charge [20].

C. Notation

We now characterize the *online* user allocation problem mathematically. We consider a PAWN serviced by N APs that are supposed to serve a set of users that arrive and depart in time. Each AP has a fixed capacity B Mb/sec. Each user j has an associated bandwidth requirement, given by a range $(b_{min,j}, b_{max,j})$, an arrival time $\tau(j)$, and a set $D_j \subseteq M$ of APs that are within RF range of the user's location.

We first consider the case of user allocation without preemption. Therefore, a user is to be assigned to exactly one of the APs in D_j upon arrival and once assigned, cannot be transferred to a different AP. The assigned AP starts processing the user's request immediately at a rate $b_{min,j} < b_j < b_{max,j}$, until the user departs the system. If no AP in D_j can admit the user's request, the user waits in a queue until capacity becomes available. The total load on AP i at time t, denoted $L_i(t)$ is the sum of the bandwidths b_j of users assigned to AP i at time t. An *online* assignment algorithm must assign a user j to a server in D_j so as to reduce the maximum load on any given AP and satisfy the user's bandwidth request subject to the capacity

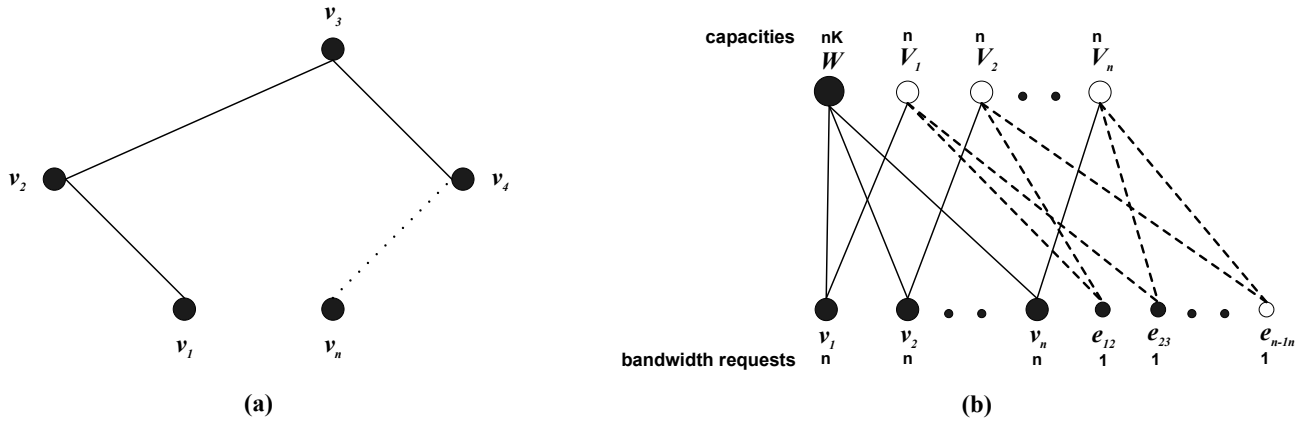


Fig. 2. The reduction of an instance of (a) Vertex Cover to an instance of (b) LCBA

constraint – the decision is made without any knowledge of future arrivals or departures [17]. For unrestricted assignment, (i.e., $|D(j) = N$) Azar et al. [17], showed that the best solution to the online problem is greedy heuristic that achieves a competitive ratio (online/optimal offline) of $log(N)$.

Before proposing algorithms to solve the *online* problem, we evaluate the complexity of the related *offline* bandwidth allocation problem, which is defined as follows. The optimal *offline* assignment algorithm assigns all arriving users knowing the entire sequence of user arrivals and departures. For a set of K total users, labeled 1, 2, ..., K, the algorithm computes the optimal allocation vector,

$c \in \{1, 2, \ldots, N\}^K : c_j \in D_j, \forall j = 1, 2, \ldots, N$

of users to APs that minimizes the maximum load on any AP and satisfies the users' bandwidth request without violating the capacity constraint of each AP. We call this *offline* problem *Load-Balanced Capacity-Constrained Bandwidth Allocation* (LCBA).

IV. NP-Completeness of LCBA

In this section, we prove that LCBA is NP-complete. We note that LCBA is an optimization (or search) problem and the theory of NP-completeness is designed to be applied only to *decision* problems [21]. However, each optimization problem can be stated as a corresponding decision problem. The proof of NP-completeness of a decision problem Π consists of two parts: (i) $\Pi \in$ NP, and (ii) for a known NP-complete problem $\Pi^{'} \in$ NP, there exists a polynomial time transformation from $\Pi^{'}$ to Π.

We assume that for each user j, ($b_{min,j} = b_{max,j} = b_j$), and $|D_j| = 2$, i.e., each user is within range of exactly 2 APs. We show that even this special instance of LCBA is NP-complete, thereby showing that the general problem is at least as hard.

A. Graph-theoretic Formulation

Given a set of K users, N APs, and the set D_j for each user j, we can construct a bipartite graph G with two sets of vertices L and R. The set L represents users ($|L| = K$) and the set R represents APs ($|R| = N$). An edge between a vertex in L and a vertex in R indicates that the user is within range of that AP. Therefore, for each user j, we have an edge connecting j and every node in D_j. The edges are labeled with capacity equivalent to the bandwidth requirement b_j of user j.

The decision version of the LCBA problem can then be described as follows:

INSTANCE: A directed graph $G = (V, E)$, where $V = L \cup R$, *and* $\forall\ j \in L$, a b_j specifying the bandwidth requirement of user j. A set of APs, D_j, within radio range of user j, where every node in D_j is in R. The edge set is $E = \{(j, D_j)\ \forall\ j,\ where\ j \in L\ and D_j \in R \forall\ j\}$. A fixed constant capacity B and a constant S.

QUESTION: Is there an allocation of users to APs such that the total load at any AP $\sum_{i \in R, j \in i} b_j \leq$ S $\leq$ B?

There are two parts to proving the NP-completeness of LCBA. First, we have to show that LCBA is in NP and next, show that $L \leq_P LCBA$ for some NP-complete language L. We show both these below.

Theorem 1: LCBA is in NP.

Proof: In order to prove membership in NP, we have to show that it is possible to verify a certificate solution to an instance of *LCBA*, in polynomial time. A certificate solution to a given instance of *LCBA* is the assignment of the each vertex in L to a single vertex in R. Formally, the verifier takes as input the graph G, the values of b_j and D_j for each vertex in L, and the assignment A. It checks that the assignment defines a map for each vertex j in L to only one vertex in D_j. Then it checks that $\sum_{i \in R, j \in i} b_j \leq$ S $\leq$ B. If these checks pass, it accepts, else it rejects. □

Theorem 2: LCBA is NP-hard, i.e., $L \leq_P LCBA$ for some NP-complete language L.

Proof: : It can be shown that the above decision problem is NP-hard through a reduction from the vertex cover problem, which is well-known to be NP-complete. Thus we choose L to be the language of graphs with vertex cover of size at most K, and show that $L \leq_P LCBA$.

The vertex-cover problem can be defined as follows:

The vertex cover of an undirected graph $G = (V, E)$ is

a subset $V^{'} \in V$ such that if $(u, v) \in E$, then $u \in V^{'}$ or $v \in V^{'}$ (or both). In other words, each vertex "covers" its incident edges, and a vertex cover for G is a set of vertices that covers all the edges in E. The size of the vertex cover is the number of vertices in it. The decision version of the vertex cover problem takes as input the graph $G = (V, E)$ and a positive integer K and is to determine whether G has a vertex cover a size at most K.

The reduction f, shown in in Figure 2, takes as input a graph $G = (V, E) : V = \{v_1, v_2, \ldots v_n\}, i.e., |V| = n$, and a positive integer K and produces a bipartite graph instance of $LCBA$ as follows. First we show the construction of the AP nodes, then the user nodes, the edges between APs and users, and then assign capacities and bandwidth requests.

1. For each vertex in V, construct an AP node V_i. We call these *vertex APs*. Then add a special $(n+1)^{th}$ AP node W.
2. For each vertex in V, construct a user node v_i. We call these *vertex users*. In addition, for each edge $e_{ij} = (v_i, v_j)$, add a user node e_{ij}. We call these *edge users*.
3. Connect each vertex user v_j to its corresponding vertex AP V_j *and* to the AP W. Connect each edge user e_{ij} to the vertex APs V_i and V_j.
4. Assign a capacity n to each vertex AP node V_j constructed from the vertices in G and assign a capacity nK to the AP node W.
5. Assign a bandwidth request of n to each vertex user v_j constructed from the vertices in G and assign a bandwidth request of 1 to each edge user e_{ij}.

It is easy to see that f is polynomial time computable. The intuition for why the reduction works is that every edge in the original graph has degree at most n, and that one or both end points of an edge are in the vertex cover.

Formally, we need to check that G has a vertex cover of size K if and only if $G^{'}$ has an admissible assignment of users to APs. So first suppose G has a vertex cover of size K that consists of nodes $\{v_i, v_j, \ldots v_{i+K-1}\}$. Assign each vertex user node in the vertex cover to the AP W, which leaves AP W filled to capacity. Now, assign the remaining vertex users (those vertices not in the vertex cover) to their corresponding vertex APs, which leaves those APs filled to capacity. Now the only user nodes that remain to be assigned are the edge user nodes, and the only APs available are those K vertex APs that form the vertex cover. It is easy to see that each of these nodes will be attached to at least one edge node (since each edge node is incident on at least one node in the vertex cover), and at most n edge nodes (since each vertex has degree at most n). Therefore, the edge nodes of unit capacity requirement can be assigned to these vertex-cover APs.

Conversely, suppose $G^{'}$ has an admissible assignment of users to APs. We let $S = nK$. Now the K users assigned to AP W will use up its capacity leaving the other vertex users (those with bandwidth request n) to be forcibly assigned to their corresponding vertex APs, again using up their capacity. This means that the edge nodes can now only be assigned to the those K vertex APs, whose correspondent vertex users were assigned to W. Since every edge user node is connected only to its end point vertices (by construction from G), this subset of K vertex APs will cover every edge user node formed from the edge set of G. In other words, these nodes form a vertex cover of size K in G. This concludes the proof. □

We note here that the reduction maps an arbitrary instance of the vertex cover problem to a defined instance of the LCBA problem, where $|D(j)| = 2$. Therefore, the problem of allocating users with arbitrarily sized requests to APs, given an arbitrary network layout is a hard problem, even for the constrained case that a user can hear only two APs.

V. Heuristic Algorithms

In this section, we describe three heuristic algorithms to solve the online LCBA problem. We focus primarily on how these heuristics manage the first-hop wireless bandwidth in the PAWN. When users request service from the network, the heuristics perform admission control at the APs and return to them the AP that they should associate with. These heuristics have previously been studied in the context of online bin packing [22].

Since the heuristics solve the online problem, they operate with no knowledge of future user requests. Their goal is to accommodate each user requests at the AP that has the capacity to service them. The criterion motivating the choice of the AP is based on: (i) achieving a balanced load distribution across APs at any instant (i.e., a greedy approach), or (ii) trading off transient load imbalance among APs in order to admit potentially larger bandwidth requests in the future. The latter approach admits requests at an already loaded AP that can still contain them, in order to reserve room at APs that may be better filled by *heavier* (i.e., higher bandwidth) future requests. In each case, the users' requests are first admitted at the lower bound, b_{min}, and any excess capacity is divided in a way that users are admitted at the level of their upper bound b_{max}, as far as possible. In other words, if $b_{avl,m}$ is the available capacity at AP m at a given instant, the excess capacity $b_{excess,j}$ above $b_{min,j}$ allocated to user j is: $b_{excess,j} = min(b_{avl,m}, b_{max,j})$. Therefore, the allocation is *fair* in the sense that all user requests assigned to a given AP get an equal excess share of the APs bottleneck capacity.

A. First-fit Allocation

The *First-Fit* heuristic allocates users to the first AP in the list that has enough available capacity. If the AP that the user associates with by default upon entering the network has enough capacity to admit his request, First-Fit retains the user at that AP, thereby performing similar to the non-load balanced approach. In general, given a set of consecutive user arrivals, First-Fit tries to admit the requests locally in the neighborhood of APs around the user. Therefore, First-Fit preferentially *fills-up* certain APs before others and gradually spreads user load from the neighborhood of the congested region through the entire network.

B. Best-Fit Allocation

The *Best-Fit* heuristic looks for the *best* AP that can still contain the user's request. The best AP refers to the most filled AP that still has enough capacity to admit the request under consideration. Intuitively, it can be seen that Best-Fit would perform worse than First-Fit in balancing the user load. However, the advantage of Best-Fit is that it minimizes overall unused capacity (i.e., wasted bandwidth) by *tightly* packing a certain heavily-loaded AP and reserving capacity at a comparatively lightly-loaded AP for heavier requests.

C. Balanced-Fit Allocation

The *Balanced-Fit* heuristic is a more intuitive approach to allocating users to APs. For a given user request, Balanced-Fit admits it at the AP that has the maximum available capacity or least load. Ties are broken arbitrarily. It is easy to see that at every step Balanced-Fit globally distributes the load through the entire network. However, Balanced-Fit can have poor worst-case performance because at any instant it creates a fragmentation of the network load among the available APs. In other words, on average every AP in the network is equally likely to admit users from a given set of incoming requests, thereby increasing the probability of denying service to a future heavy request. In contrast, both First-Fit and Best-Fit have better worst case performance in being able to admit more users at their admissible bandwidth levels. The advantage of Balanced-Fit lies in its efficient use of available resources to maximize instantaneous network utilization. Therefore, it always has better average-case performance.

D. Discussion

The heuristics described in this section operate on the assumption that users more or less stay localized in a certain region of the network, which is true the case as reported in the PAWN workload characterization studies involving laptop users [3], [4], [5], [6]. However, if users are very mobile the bandwidth provisioning problem may need trajectory prediction and advance bandwidth reservation in the wireless hop [23]. Therefore, the network has the opportunity to provide users feedback about where in the network (i.e., through which AP) their service requests will be best met using one of the above heuristics. If the AP selected by admission control is different from the one the users are currently associated to, they would be required to change their point of attachment to that AP. While the heuristics described in this section determine the best AP that can service the user's request, there still needs to be a mechanism by which users actually change their association with the APs. The detailed description of these mechanisms is beyond the scope of this paper, but we briefly discuss two mechanisms in the following section.

VI. Online Load Balancing in WLANs

The bandwidth allocation heuristics described in the previous section achieve load balancing by redistributing user load either locally among neighboring APs around the user (e.g., First-Fit, Best-Fit), or globally throughout the entire network (e.g., Balanced-Fit). Through these approaches the network explicitly incorporates user service requests while associating users with APs. Although users initially submit their requests to the network through a default AP association (i.e., one based on strongest signal strength), these approaches may require users' connections to be routed through a different AP that better accommodates their workload.

In this section, we briefly describe two approaches that can be used to provide feedback to users about which AP they are to associate with and how they perform this association. Depending on the admission control heuristic used, this can be done either: (i) by transparently changing the user-AP associations in place without requiring the user to move (*explicit channel-switching*), or (ii) by providing feedback to the user about the location of the AP that provides the service (*network-directed roaming*) [24].

A. Explicit Channel Switching

Figure 3 depicts a WLAN installation with three APs within a subnet providing overlapping coverage in a region, thereby ensuring continuity of network access as users change their location within the network. In order to minimize channel interference, neighboring APs are often configured to operate on different RF channels.

We now consider heuristics that distribute load locally among neighboring APs. In this case, the mobile user is at the periphery of the transmission range of Access Point 1 and within hearing range of APs 2 and 3. When the user submits a service request he is initially associated with AP 1, which is unable to handle his service requirement (as indicated in the $[b_{min}, b_{max}]$ range). The user also records the received signal strength (*Rssi*) of beacon signals received from the other APs and sends the list of APs (AP 2 and AP 3, in this case) during the QoS negotiation phase. Once the network determines the AP that can service the user's request, it returns the AP's identity (SSID, MAC address) and its operating channel to the user. The user now transparently associates with this new AP, by merely changing the RF channel to that of the new AP. The operation of dynamically switching the user's RF channel is supported in current hardware and software.

Explicit channel switching, thus achieves localized load balancing among APs that provide overlapping coverage in the neighborhood of the user. This algorithm trades off signal strength with load by forcing the user to switch from an overloaded AP that has the strongest RF signal to a neighboring lightly loaded AP to which the signal may possibly be weaker.

B. Network Directed Roaming

With explicit channel switching, the network locally redistributes load across neighboring APs by requesting user wireless devices to explicitly change their association from an overloaded AP to a less loaded neighboring AP that can admit the service request. This algorithm relies on the

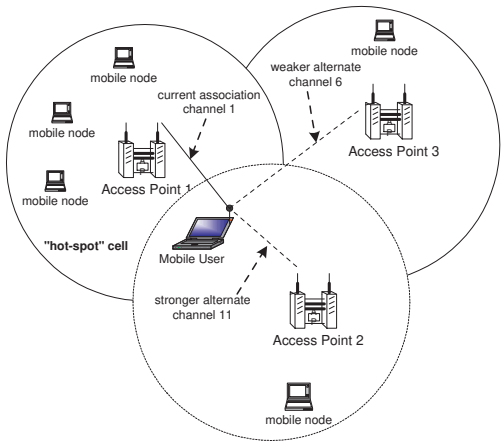


Fig. 3. A WLAN showing overlap between neighboring APs. The dotted lines indicate potential channels that the mobile user can switch to.

existence of at least one AP within range of the user that has enough capacity to honor the QoS requirement. However, complete overlapping coverage may not be available in all scenarios (e.g., in the ends of corridors of an building). Furthermore, none of the APs in the neighborhood of the user may be able to admit the user at the requested service level. Or, the user may not be able to hear a clear signal from any other APs, possibly due to the logistical constraints imposed by her location (like obstructions between her and the AP, causing the SNR value to go below the operable threshold).

When neighboring APs cannot handle user admission requests using explicit channel switching, the network can instead provide feedback suggesting potential locations to which users can roam to get the desired level of service. We call this technique network-directed roaming.

When the network cannot handle a user's service request in the user's current location, the user is likely to roam in the network to find an AP with connectivity. Since the network knows both the locations of APs with available capacity as well as the user's current location, it is ideally situated to direct the user to the AP where requested service can be provided. The Balanced-Fit heuristic, which performs admission control at all APs in the network, can determine which AP, if any can provide this service. Furthermore, with the flexibility to potentially direct users to any AP, the network has the ability to globally balance load across all APs. Of course, this depends upon the cooperation of the user, but it is in the user's best interest to follow the network's roaming suggestion to get service. If the user did not wish to undertake the overhead of physically moving, he could renegotiate the service in the same location with a lower b_{min}.

Network-directed roaming fundamentally depends upon the ability of the network to determine a user's location, and the ability to direct the user to locations with available capacity. There are many techniques that can be used to determine the user's geographic location, each with a varying level of accuracy [25], [26]. Once the user's location is known, a visual way of directing the user to the desired location is to use an indoor navigation map (e.g., an active map) of the coverage area [27]. Alternatively, the network, using pre-defined associations, could translate the destination AP names into specific location names within the network that can aid the user while roaming. For instance, gate numbers could be used in an airport network to indicate roaming destinations to users. The roaming decision also depends upon factors like natural obstacles in the environment, which can be depicted in the active map.

VII. Performance Evaluation

In this section, we investigate the performance of the heuristics described in the previous sections via trace driven simulations. Since the admission control and load balancing heuristics seek to satisfy individual user QoS requirements and distribute load across the network, we use performance metrics to experimentally answer to the following basic questions:

1. What is the effect of performing admission control at each AP on the bandwidth received by users?

2. How does the net offered load at a heavily-loaded AP change as a result of re-allocating users to lightly-loaded APs?

3. What is the effect of these heuristics on overall network utilization?

We begin by describing our simulation methodology and the metrics that we use to quantify the performance of the heuristics, and then present results for three different simulation scenarios. These scenarios use three real workloads from conference [3], corporate [5], and campus [4] WLAN environments.

A. Simulator Setup

We designed a simulator that implements the admission control heuristics on all arriving users in the PAWN. The simulation parameters that can be configured during input are: (i) the number of simulation iterations (ii) the number and location of APs, (iii) the user arrival model, (iv) the location of users relative to the APs, (v) the peak bandwidth at the APs, and (vi) the admission control heuristic to be employed. The simulation parameters that we incorporate directly from the trace are user arrival rate, user data rates, and user session durations.

In all scenarios, we set the capacity of the APs to be the practical achievable limit of 6 Mb/sec [28]. The simulator generates users according to an arrival model that is specified during initialization. For CBR traffic, users generate data according to the actual data rate of the application. For VBR and bursty traffic, we choose the data rates from the three workload studies. We model the size and dimensions of the network from the studies again, but only analyze representative network domains in the larger (i.e., corporate and campus) scenarios.

B. Performance Metrics

To quantify the benefits achieved by admission control on the QoS provided to users, we define the *normalized bandwidth* as the ratio of the actual allocated bandwidth to the maximum desired bandwidth of users. When the APs have adequate capacity to admit all users at their upper data rate bound, b_{max}, the normalized bandwidth approaches 1. Normalized bandwidth reduces as APs are driven close to saturation and incoming users are admitted at data rates much lower than their upper bound. Furthermore, a user's normalized bandwidth is inversely proportional to the time the user spends in the system.

To quantify the benefits achieved by redistributing load across the network, we use the net offered load at the APs and monitor its variation as users get reallocated from a heavily-loaded AP to a lightly-loaded AP.

To further quantify the effect of inter-AP load balancing, we adapt the concept of *balance index* introduced in [29] to reflect the used capacity (bandwidth) in each AP. Suppose B_i is the total throughput of AP i, then we define the balance index β to be:

$$\beta \quad = \quad (\sum B_i)^2/(n * \sum B_i^2)$$

where n is the number of cells over which the load is being distributed. When the load across APs is more or less balanced, the balance index approaches 1. On the other hand, β approaches $1/n$ in the case of heavily unbalanced network load.

In all our results we compare the performance of the three AP allocation heuristics with the base-case approach of default association with an AP based on strongest received signal strength (*Rssi*).

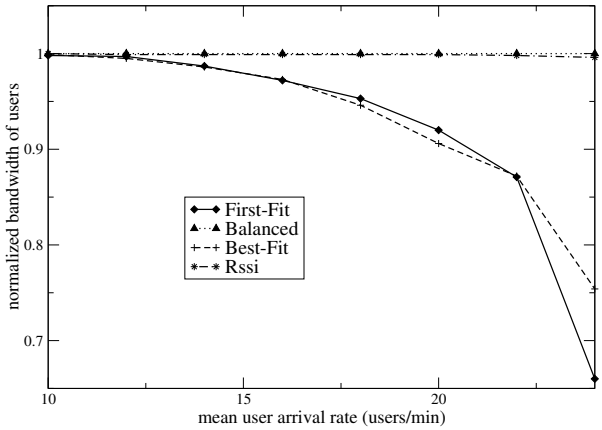
C. Scenario 1 – Conference Room

The first trace that we use to populate our simulation models was collected by Balachandran et al. [3] over three days at the ACM SIGCOMM conference in 2001. The high-level characteristics of the trace are:

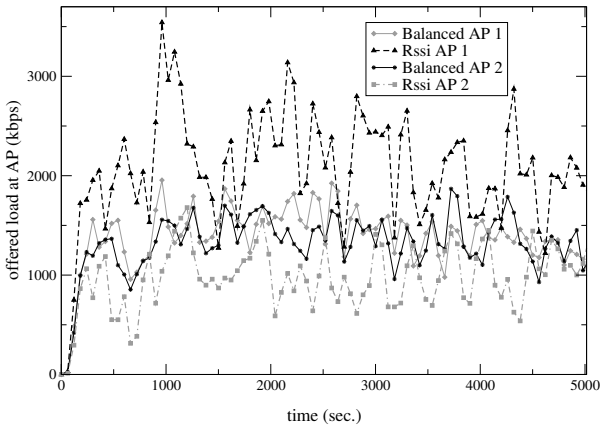
- User arrivals closely follow the conference schedule and are modeled as a Markov-modulated Poisson process (MMPP).
- Users are more or less equally distributed across APs in the conference room. However individual workloads vary widely.
- Users are broadly classified as light, medium, and heavy users depending on their average data rates. Light users have an average data rate of 30 kbps, medium users around 80 kbps, and heavy users around 175 kbps.

Our conference room is a network of area 30m by 30m with three APs linearly placed linearly in the room. We incorporate the user workloads (i.e., light, medium and heavy users), and inter-arrival times (τ = 38 sec.) directly from the trace.

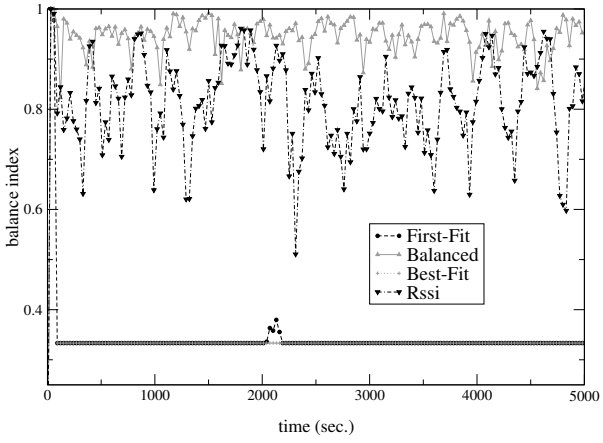
We now study the effect of the heuristics on normalized bandwidth, offered load, and balance index, and then present a discussion of the observations. We study the variation of these parameters over a single conference session that lasts 90 min., with users arriving according to an



(a) The normalized bandwidth of users as a function of system load.



(b) The offered load at an AP during a conference session.



(c) The balance index of the network during a conference session

Fig. 4. The effect of heuristics on user QoS and overall network utilization for the conference WLAN workload.

MMPP (*ON* period = 84 min., mean arrival rate, $\lambda = 1.58$ users/min.).

C.1 Bandwidth Allocated to Users

Figure 4(a) shows the variation in normalized bandwidth allocated to users as a function of increasing system load for a single conference session. The system load on the x-axis is increased by increasing the mean arrival rate of the Poisson process. Balanced-Fit performs the best providing users with nearly 100% of their maximum required capacity. Balanced-Fit is closely followed by Rssi, which performs better than the First-Fit and Best-Fit algorithms. This is because of the fact that in the conference network users are in a constrained space and are equally likely to associate with any one AP. Therefore, the user and workload distribution at all APs are more or less the same except at times when one AP gets a significantly higher share of *heavy* users as seen at times in the trace [3]. In these situations, Balanced-Fit outperforms Rssi.

C.2 Offered Load at the AP

Figure 4(b) plots the offered load at two representative APs, AP2 and AP3, when a conference is in session. AP2 is placed in the center of the room, whereas AP3 is a corner AP providing coverage only to a smaller geographical region around it. The plot shows curves only for Balanced-Fit and Rssi. As users enter the network, Balanced-Fit keeps the offered load at both APs almost constant at around 1.5 mbps (i.e. effective load balancing). On the other hand, Rssi admits users *in place* at the AP closest to their current location and thus witnesses a greater load imbalance between AP2 and AP3. The lower average offered load at AP3 for the Rssi approach is also because AP3 is a corner AP and by default has fewer users that associate to it.

C.3 Balance Index

Figure 4(c) shows how the balance index in the network varies during a conference session. Again, as expected, Balanced-Fit performs near optimal for the given conference workload. The balance index for the Rssi approach has a bursty variation following the change in the offered load at each AP. Comparing with Figure 4(b), we can see that whenever the difference in the offered load at AP1 and AP2 is high, the balance index drops to 0.6 or below, when using Rssi.

C.4 Discussion

The performance of the admission control heuristics on individual user bandwidth allocation and overall network utilization reflects the following characteristics of the conference room environment. First, since it is a constrained space where APs are symmetrically placed in the network, users are equally likely to associate with any one of the APs. Second, since the percentage of users that contribute to significantly larger data transfers is small, an even user distribution is *almost* as good as a load balanced approach. Therefore, Rssi performs almost as well as Balanced-Fit. On the other hand, such workloads do not favor the use of Best-Fit and First-Fit approaches, which are both designed to perform better for a greater variation in the workload distribution among APs.

One implication of the above results concerns capacity planning. Although network designers for such conference-room scenarios may deploy APs to symmetrically cover the space, it may not be sufficient to achieve load balancing. If the network witnesses a greater proportion of heavy users at one particular AP, resulting in a greater disparity in the workload distribution among APs, intelligent load balancing schemes will need to be implemented.

Other PAWN settings like airport gate areas and lounges physically and geographically resemble a conference-room network due to the existence of a constrained space and scheduled times of use. However, the two scenarios have important differences. First, users are more likely to localize themselves to certain particular areas of the network for various reasons such as the proximity of power outlets, or geographic constraints of other services (e.g., gate areas with arriving or departing flights). Second, such networks are highly likely to see a greater variation in workload distribution among APs (e.g., a large group of MP3 downloads, online games) resulting in *hot-spots*. In such cases, using Rssi for allocating users to APs will lower the normalized bandwidth of users and leave the network underutilized. Therefore, the network will benefit by implementing dynamic load balancing.

D. Scenario 2 – Corporate Office Building

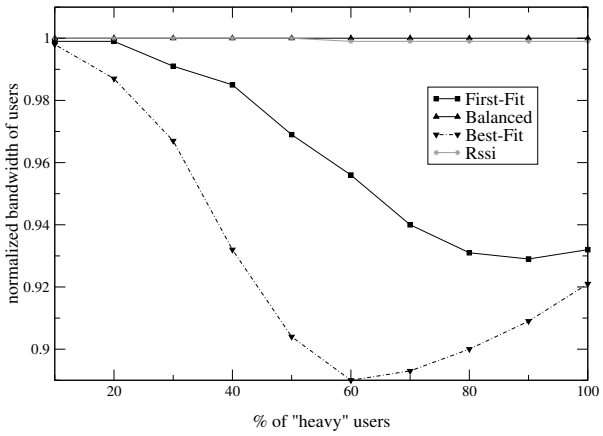
The second scenario we study is a four-week trace collected at a corporate wireless network deployed in three research buildings at the IBM T. J. Watson Research Center [5]. This is a larger trace than the conference network both in terms of the size of the network and the user population. In this trace, Balazinska et al., found that:

- A bulk of the data transfers (over 40%) is accounted for by a very small fraction of the users (< 10%).
- The user data rates and session durations both follow a power law.
- User arrivals follow the regular office schedule.
- Heavy user workloads have average data transfer rates of about 1 Mbps and light users have data rates of around 10 kbps.

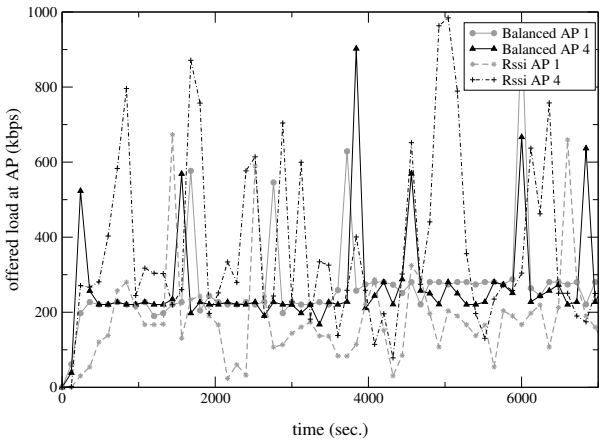
We model the busiest and largest building in the corporate network for which detailed characterizations were available. We model a single floor of the building spanning an rectangular area of 50m by 20m with 8 APs. Four APs are placed in the four corners of the floor and the other 4 APs are symmetrically placed in the hallways in the middle. Using the power law distribution of user workloads in the trace, our network has 10% of heavy users.

Figure 5 presents our results for normalized bandwidth, offered load, and balance index between the hours of 11 am to 1 pm, which witness peak user activity during the day.

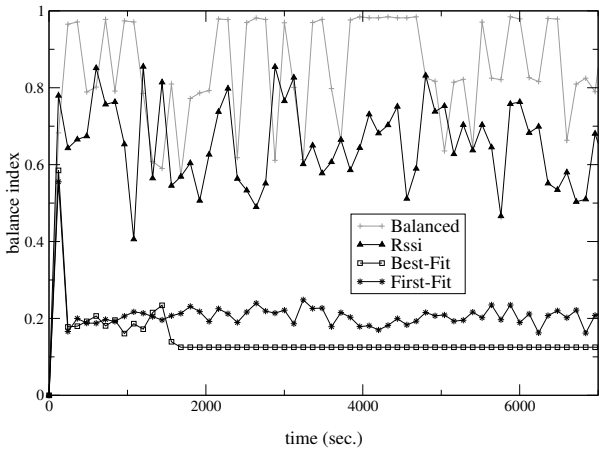
Ideally, we would like to have evaluated the three simulation scenarios using the same invariants. Unfortunately however, the traces have been independently analyzed by three different research groups and are not similar in their



(a) The normalized bandwidth of users as a function of system load.



(b) The offered load at an AP as a function of time.



(c) The balance index of the network as a function of time.

Fig. 5. The effect of heuristics on user QoS and network utilization for the corporate WLAN workload.

characterization. For example, the corporate PAWN trace does not characterize user arrivals during the day. To address this situation, we find the parameter in the trace that best captures the system load – percentage of heavy users in the network.

D.1 Bandwidth Allocated to Users

We now study the variation in normalized bandwidth as a function of the percentage of heavy users in the network. Figure 5(a) plots the normalized bandwidth of users as the number of heavy users in the system increases. As in the conference-room case, Balanced-Fit and Rssi outperform the other admission control approaches, albeit only by a 10% margin. Furthermore, as the percentage of heavy users in the system increases, the normalized bandwidth provided by Best-Fit first decreases and then increases (at over 60% heavy users). First-Fit also sees similar improvement, albeit only with a much higher percentage of heavy users.

This phenomenon can be explained as follows. Best-Fit tries to *tightly* pack an AP until it reaches peak capacity. As the number of heavy users increases, the normalized bandwidth decreases until such time that Best-Fit allocates them to the AP that is being filled. However, once this AP reaches a capacity at which no further heavy user request can be admitted, Best-Fit starts filling the next AP that has least capacity greater than this user's request. This causes the normalized bandwidth to rise again. First-Fit sees a delayed improvement because it takes longer to fill-up an AP to capacity.

D.2 Offered Load at the AP

Figure 5(b) shows the variation in offered load in the network at two APs, AP1 and AP4. Again, the plot compares Balanced-Fit and Rssi approaches only. As would be expected, Balanced-Fit keeps the offered load relatively equal at both APs except during sudden bursts in the offered load (just before $t = 4000$). However, even such situations stabilize rather soon. The Rssi approach, on the other hand, performs poorly with load differences of over 80% between the two APs (at $t = 5000$). Further, it can be seen that the offered load does not stabilize with Rssi because users are not reallocated from AP4 to AP1.

D.3 Balance Index

Figure 5(c) shows the balance index of the network as a function of time. Balanced-Fit spreads the load in the best way possible and hence outperforms the other heuristics, while Best-Fit and First-Fit perform little or no load balancing by preferentially loading an AP. As the percentage of heavy users is higher, the net offered load at these APs is also higher, resulting in a lower balance index than in the conference-room case.

D.4 Discussion

Our performance evaluation of the corporate WLAN scenario indicates that the Best-Fit and First-Fit heuristics perform well when there is a greater proportion of heavy

data transfers in the network. The trace [5] we used witnessed that some users (about 10%) on average transfer over 1 mbps of data, and that average user data rates follows a power law distribution with exponent 0.85 (i.e., $1/x^{0.85}$). We envision that in the future more users will have data-intensive average workloads, thus decreasing the the exponent of this distribution. In such situations, the performance of First-Fit and Best-Fit heuristics will be comparable to the Balanced-Fit approach. This is because as heavy user requests use up significant available capacity in an AP, both First-Fit and Best-Fit naturally start allocating users to other APs, thus gradually spreading the workload across the network and achieving load balancing. On the other hand, the Rssi approach is able to provide a higher normalized bandwidth only as long as the AP has adequate capacity to accommodate the user request.

E. Scenario 3 – University Campus

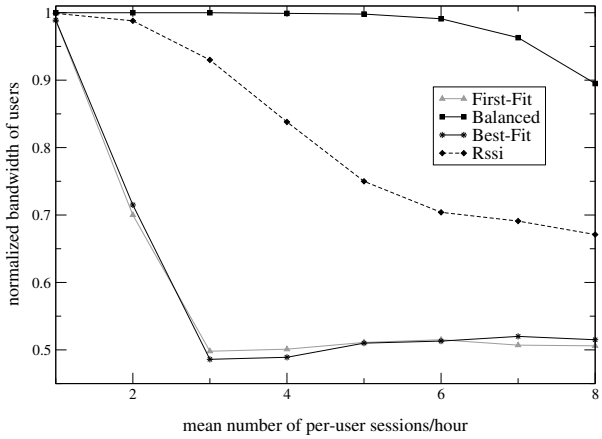
The third scenario that we use in our simulation is a campus WLAN trace collected at several parts of the Dartmouth College campus [4]. This is the largest and most comprehensive trace of a public wireless network spanning 11 weeks and captures the activity of over 2000 users. In this trace, Kotz et. al., discovered that:

- Residential traffic in dormitories dominate all other traffic.
- Network backup and file sharing contribute to a large fraction of the generated traffic.
- Cross-subnet roaming frequently occurred.

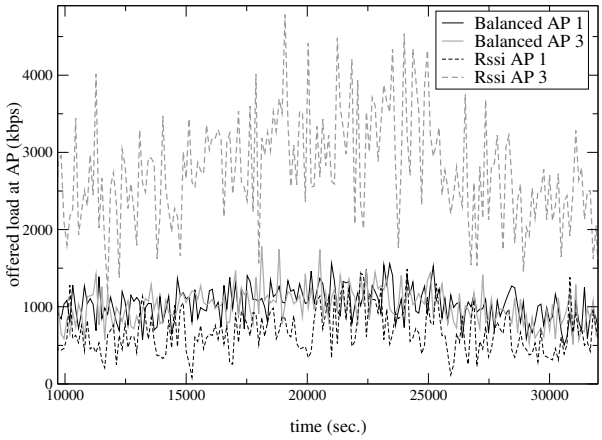
We note that the areas of campus around the classrooms are similar to a conference room setting with constrained space and scheduled times of use. Therefore, we used the dorm as the PAWN for this scenario. Our network spans 35m by 20m with 5 APs in the coverage area. The placement of APs is based on a simple rectangular geometry – one AP in each corner and the fifth AP in the center of the rectangular region.

The study mentions that the dorm had a more or less constant number of users (about 400, on average) during the night hours, which are the 10 hours of peak activity. Therefore, we model a constant user base of 400 users. The study also observed that during the night hours certain parts of the dorms were hot-spots, witnessing heavy average data transfers (e.g., due to high-bandwidth *KaZaA* downloads). We model the central AP to be the one that these heavy users are associated to. As in the corporate WLAN trace, this trace also unfortunately does not characterize user arrivals and user session activity. Therefore, in order to effectively vary system load, we vary the number of per-user (light or heavy) sessions during the 10-hour period of the simulation. As the number of per-user sessions increases, the offered load in the system increases. This simulates the effect of increasing user arrivals.

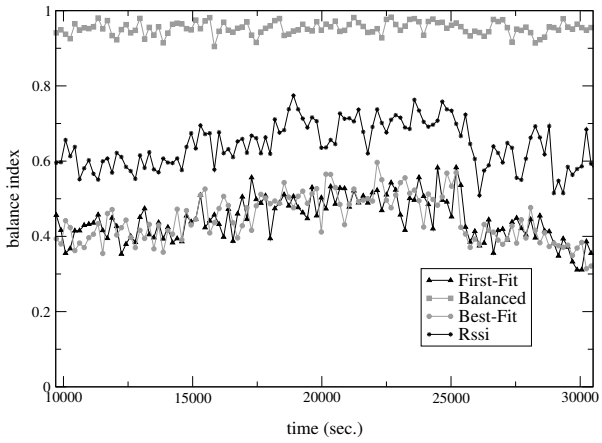
Figure 6 shows our results for normalized bandwidth, offered load, and balance index for one hour of user activity.



(a) The normalized bandwidth of users as a function of active sessions.



(b) The offered load at an AP as a function of time.



(c) The balance index of the network as a function of system load.

Fig. 6. The effect of heuristics on user QoS and network utilization for the campus WLAN workload

E.1 Bandwidth Allocated to Users

Figure 6(a) shows the normalized bandwidth of users as the system load increases. As the load in the system increases, the normalized bandwidth allocated by Rssi decreases steadily. As the number of per-user sessions increases to about 8/hr., the drop in normalized bandwidth for the Rssi is over 30%. This is because there are a greater percentage of heavy users at one particular location in the system. When Rssi is used for associating with an AP in this hot-spot, the AP is unable to satisfy user requests once its capacity is fully used. On the other hand, Balanced-Fit achieves better performance. The normalized bandwidth provided by the Best-Fit and First-Fit heuristics initially decreases with increasing load and then increases as these heuristics start gradually re-allocating users to neighboring APs.

E.2 Offered Load at the AP

Figure 6(b) plots the variation in offered load as a function of time across two APs, AP1 and AP3. AP1 is a corner AP, whereas, AP3 is the hot-spot AP at the center of the PAWN which handles a peak offered load of nearly 5 mbps. As in the previous two scenarios, we compare Rssi and Balanced-Fit only. With association based on Rssi, this AP quickly gets saturated, leaving the network unbalanced and denying further user requests. On the other hand, Balanced-Fit spreads the load among the available APs keeping it nearly balanced over time.

E.3 Balance Index

Lastly, we study the variation in balance index. Figure 6(c) shows the variation in balance index as a function of system load. Again, system load is increased by increasing the average number of per-user sessions per hour. Balanced-Fit performs the best as it achieves the maximum load balancing among APs. It is interesting to see that First-Fit and Best-Fit have very similar performance. This behavior is a result of the inherent workload distribution across APs. APs 1, 2, 4, and 5 are all evenly loaded with the same proportion of light and heavy users, while AP 3 is the hot-spot, with a greater percentage of heavy users. Therefore, both First-Fit and Best-Fit are equally likely to choose the same lightly-loaded AP (1, 2, 4, or 5) to allocate users to, since there is no inherent ordering among these APs. Rssi performs better than Best-Fit and First-Fit, but has an average balance index of 0.6 because of the heavy load in AP 3.

E.4 Discussion

Among the three traces that we used to populate our simulation models, the campus dorm trace had the greatest disparity in user workload distribution among APs. Two high-level characteristics of the user behavior were: (i) users have a wide variation in their workloads, and (ii) certain specific regions in the network witness higher-bandwidth data transfers than others, creating localized hot-spots. Under such conditions, the Rssi approach fails to provide users with their requested bandwidth once the overloaded AP reaches capacity. Furthermore, it does not improve the imbalance in the offered load across APs. This is not the case with the conference-room and corporate traces where, for a bulk of the trace, users are fairly evenly distributed across APs, and users have more or less similar workloads.

We now discuss how the correlation between number of users and the offered load at an AP can influence the decision on the admission control heuristic to be used. Whenever there is a *weak* correlation between number of users associated with an AP and the offered load at those APs, as in the campus and corporate WLAN traces, it creates hot-spots in the network where the APs are more likely to get saturated. In such situations, the Rssi approach will not perform well and the network will benefit from explicitly re-allocating users using a heuristic such as Balanced-Fit. On the other hand, if the correlation between number of users and workload improves and the network has a symmetric distribution of APs where users are equally likely to associate with any AP, Rssi is as effective as Balanced-Fit.

Lastly, we discuss scenarios where the Balanced-Fit may not perform well in offering high normalized bandwidth to users. Consider a network where a group of many small bandwidth requests are followed by a group of large (i.e. around $B/4$, where B is the AP's capacity) requests arrive in the network. A Balanced-Fit approach would spread the small requests across all APs keeping the offered load balanced across APs. This form of allocation uses up capacity nearly equally at all APs, not leaving adequate capacity anywhere for the second group of large user requests. As a consequence, the larger user requests cannot be admitted to any of the partially filled APs. In such, situations, approaches like Best-Fit and First-Fit will more optimally use the overall network capacity.

VIII. Conclusions and Ongoing Work

This work has been motivated by three key observations made in three recent PAWN workload characterization studies: (i) user loads are often time varying and location-dependent; (ii) user load is often unevenly distributed across access points (APs); and (iii) the load on the APs at any given time is not well correlated with the number of users associated with those APs. In order to address this problem, we propose heuristics to adaptively and dynamically vary the bandwidth allocated to users in the wireless hop within certain bounds. Furthermore, these heuristics change user-AP associations and thus alleviate user congestion at popular locations, providing inter-AP load balancing.

This paper makes the following contributions:

1. We present the problem of first-hop wireless bandwidth allocation as a special case of the well-known online load balancing problem and present three online heuristics for first-hop bandwidth allocation. These heuristics improve the degree of balance in the system by over 45% and allocate over 30% more bandwidth to users than current ap-

proaches;

2. We prove that the general *offline* problem (i.e., where we have global knowledge of user arrivals and requests) of finding an optimal assignment of users to APs in an arbitrary network with arbitrarily sized user bandwidth requests, is NP-complete;

3. We propose three different heuristics for allocating users to APs based on their bandwidth requirements and evaluate their performance via trace driven simulations.

Our high-level results indicate that for all three scenarios Balanced-Fit outperforms all the other admission control heuristics and the base case approach of association based on received signal strength (Rssi). On average, Balanced-Fit, allocates over 30% more normalized bandwidth to users and improves the network balance index by over 45%. Rssi performs well, in scenarios with even user distribution across APs and when the number of users and offered load at the APs are relatively well correlated. Best-Fit and First-Fit improve in their ability to allocate bandwidth to users as the proportion of heavy data transfers increases. To the best of our knowledge, ours is the first study of wireless LAN bandwidth provisioning incorporating real WLAN workloads in performance evaluation.

References

[1] IEEE, "802.11b/d3.0 Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification," August 1999.

[2] P. Bahl, A. Balachandran, A. Miu, W. Russell, G. M. Voelker, and Y-M. Wang, "PAWNs: Satisfying the Need for Secure Ubiquitous Connectivity and Location Services," *IEEE Wireless Communications Magazine*, pp. 40–48, February 2002.

[3] A. Balachandran, G. M. Voelker, P. Bahl, and P. V. Rangan, "Characterizing User Behavior and Network Performance in a Public Wireless LAN," in *Proc. ACM SIGEMTRICS'02*, June 2002.

[4] D. Kotz and K. Essien, "Characterizing Usage of a Campus-wide Wireless Network," in *Proc. ACM MobiCom'02*, March 2002, pp. 107–118.

[5] M. Balazinska and P. Castro, "Characterizing Mobility and Network Usage in a Corporate Wireless Local-Area Network," in *Proc. MobiSys'03*, May 2003 (To Appear).

[6] D. Tang and M. Baker, "Analysis of a Local-Area Wireless Network," in *Proc. ACM MobiCom'00*, August 2000, pp. 1–10.

[7] A. Ayyagiri, Y. Bennet, and T. Moore, "IEEE 802.11 Quality of Service," February 2000.

[8] L. Qiu and P. Bahl and A. Adya, "The Effect of First-Hop Wireless Bandwidth Allocation on End-to-End Network Performance," in *Proc. NOSSDAV'02*, May 2002, pp. 85–93.

[9] M. Barry, A. T. Campbell, and A. Veres, "Distributed Control Algorithms for Service Differentiation in Wireless Packet Networks," in *Proc. IEEE Infocom'01*, April 2001.

[10] S. Lu, V. Bhargavan, and R. Srikant, "Fair Scheduling in Wireless Packet Networks," in *Proc. ACM Sigcomm'97*, August 1997, pp. 63–74.

[11] N. H. Vaidya, P. Bahl, and S. Gupta, "Distributed Fair Scheduling in a Wireless LAN," in *Proc. ACM MobiCom'00*, August 2000, pp. 167–178.

[12] E. Ng, I. Stoica, and H. Zhang, "Packet Fair Queuing Algorithms for Wireless Networks with Location-Dependent Errors," in *Proc. IEEE Infocom'98*, March 1998, pp. 1103–1111.

[13] Agere Systems, "Firmware Update for ORINOCO PC Cards v7.28 - Spring 2001 release," April 2001.

[14] Cisco Systems Inc., "Data Sheet for Cisco Aironet 350 Series Access Points," June 2001.

[15] P. Hsiao, A. Hwang, H. T. Kung, and D. Vlah, "Load-Balancing Routing for Wireless Access Networks," in *Proc. IEEE Infocom'01*, April 2001, pp. 986–995.

[16] S. V. Hanly, "An Algorithm for Combined Cell-Site Selection and Power Control to Maximize Cellular Spread Spectrum Capacity," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1332–1340, September 1995.

[17] Y. Azar, A. Z. Broder, and A. R. Karlin, "On-line load balancing," in *Proc. 33^{rd} IEEE Annual Symposium on Foundations of Computer Science*, October 1992, pp. 218–225.

[18] S. Phililips and J. Westbrook, "Online Load Balancing and Network Flow," in *Proc. ACM STOC'93*, June 1993, pp. 402–411.

[19] S. Lu, K. W. Lee, and V. Bhargavan, "Adaptive Service in Mobile Computing Environments," in *Proc. IFIP IWQoS'97*, May 1997, pp. 25–36.

[20] Wayport Passes 1 Million Connections, "*http://www.80211-planet.com/*," 2002.

[21] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman and Company, 1979.

[22] S. Martello and P. Toth, *Knapsack Problems: Algorithms and Computer Implementations*, Wiley, 1990.

[23] T. Liu, P. Bahl, and I. Chlamtac, "Mobility Modeling, Location Tracking, and Trajectory Prediction in Wireless ATM Networks," *IEEE Journal of Selected Areas in Communications*, vol. 16, no. 6, pp. 922–936, August 1998.

[24] A. Balachandran, G. M. Voelker, and P. Bahl, "Hot-Spot Congestion Relief in Public-Area Wireless Networks," in *Proc. Workshop on Mobile Computing Systems and Aplications, WMCSA'02*, June 2002, pp. 70–80.

[25] P. Bahl, V. N. Padmanabhan, and A. Balachandran, "A Software System for Locating Mobile Users: Design, Evaluation and Lessons," Tech. Rep. MSR-TR-2000-12, Microsoft Research, February 2000.

[26] S. Y. Seidel and T. S. Rapport, "914 MHz Path Loss Prediction Model for Indoor Wireless Communication in Multi-floored Buildings," *IEEE Transactions on Antennas and Propagation*, vol. 40, no. 2, pp. 207–217, February 1992.

[27] B. N. Schilit and M. Theimer, "Disseminating Active Map Information to Mobile Hosts," *IEEE Network*, vol. 8, no. 5, pp. 22–32, September 1994.

[28] A. Vasan and U. A. Shankar, "An Empirical Characterization of Instantaneous Throughput in 802.11b WLANs," Tech. Rep. CS-TR-4389, University of Maryland, November 2002.

[29] D. Chiu and R. Jain, "Analysis of the Increase and Decrease Algorithms for Congestion Avoidance in Computer Networks," *Journal of Computer Networks and ISDN Systems*, vol. 17, no. 1, pp. 1–14, June 1989.