

# Hot-Spot Congestion Relief in Public-area Wireless Networks

Anand Balachandran  
U. C. San Diego  
9500 Gilman Dr. 0114  
La Jolla, CA 92093  
anandb@cs.ucsd.edu

Paramvir Bahl  
Microsoft Research  
One Microsoft Way  
Redmond, WA 98052  
bahl@microsoft.com

Geoffrey M. Voelker  
U.C. San Diego  
9500 Gilman Dr. 0114  
La Jolla, CA 92093  
voelker@cs.ucsd.edu

## Abstract

*Wireless LAN administrators are often called upon to deal with the problem of sporadic user congestion at certain popular spaces (“hot-spots”) within the network. To address this problem, we describe and evaluate two new approaches, explicit channel switching and network-directed roaming for providing hot-spot congestion relief while maintaining pre-negotiated user bandwidth agreements with the network. The goals of these algorithms are: (i) to accommodate more users by dynamically providing capacity where it is needed, when it is needed; (ii) to improve overall network utilization by making more efficient use of deployed resources; and (iii) to guarantee at least a minimum amount of bandwidth to users. We propose that both the network and its users should explicitly and cooperatively adapt themselves to changing load conditions depending on their geographic location within the network. We describe how these algorithms enable the network to transparently adapt to user demands and balance load across its access points (APs). We evaluate the effectiveness of these algorithms on improving user service rates and network utilization using simulations. Our algorithms improve the degree of load balance in the system by over 30%, and user bandwidth allocation by up to 52% in comparison to existing schemes that offer little or no load balancing.*

## 1. Introduction

Advances in communication technology and the proliferation of lightweight, hand-held devices with built-in, high-speed radio access are making wireless access to the Internet the common case rather than an exception. The significant performance benefits of wireless LANs [2, 13] have made them an ideal networking platform for offices, homes, and public places like airports, malls, hotels, etc. A key challenge to the host organization deploying these pub-

lic wireless networks is capacity planning, making the best use of the available network resources to derive the best return on its investment while at the same time satisfying user service demands.

Recent studies of deployments of public-area wireless networks 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 wireless access points (APs) [26, 27]. Users tend to localize themselves in particular areas of the network for various reasons, such as the availability of favorable network connectivity, the proximity of power outlets, or geographic constraints of other services (e.g., airport gate areas with arriving and departing flights). A key consequence of this behavior is sporadic user congestion at certain popular spaces (“hot-spots”) within the network. At any one time, a large percentage of the mobile users communicate with a small subset of the APs in the wireless LAN. These user concentrations create an unbalanced load in the network, and complicate the capacity planning problem, making it difficult to accommodate heavy, concentrated load in different parts of the network without significant, and costly, over-engineering.

To address this problem, we describe and evaluate two new approaches for providing hot-spot congestion relief while maintaining pre-negotiated user bandwidth agreements with the network. The goals of these algorithms are: (i) to accommodate more users by dynamically providing capacity where it is needed, when it is needed; (ii) to improve overall network utilization by making more efficient uses of deployed resources; and (iii) to guarantee at least a minimum amount of bandwidth to users. We propose that both the network and its users should explicitly and cooperatively adapt themselves to changing load conditions depending on their geographic location within the network. When a user requests service from the network in an overloaded region, the network tries to adapt itself to handle the user service request by readjusting the load across its APs. If the network cannot adapt itself to handle the user’s request, it provides feedback to the user about where the user

can move to get the service requested. As a result, overall network utilization increases, and users get the QoS they request, either transparently or by explicitly moving to specific locations within the network.

This paper makes the following contributions:

1. We exploit mobile-host radio frequency (RF) channel agility through *explicit channel switching*. This algorithm assists in changing user-AP associations on the fly and trades off signal strength with load by forcing a mobile user to change association from an overloaded AP with a stronger signal to a neighboring lightly loaded AP with a possibly weaker signal.
2. We exploit mobile-host location agility through *network directed roaming*. When the user service requests are beyond the capability of the network to transparently adapt to through channel switching, the network balances load by providing explicit feedback to users about where to roam to get the services they require and what the network can provide.
3. We describe a QoS-negotiation and admission control protocol that operates in conjunction with the aforementioned load-balancing algorithms. This admission protocol provides users with a bounded QoS guarantee and dynamically provisions the available resources in each cell, maintaining QoS within those bounds.

We evaluate the benefit of the load-balancing algorithms on network utilization, using simulations of a public-area wireless network. The simulation results show that our algorithms perform well in a variety of user configurations. We use a parameter called balance index to evaluate the extent of balance achieved between the cells in the network. Our algorithms improve the balance index by over 30%, and user bandwidth allocation by up to 52% in comparison to existing schemes that offer little or no load balancing. We analyze in detail the costs involved in implementing our algorithms and show that our algorithms are scalable, and that the benefits derived outweigh user and network overhead. Based upon these results, we conclude that public-area wireless networks would benefit greatly from the use of these algorithms.

The rest of this paper is organized as follows. In Section 2, we discuss related work. In section 3, we discuss the issues involved in providing service differentiation in public wireless networks. In Section 4, we present the design of our adaptive load balancing algorithms. In Section 5, we evaluate the performance benefits of our techniques via simulation. Finally, we conclude in Section 6.

## 2. Related Work

The state of the art for channel access in wireless LANs is the IEEE 802.11 CSMA/CA protocol with the *Distributed Coordination Function* (DCF) for media access [13]. DCF itself does not guarantee anything more than best-effort service for the mobile hosts. To support real-time services, the standard provides a polling based media access in the *point coordination function* (PCF) mode. However, PCF is not supported by most wireless vendors and has been shown to perform poorly in the presence of DCF [29]. As a result, the 802.11 Working Group is 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 [3]. However, this scheme does not take into account the dynamically varying nature of the wireless medium.

There have been a number of other proposals to enhance or modify the MAC protocol in wireless LANs to provide service differentiation using centralized and distributed schemes [8, 18, 28]. 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 [1, 11]. APs supporting this feature maintain a measurement of the load in their respective cells and broadcast beacons containing this load to users in the cell. New users receive beacons from multiple access points and use this information to determine and associate with the least-loaded AP. However, these techniques do not take into account explicit user service (QoS) requirements and are local in scope, distributing users only across available overlapping cells.

In [12], the authors present load-balancing algorithms for efficient routing in multi-hop wireless access networks. Although some of the ideas expressed by them are similar to the algorithms described in this paper, there are some basic differences. First, 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. Second, they do not consider how network load changes with arriving and departing users; this cannot be neglected in public-area wireless networks. Furthermore, many of the assumptions made by the authors relate to

multi-hop wireless networks and do not apply to the case of public-area wireless networks.

Earlier work has incorporated user location into a different network setting, routing algorithms for ad-hoc networks. In [16], the authors propose that the network asks nodes to change their roaming direction to assist in the delivery of packets among nodes in a disconnected, ad-hoc network. Although we use the same basic idea of having the network suggest that users roam, in our network-directed roaming the network makes the suggestion for the direct benefit of the roaming node, rather than other nodes in the network. And [15] shows that ad-hoc routing algorithms can incorporate user location to improve routing performance.

Our contributions differ from related work in three significant ways: (i) we capitalize on typical user behavior in public-area wireless networks and thus focus on providing QoS to users in the network as a whole rather than within one specific cell; (ii) we focus on improving network utilization by redistributing users from heavily loaded cells to less heavily loaded neighboring cells, and thus, (iii) we increase the chances of being able to guarantee a minimum QoS level to users in the network depending on the degree of their channel and location agility<sup>1</sup>.

We would like to emphasize that our algorithms are not a new QoS protocol. Rather, our techniques can benefit from any of the QoS-aware MAC or higher layer protocols. At the same time, our techniques distribute user loads within the entire network to achieve high utilization.

### 3. Providing Per-User QoS in Wireless LANs

In order to adequately support both traditional data services together with emerging multimedia services (mobile IP telephony, streaming audio and video, etc.), future wireless network infrastructures need to:

- explicitly establish service level agreements (SLAs) with each mobile user at the beginning of service and repeatedly make admission control decisions on user requests as users move within the network (and thus change their point of attachment), and
- implement QoS-aware MAC algorithms that prioritize channel access for traffic classes with specific QoS (throughput, delay, jitter) needs.

In this section, we describe a simple model for how users can negotiate their QoS needs with the network and how the network uses admission control to accept or deny user service requests. While admission control helps the network

<sup>1</sup>We note here that we are primarily interested in providing per-user statistical guarantees rather than per-application deterministic guarantees. We leave it to the user to use the allocated bandwidth among applications in the most appropriate way.

to effectively plan the capacity in each cell, the bandwidth thus negotiated with each user is provisioned through MAC layer service disciplines. We begin by introducing the notion of QoS bounds, which users specify in order to indicate their service requirements to the network.

#### 3.1. Service Level Specification

Since the last-hop bandwidth in a wireless network is a scarce, shared resource, providing acceptable QoS to contending users necessitates some form of negotiation between users and the network. While wired networks provide users with fixed levels of deterministic or statistical QoS guarantees, through bandwidth reservation, many aspects of wireless networks preclude exact control over the network bandwidth. First, wireless networks are characterized by time-varying and location-dependent errors in the channel [20]. Second, users in a wireless network tend to be mobile and the QoS that has been negotiated in one cell may not be honored as the user moves to other cells because those cells may not be able to provide the required capacity [19]. We envision that organizations deploying public-area wireless networks would want to support a wide range of service models from plain connectivity without guarantees (best-effort service) to differentiated QoS (as is provided by the Diffserv model [9] in the wired Internet).

To initiate QoS negotiation, users establish a *Service Level Specification* (SLS) with the network before starting their session. Each SLS specifies a minimum and a maximum bound on the bandwidth  $min, max$  that the user expects to be provided at that level. To aid the users in making a decision about their SLS, the network broadcasts service announcements in each cell advertising the available capacity. Alternatively, the SLS for users could be driven by some pre-negotiated policy between the host organization deploying the network and other corporations. For example, a corporation might negotiate a service package with a local airport such that, whenever any of its employees accesses their network, the airport would provide a minimum level of connectivity at a pre-determined charge. Providing a bandwidth range in the SLS enables the network to adaptively vary the level of QoS provided to the user as the effective capacity of each cell changes with time due to the dynamics of the wireless environment; the network attempts to guarantee the user a data rate of  $min$  with possible provisioning up to  $max$ . If the user does not specify QoS bounds in the SLS, the network assumes a best-effort service request. Each cell in the network has a certain fraction of its capacity reserved for best-effort users. Reserving bandwidth for best-effort connections allows the network to be backward compatible with existing schemes. Users who do not request any specific service guarantee can continue to obtain service without any upgrades to their hosts.

The negotiation of service levels using QoS bounds in the wireless access network has a number of advantages. First, it enables the host organization deploying public wireless networks to adaptively plan its capacity to increase network utilization (the primary goal of our algorithms) and thus maximize the return on its investment. Second, service negotiation using an SLS helps users negotiate a pipe to the wired backbone, with a guaranteed minimum bandwidth and excess capacity provisioning beyond  $min$ , as available. Third, QoS bounds can be used to characterize user workloads for both real-time multimedia (voice and video) and bursty data traffic. Fourth, different service levels allow the creation of a tiered service model that benefits preferred customers. The SLS can be used to specify other QoS parameters like delay, but we focus our discussion on bandwidth only.

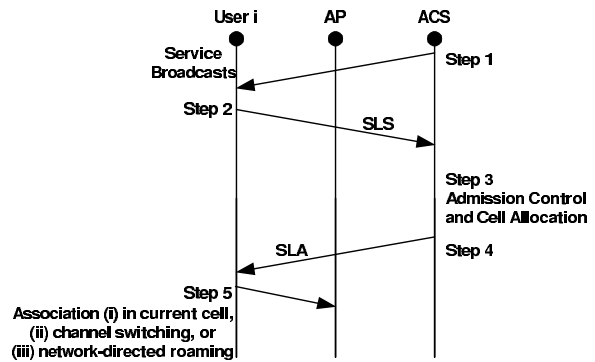
### 3.2. Negotiation and Admission Control

The admission control and load-balancing algorithms rely on the availability of state information about the local network, such as available capacity in each cell, number of users per cell, QoS bounds of admitted users, etc. Whether this information is stored in the APs in every cell (distributed) or in a single access server in the network (centralized), is a design choice.

In the centralized approach, there is an *admission control server* (ACS) that receives and processes the SLS requests from users. There are a number of benefits in using this approach. First, since the ACS maintains all per-cell and per-user state for the network it can monitor and control the use of the wireless bandwidth in the entire network. Global knowledge of system state enables the ACS to easily identify hot-spots. Second, moving state away from the access points to the ACS keeps the APs lightweight and avoids the need for inter-AP communication when redistributing users. In addition, it helps to keep the system hardware agnostic, independent of the firmware and access technology supported in the AP. With a decentralized scheme, APs have to continuously exchange state information, potentially as often as the state changes in the network. Finally, establishing an SLS with the central server helps users to create a context for their service, which can be broadcast to relevant APs as the user roams in the network, thereby simplifying context-transfer between cells [25].

The decentralized approach has advantages too. First, it stops all unauthenticated traffic at the edge of the network and is thus a more secure design. Second, managing state at each individual AP is both modular and scalable. Third, distributed state maintenance reduces the network management overhead due to an additional server (ACS) that would otherwise be involved in the centralized approach.

**3.2.1. Sequence of Operations.** We now describe the de-



**Figure 1. Diagram showing the sequence of steps involved in QoS negotiation and admission control.**

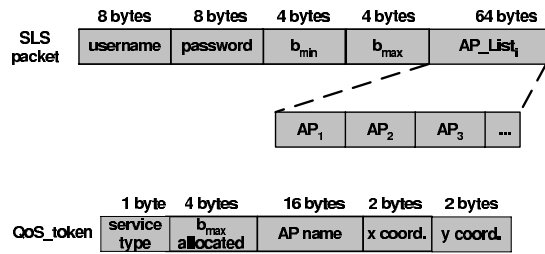
tailed steps involved as a user  $i$  negotiates her QoS level with the network, as shown in Figure 1.

**Step 1:** Upon entering the access network, user  $i$  discovers the existence of a service through a pre-existing configuration protocol (e.g., DHCP with options) or via periodically broadcasted beacons in the local network. The user's wireless adapter associates, by default, to the AP from which it senses the strongest signal.

**Step 2:** After detecting the network, the user performs authentication and service level specification. To reduce the communication overhead for the user and the latency involved in getting network access, the user submits the initial service specification,  $i$ , while authenticating with the access network. The SLS, shown in Figure 2, includes: the user's credentials required for authentication (username, password, digital certificate etc.); the QoS bounds  $min,i$ ,  $max,i$ ; and a list of the APs,  $i$ , that are within communication range of the user. The latter information is obtained using the inherent ability of the user's wireless adapter to scan the local network to locate the AP with the strongest signal.  $i$  contains only those APs whose SNR is above a certain threshold, which varies depending on the user's geographic location in the network. The APs in  $i$  are ranked in decreasing order of SNRs. The user's mobile host also records the RF channel of each AP in  $i$ , but does not include this information in the SLS.

**Step 3:** If the SLS includes no QoS bounds, no admission control is performed and the user competes for a fair share of the available reserved capacity for best-effort service. Otherwise, the ACS uses the information in the SLS to perform an admission test to determine which of the APs in  $i$  can admit the user's connection at the minimum capacity  $min,i$ . The goal here is to identify the cell (if one exists) where the user's QoS bounds can be adequately met.

The admission test is initially done using the lower band-



**Figure 2. The format of the user's SLS and the QoS token returned to the user after admission control.**

width bound,  $min_i$ . If more than one AP can admit the request at this rate, the ACS determines which of those has most available capacity. By choosing the AP with highest capacity, the admission control procedure tries to admit each user  $i$  at a capacity allocation as close to  $max_i$  as possible, thereby maximizing the total utilization in the network<sup>2</sup>. Once the test succeeds, the ACS admits the user upon successful authentication. Thus, the admission control phase tries to determine the best allocation of users to cells to achieve two goals: (i) users are allocated to the cells where their capacity requirements can best be met, and (ii) allocation of users to lightly loaded cells helps to reduce the imbalance between the cell loads.

**Step 4:** The next step is to inform the user of the level of service that she has been granted through a *Service Level Agreement* (SLA) returned to the user. The SLA includes user  $i$ 's network access key and a QoS token (see Figure 2). The QoS token is valid for a time  $t_i$  indicated in the SLA, after which renegotiation is necessary. The QoS token contains: the permissible QoS bound after admission control (note that the  $max$  in this bound can be different from the  $max_i$  indicated in the SLS); a  $roam$  field indicating if the service is provided in place or if it requires roaming; the AP which provides the service; and the physical  $(x, y)$  coordinates of the AP.

**3.2.2. Incorporating Channel Errors.** Since wireless networks are prone to location-dependent, time-varying channel errors, the effective bandwidth negotiated in a cell may not be available to users throughout their session. Hence no absolute guarantees can be made on bandwidth and delay. Although QoS-aware, MAC-level scheduling algorithms can ensure long-term fairness for user connections even in the presence of errors [18, 20], we enable users to adapt to channel errors as described below. Upon receiving

<sup>2</sup>Getting an optimal allocation of users to cells in their APList (location constraints) that maximizes overall network utilization requires complete knowledge of all future events and is an NP-hard problem [4]. Therefore, with only past and current state information, we use the greedy strategy described in Step 3.

network access, the user's mobile host constantly monitors the channel for errors by keeping track of the number of MAC-level retransmissions over a period of time. If the retransmissions cross a certain pre-determined threshold during the user's session due to a poor channel, she can renegotiate her service with the network by issuing a new SLS that does not include the current AP. The network now again performs admission control as described. We are investigating the impact of channel errors on the performance of our algorithms as part of future work.

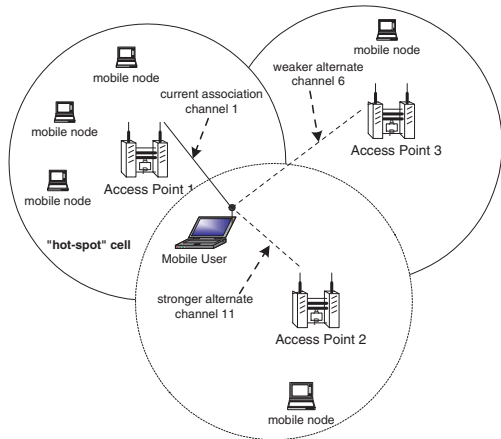
## 4. Overview of Adaptive Load Balancing Algorithms

When the network receives the user's SLS, it determines whether (i) it can provide the requested service in the user's current cell without violating the QoS bounds for admitted users (no action required), (ii) it can transparently handle the user's service requirement by redistributing load among neighboring cells (explicit channel switching), or (iii) it should provide feedback to the user about the closest cell that can handle the requested service (network-directed roaming). Note that explicit channel switching locally distributes load within the neighborhood of access points around the user, whereas network-directed roaming has the flexibility to globally distribute load throughout the entire network. These algorithms operate on the assumption that users more or less stay localized within a single cell, which is true with the case of laptop users in many public-area wireless networks [7, 14, 27]. However, if users are very mobile the bandwidth provisioning problem may need trajectory prediction and advance bandwidth reservations in cells [17]. We now describe each of our algorithms in detail.

### 4.1. Explicit Channel Switching

In most wireless LAN installations, neighboring APs within a subnet often provide overlapping coverage in the region, thereby ensuring continuity of network access when users roam. To maximize system capacity and keep the interference to a minimum, neighboring APs are configured to operate on different RF channels as shown in Figure 3. The mobile user is at the boundary of Access Point 1 and within hearing range of APs 2 and 3.

When the user submits the SLS to the network, the APList field in the SLS will contain APs 1, 2 and 3. As mentioned in Step 3 of the admission control algorithm, the AP to which the user is initially associated (AP 1, in this case) may not be able to handle her QoS requirement (indicated in the  $min, max$  range). After performing admission control at all the three cells, the network determines



**Figure 3. A Wireless LAN showing overlap between neighboring APs. The dotted lines indicate potential channels that the mobile user can switch to.**

which cell (if any) can admit the user. If this is different from the cell in which the user currently is, the user will have to switch her operating channel to that of the new cell. Both the AP that provides the service and its operating channel are conveyed in the SLA that the network returns to the user. The user now transparently associates with the AP indicated in the SLA. If more than one AP can admit the user's service request, the one with the strongest signal is used. With explicit channel switching, association with an AP is not merely on the basis of signal strength, but is determined by whether that cell can accommodate the user's workload. The algorithm trades off signal strength with load by forcing the user to switch from an overloaded cell containing the AP with a stronger signal to a neighboring lightly loaded cell where the signal to the AP may possibly be weaker.

## 4.2. 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 existence of at least one AP within range of the user that has enough capacity to honor the QoS requirement. However, this assumption may not always be valid. For example, none of the APs indicated in the *APList* field of the user's SLS 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 a cell 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 a cell where requested service can be provided. 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, she could renegotiate the service through a new SLS with a lower *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. We describe how the network can do both tasks in the following sections.

**4.2.1. Determining Current User Location.** There are many techniques that can be used to determine the user's location, each with a different level of accuracy [6, 24]. The choice of which location estimation algorithm to use is a trade off between accuracy of location resolution and ease of implementation. One approach that we have previously investigated, called RADAR, uses signal strengths from a network of APs to estimate user location [6]. Our study showed that this technique can estimate user location to within a few meters, a degree of accuracy that is more than sufficient for network-directed roaming. Since RADAR requires no additional hardware, it is a technique that can be deployed in any public-area wireless network.

**4.2.2. Directing Users to New Locations.** Recall that in Step 3 of the admission control algorithm, the network returns an SLA giving the ( , ) coordinates of the AP that can service the user's request. If the roaming flag in the SLA is set, a software module on the mobile host determines its location using one of the algorithms described in the previous section. One possible visual way of directing the user to the desired location is to use an indoor navigation utility (e.g. an active map) of the coverage area [23]. Both the user's location and the ( , ) coordinates of AP (from the SLA) are indicated on the active map. If there is more than one AP that can service the user's QoS request, the active map includes the distances to each AP sorted from nearest to farthest with respect to the user's current location.

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.

## 5. Experimental Evaluation

We now investigate the performance of the algorithms presented in the last two sections. Since the algorithms seek to redistribute load across the network and satisfy individual user service requirements, our goal is to experimentally answer two basic questions:

- What is the effect of the load balancing algorithms on overall network utilization?
- What is the effect of performing network-wide admission control of user's requests on the QoS guarantees received by different classes of users?

To quantify the benefits achieved by redistributing load across the network, we adapt the concept of *balance index* introduced in [10] to reflect the used capacity (bandwidth) in each cell. Suppose  $T_i$  is the total throughput of cell  $i$ , then we define the balance index to be:

$$BI = \frac{\sum_i T_i}{N \cdot \sum_i T_i}$$

where  $N$  is the number of cells over which the load is being redistributed. In the case of channel switching,  $N$  is the number of cells in the subnet, while in network-directed roaming,  $N$  is the number of cells in the entire network. The balance index has the property that it is 1 when all cells have exactly the same throughput. When the cells are heavily unbalanced, it gets closer to 0.

### 5.1. System Parameters

For our experimental evaluation, we use a model of a public-area wireless network constructed using the Opnet simulation tool [21]. The physical and MAC layers of the wireless network are modeled according to the IEEE 802.11 standard with direct sequence spread spectrum in physical layer and DCF in the MAC layer. We use a network comprised of six wireless cells that spans a rectangular area of 300m by 300m. Each cell is centrally managed by an AP and the six APs provide overlapping coverage in the entire region. Each AP operates at a raw data rate of 2 Mbps and a power level of 100 mW, providing an operating range of 100m. We are aware that the current IEEE 802.11 standard

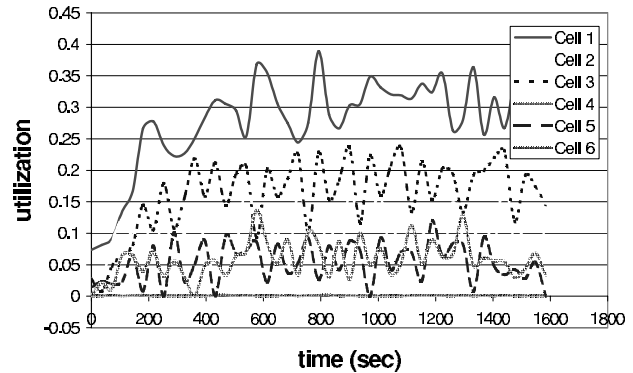


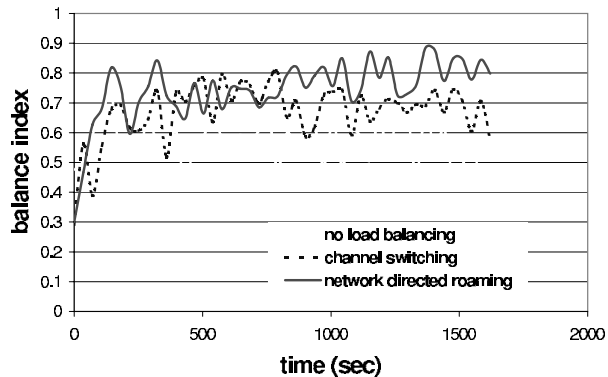
Figure 4. The network utilization in each cell.

supports a data rate of 11 Mbps. However, due to the high memory demands in generating workloads to load a network of 6 cells, we chose to scale our tests down to the 2 Mbps data rate. We conducted the basic experiment using a data rate of 11 Mbps; the results showed no significant departure from those using the 2 Mbps model. Neighboring cells are configured on different RF channels in order to avoid inter-cellular interference. To model channel error, we use the basic signal propagation and channel error models provided by Opnet that includes thresholds on packet bit error rate.

We model four classes of users, and each class has a different application profile reflecting the traffic mix generated by users of that class:

- E-commerce user: Represents users with high workloads that generate a bursty traffic mix consisting of heavy Web browsing and email. The QoS bounds for these users are [100kbps, 400kbps].
- Researcher: Represents users with slightly higher workloads than the E-commerce class. Their traffic mix is characterized mainly by Web browsing, telnet and email. Their QoS bounds are [200kbps, 500kbps].
- Sales User: Represents best-effort users that generate light web traffic. There is no admission control on their traffic and they get a fair share of the best-effort bandwidth.
- Voice User: Represents users that predominantly generate Voice over IP connections, each with a QoS range of [60kbps, 120kbps]. In addition, voice connections tolerate a maximum delay of 25ms.

Heavy web traffic is characterized by a 50 KB page size (including embedded graphics) and 1 min. inter-arrival times, while light web traffic has 10 KB page sizes and 5 min. inter-arrival times [22]. Heavy email traffic has an average email size of 60 KB and a uniform 1 min. inter-arrival



**Figure 5. The balance index of the network as a function of simulation time.**

time. Voice traffic is modeled using an on/off source with exponentially distributed on and off periods of 350ms and 650ms respectively. Traffic is generated during the on periods at the rate of 60kbps. We obtained these values from Opnet's default application configuration [21].

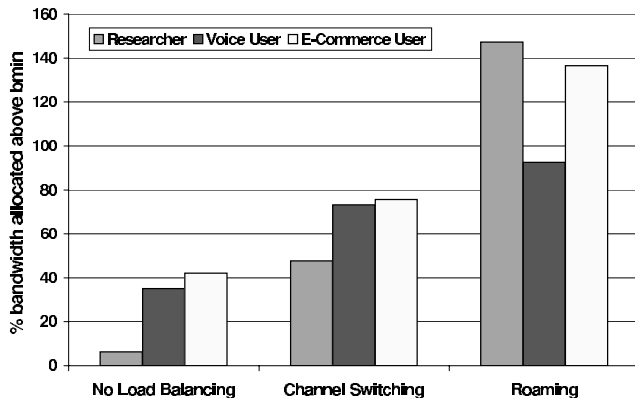
## 5.2. Experimental Scenario

We evaluate the performance of the load-balancing admission control using an illustrative example. We choose a user distribution model to resemble that of a public-area wireless network in two respects: (i) users tend to remain at a particular location once they start to use the network, and, as a result, roaming events are relatively infrequent [7, 14, 27]; and (ii), the user activity in particular cells increases over time, eventually creating hot-spots that require congestion relief.

We pick two of the six cells (1 and 3) to serve as hot-spots. One hundred mobile users are randomly placed in the network, with the constraint that each hot-spot cell gets 30 users each and the remaining cells get 10 users each. User arrivals are simulated by varying the start time of each user's network activity within the cell; these start times follow an independent exponential distribution. Each cell has 10% of its capacity reserved for best-effort connections. Cell 1 has 3 voice users, 10 E-commerce users, 3 researchers, and 14 sales users. Cell 3 has an equal distribution of E-commerce users and sales users.

Figure 4 shows the network utilization in each cell. Note that the utilization of cells 1 and 3 is much higher than the rest of the cells due to the higher concentration of users. The goal of the load balancing algorithms is to reduce this imbalance.

Figure 5 shows the balance index as a function of simulation time. The curves show the effect of the load-balancing algorithms on the overall network utilization. The balance

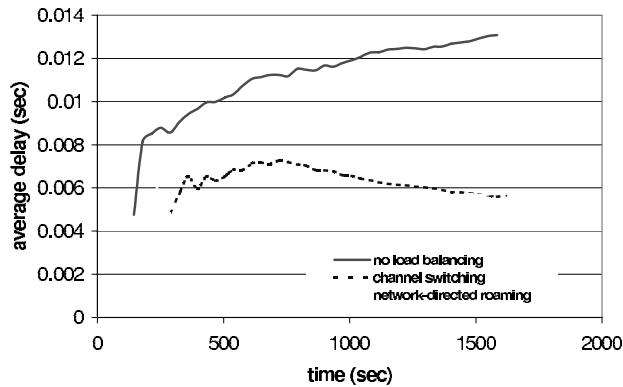


**Figure 6. Percentage bandwidth allocation above  $min$  as a result of network-wide admission control for each class of users.**

index is 0.55 without load balancing and improves to 0.69 with channel switching and reaching 0.85 with network-directed roaming. This is not surprising because, in the case of channel switching, the network adapts by redistributing load only among those cells whose APs are within range of the hot-spot. With network-directed roaming, however, the load can be spread across the entire network and achieve greater balance and correspondingly higher utilization. We assume here that users agree to roam to the cell indicated. In the case of roaming, the plots shown depict the system after it reaches steady state and do not include the transient period while users physically roam to the new cell.

To demonstrate the effect of performing network-wide admission control and load balancing on the QoS perceived by individual connections, we measure (i) the percentage of bandwidth above  $min$  allocated to each user (Figure 6), and (ii) the average round-trip delay perceived by voice connections (Figure 7). It can be seen in Figure 6 that the overall bandwidth allocation increases with the reallocation of users through channel switching and roaming. In particular, researchers and e-commerce users see a marked increase in allocated bandwidth as a result of channel switching and roaming as compared to voice users. This is because voice users have a lower  $min$  requirement (60 kbps) compared to researchers and e-commerce users. The greater QoS requirement of researchers and e-commerce users is better met with network-directed roaming, resulting in gains of over 100%. Figure 7 compares the average round-trip delay of voice connections over time with and without load balancing. From the figure, we see that that, in the absence of any load balancing, the delay of voice calls reaches 14ms. Channel switching reduces the delay to 6.5ms, and network-directed roaming drops the delay even further to 5 ms. Both techniques switch voice users from cell 1 to neighboring





**Figure 7. The average round-trip delay of voice calls over time with and without load balancing.**

less congested cells.

With channel switching, 8 users are switched to 3 neighboring cells. With network-directed roaming 12 users are distributed across all cells; each user moves once for a distance of about 40m. We note that when both channel switching and roaming are used together, the balance index increases to about 0.93 as this achieves the greatest degree of user redistribution in the network. Overall, our algorithms improve the degree of load balance by over 30%, and user bandwidth allocation by up to 52% as compared to the base case.

### 5.3. Performance Implications

We have seen how performing cell-based admission control aids in guaranteeing a minimum level of QoS to users, and in preventing excessive user congestion at specific locations within the network. However, any benefit derived from these algorithms comes at a certain cost to the user and the network. To quantify the trade-offs involved, we answer the following questions:

- Are there any system costs/overheads (memory, signaling traffic) involved that offset the benefit derived from these algorithms?
- What are the implications of running these algorithms on network capacity planning?
- How do the load-balancing algorithms affect the startup latency perceived and relative utility gained by users accessing the network?

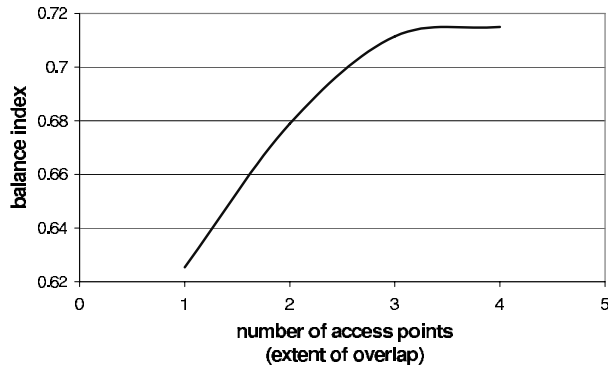
**5.3.1. System and Network Costs.** The load-balancing algorithms require the network to maintain additional state

for input to the algorithms. Since we advocate a centralized architecture with an admission control server (ACS), the ACS is the ideal location to maintain and update this state information. The algorithms require two kinds of state information at the ACS: (i) per-AP state that includes the total load handled by the APs in each cell, and (ii) per-user state that includes the user's access key, session duration, and QoS token.

The ACS maintains information about each cell in its AP state table. Part of the information in the AP state table (e.g. AP name, IP/hardware address, and location) is static. In addition, the state table also maintains dynamically varying information like the number of associated mobile nodes in every cell and the aggregate throughput at each AP. This information needs to be updated at regular intervals of few tens of seconds. The ACS sends periodic requests to the APs in the network in order to keep the dynamic state updated. This amounts to a total of 40 bytes of state per-AP of which only 8 bytes is dynamically varying. As a result, the per-AP state information does not impose heavy demands on the system both in terms of storage and communication overhead due to update traffic.

Per-user state includes the user-specific access key and the QoS token issued to the user as a result of admission control. Again, per-user state needs to be updated regularly as users terminate or renegotiate their service with the network. A key challenge in updating user-state information in a centralized architecture lies in being able to detect if a user is still associated to a certain AP. Since there is no explicit disconnect operation in a wireless network, the ACS has to rely on some higher-layer mechanism to keep track of the user state changes. We have mentioned that the network periodically broadcasts service beacons in the network to make users aware of the existence of an access service. These beacons can also help to poll the clients at regular intervals; a response from a particular mobile host will indicate that the user is still accessing the network. Responding to network beacons at a pre-determined periodic interval (once every 30s) keeps the amount of signaling traffic in the system minimal.

**5.3.2. Network Capacity Planning.** One of the requirements of the channel switching algorithm is the existence of overlapping coverage provided by APs in the network. To provide overlapping coverage, network administrators need to effectively plan the positioning of APs so as to avoid *dead spots* within the network. There are two main disadvantages to over-provisioning wireless coverage in the network. First, having more APs implies higher installation and maintenance costs. Second, with a limited set of orthogonal RF channels (three, as per the IEEE 802.11 standard in the US) a network with densely packed APs may result in two neighboring cells being configured on the same chan-

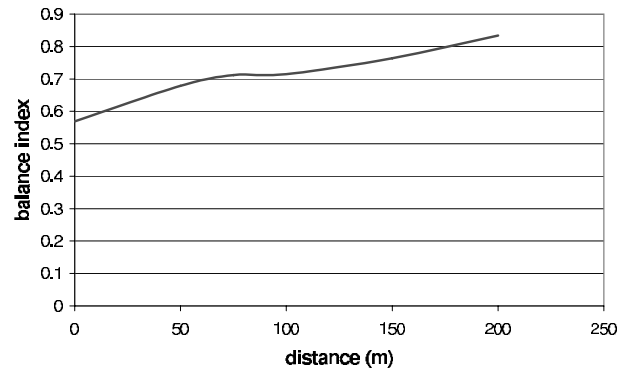


**Figure 8. The effect of extent of coverage on the performance of explicit channel switching.**

nel, thereby degrading the effective system capacity. This directly impacts our goal of providing higher capacity to handle load.

To investigate the effect of overlapping coverage on the performance of the load balancing algorithms, we ran the simulation described in the previous subsection under multiple network configurations. In each configuration, the coverage in the network was improved by adding an extra AP. Figure 8 shows the results, and we can see that the balance index improves when the coverage changes from 1 AP to 2 APs and again from 2 APs to 3 APs, and finally reaches an asymptote of 0.72 with 4 APs.

**5.3.3. Startup Latency and Roaming Overhead.** Another aspect of performing explicit QoS negotiation and admission control is the increased startup latency that users could experience due to the additional communication required by the load balancing algorithms. Part of the startup latency comes from user authentication, which is a mandatory cost. The additional overheads are due to: (i) users communicating their QoS bounds to the ACS, (ii) the ACS performing admission control to determine the best location where the user can get the desired service, and (iii) the user switching or roaming to the new cell as necessary. The negotiation of QoS bounds involves a lightweight message exchange (72 bytes) between the user and the ACS and can be coupled with the authentication phase. With this nominal load, we have found that one ACS can handle the traffic from about 10 APs, making our system scalable [5]. The admission control test is not computationally intensive since it merely involves a serial lookup in the AP state table to determine which of the APs (in the list sent as part of the user's SLS) can best provide the desired level of service. That leaves us with determining the costs for the explicit channel switching and network-directed roaming.



**Figure 9. The effect of roaming radius on the performance of network-directed roaming.**

For explicit channel switching, scanning RF channels and switching to the channel with highest signal strength is a straightforward operation that can be accomplished on the order of tens of milliseconds with current wireless LAN hardware [6].

For network-directed roaming, the cost to users is the distance that they have to travel to reach an AP that can accommodate their service request. To explore the tradeoff between roaming distance and the effectiveness of network-directed roaming, we simulated a network where the radius within which users could roam was progressively increased. We also divided the network into two different subnets and restricted the hot spot to one subnet, keeping the second subnet lightly loaded. The results, plotted in Figure 9, show an increase in balance index with roaming radius that begins to level off at 100m. As the roaming radius increases beyond 100m (which corresponds to roaming into the second subnet), the balance index once again increases until it reaches a maximum of 0.83. However, the benefit of this increase and the correspondingly higher user QoS comes at the cost of explicitly roaming to the new cell.

## 6 Conclusions and Future Work

In this paper, we have described and evaluated two new approaches for providing hot-spot congestion relief while maintaining pre-negotiated user bandwidth agreements with the network. These algorithms accommodate more users at pre-negotiated service levels and improve network utilization by making more efficient use of deployed resources. We describe a QoS-negotiation and admission control protocol that enables users to negotiate service levels. Finally, we describe a unified QoS management architecture that provides differentiated last-hop service and monitors the network against unauthorized use of allocated resources.

We evaluate the benefit of the load balancing algorithms and admission control using simulations. The simulation results show that our algorithms perform well in a variety of user configurations. We use a parameter called balance index to evaluate the extent of balance achieved between the cells in the network. Our algorithms improve the balance index by over 30%, and users bandwidth allocation by up to 52%, in comparison to existing schemes that offer little or no load balancing. We analyze in detail the costs involved in implementing our algorithms and show that our algorithms are scalable, and that the benefits derived outweigh user and network overhead. Based upon our results, we conclude that such networks would benefit greatly from the use of these algorithms.

As part of future work, we are investigating the effect of more sophisticated channel error models (like multipath) on our load balancing and admission control algorithms. We are also investigating means to exploit AP power agility and adjust AP transmit power thus dynamically reconfiguring cell boundaries to adjust network load.

## References

- [1] Agere Systems. Firmware Update for ORINOCO PC Cards v7.28 - Spring 2001 release, April 2001.
- [2] Aironet Wireless Communications Inc. Developer's Reference Manual: PC4500/PC4800 PC Card Wireless LAN Adapter, 1999.
- [3] A. Ayyagiri, Y. Bennet, and T. Moore. IEEE 802.11 Quality of Service, February 2000.
- [4] Y. Azar, A. Z. Broder, and A. R. Karlin. On-line load balancing. In *Proc. 33<sup>rd</sup> IEEE Annual Symposium on Foundations of Computer Science*, pages 218–225, October 1992.
- [5] P. Bahl, A. Balachandran, and S. Venkatachary. Secure Wireless Internet Access in Public Places. In *Proc. IEEE ICC'01*, pages 3271–3275, June 2001.
- [6] P. Bahl, V. N. Padmanabhan, and A. Balachandran. A Software System for Locating Mobile Users: Design, Evaluation and Lessons. Technical Report MSR-TR-2000-12, Microsoft Research, February 2000.
- [7] 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 SIGMETRICS'02*, June 2002 (To Appear).
- [8] 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.
- [9] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and E. Weiss. An Architecture for Differentiated Services. *IETF RFC 2475*, December 1998.
- [10] 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*, 17(1):1–14, June 1989.
- [11] Cisco Systems Inc. Data Sheet for Cisco Aironet 350 Series Access Points, June 2001.
- [12] P. Hsiao, A. Hwang, H. T. Kung, and D. Vlah. Load-Balancing Routing for Wireless Access Networks. In *Proc. IEEE Infocom'01*, pages 986–995, April 2001.
- [13] IEEE. 802.11b/d3.0 Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, August 1999.
- [14] D. Kotz and K. Essien. Characterizing Usage of a Campus-wide Wireless Network. Technical Report TR2002-423, Dartmouth College, March 2002.
- [15] J. Li, J. Jannotti, D. S. J. De Couto, D. R. Karger, and R. Morris. A Scalable Location Service for Geographic Ad Hoc Routing. In *Proc. ACM MobiCom'00*, pages 120–130, August 2000.
- [16] Q. Li and D. Rus. Sending Messages to Mobile Users in Disconnected Ad-hoc Wireless Networks. In *Proc. ACM MobiCom'00*, pages 44–55, August 2000.
- [17] 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*, 16(6):922–936, August 1998.
- [18] S. Lu, V. Bhargavan, and R. Srikant. Fair Scheduling in Wireless Packet Networks. In *Proc. ACM Sigcomm'97*, pages 63–74, August 1997.
- [19] S. Lu, K. W. Lee, and V. Bhargavan. Adaptive Service in Mobile Computing Environments. In *Proc. IFIP IWQoS'97*, pages 25–36, May 1997.
- [20] E. Ng, I. Stoica, and H. Zhang. Packet Fair Queuing Algorithms for Wireless Networks with Location-Dependent Errors. In *Proc. IEEE Infocom'98*, pages 1103–1111, March 1998.
- [21] Opnet Technologies Inc. Wireless LAN Model Description, November 2000.
- [22] J. E. Pitkow. Summary of WWW Characterizations. *Journal of Computer Networks and ISDN Systems*, 30(1-7):551–558, April 1998.
- [23] B. N. Schilit and M. Theimer. Disseminating Active Map Information to Mobile Hosts. *IEEE Network*, 8(5):22–32, September 1994.
- [24] 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*, 40(2):207–217, February 1992.
- [25] H. Syed, G. Kenward, P. Calhoun, and M. Nakhjiri. General requirements for a context transfer framework. *IETF Draft*, January 2002.
- [26] D. Tang and M. Baker. Analysis of a Metropolitan-Area Wireless Network. In *Proc. ACM MobiCom'99*, pages 13–23, August 1999.
- [27] D. Tang and M. Baker. Analysis of a Local-Area Wireless Network. In *Proc. ACM MobiCom'00*, pages 1–10, August 2000.
- [28] N. H. Vaidya, P. Bahl, and S. Gupta. Distributed Fair Scheduling in a Wireless LAN. In *Proc. ACM MobiCom'00*, pages 167–178, August 2000.
- [29] M. A. Visser and M. E. Zarki. Voice and Data Transmission over an 802.11 Wireless Network. In *Proc. PIMRC'95*, pages 648–652, September 1995.