Towards All-IP Wireless Networks: Architectures and Resource Management Mechanism

Majid Ghaderi and Raouf Boutaba

School of Computer Science

University of Waterloo

Waterloo, ON N2L 3G1, Canada

Tel: +519 885 5412

Fax: +519 885 1208

{mghaderi,rboutaba}@uwaterloo.ca

Abstract

Future wireless Internet will consist of different wireless technologies that should operate together in an efficient way to provide seamless connectivity to mobile users. The integration of different networks and technologies is a challenging problem mainly because of the heterogeneity in access technologies, network architectures, protocols and service demands. First, this paper discusses three alternative architectures for an all-IP network integrating different wireless technologies using IP and its associated service models. The first architecture, called ISB, is based on a combination of DiffServ and IntServ models appropriate for low-bandwidth 3G cellular networks with significant resource management capabilities. The second architecture, called DSB, is purely based on the DiffServ model targeted for high-bandwidth wireless LANs with little resource management capabilities. The last architecture, called AIP, combines ISB and DSB architectures to facilitate the integration of wireless LAN and 3G cellular networks towards a uniform architecture for all-IP wireless networks. Second, this paper proposes a flexible hierarchical resource management mechanism for the proposed all-IP architecture which aims at providing connection-level quality of service for mobile users. Simulation results show that the proposed mechanism satisfies the hard constraint on connection dropping probability while maintaining a high bandwidth utilization.

Towards All-IP Wireless Networks: Architectures and Resource Management Mechanism

I. Introduction

Today's wireless networking involves various wireless technologies and networks that serve mobile users across the globe. Existing wireless systems, ranging from wireless local area networks to wide area cellular systems to satellite-based communications, are not compatible with each other, making it difficult for a user to roam from one radio system to another. There is no single commonly agreed upon universal solution for wireless communications. Various technological, commercial and political interests involved in standardization processes prevent from creating such a universal solution. As a result, a variety of wireless networks co-exist and can sometimes complement each other. Therefore, integrating these networks and technologies will empower mobile users to be always connected to the most appropriate network using the most appropriate technology that suits their needs.

Unfortunately, the integration of heterogeneous networks and technologies is a challenging problem mainly because of the following issues [1]:

- **Access technologies:** Different networks apply different radio technologies for the air interface.
- **Network architectures and protocols:** Different networks have different architectures and protocols for transport and routing, resource and mobility management.
- **Service demands:** Mobile users demand different services with different resource and quality of service requirements.

To cope with these heterogeneities a common interconnection protocol which makes no assumptions about the characteristics of the underlying technologies is required. The Internet protocol (IP) provides a universal network-layer protocol for wireline packet networks, and is viewed as an attractive candidate to play the same role in wireless systems. An all-IP wireless network, i.e. IP-based wireless access and fixed core, could make wireless networks more robust,

scalable, and cost effective [2]. It will also enable the applications and software technologies developed for wired IP networks to be used over wireless networks. An IP-enabled mobile device supporting multiple air interfaces could roam seamlessly among different wireless systems if IP is adopted as the common network layer protocol.

One important component of any network architecture is resource management. Typically, wireless networks struggle with limited radio resources which intensifies the need for efficient resource management. The goal of efficient resource management is to achieve maximum radio resource utilization while providing a desired level of quality of service (QoS) to users. In this paper, we investigate the design of efficient resource management techniques that take advantage of IP-based technologies to achieve global roaming in heterogeneous networks. The integration and interoperation of heterogeneous resource management mechanisms is of paramount importance for seamless roaming.

A large research effort has been dedicated to adding resource management capabilities into wired IP networks. As a result a number of proposals have been made. The IETF has adopted two architectures for providing end-to-end quality of service in IP networks: Integrated Services (IntServ) [3] and Differentiated Services (DiffServ) [4]. These architectures differ significantly in terms of router behavior. IntServ operates on a per-flow basis and maintains individual states for all accepted flows, which raises a scalability issue. In turn, DiffServ merges individual flows into fewer aggregates and is hence more scalable. Each aggregate is associated with a particular forwarding behavior known as the Per Hop Behavior (PHB) [4]. These PHBs are local and their link-by-link connection results in end-to-end quality of service.

IntServ and DiffServ have been designed for wired IP networks. They must be extended in order to operate in wireless networks. Several attempts have been made to adapt IntServ and DiffServ for use in wireless networks. Assuming that user mobility is predictable so that the set of cells a mobile user is expected to visit during the life time of the connection[1] can be determined, Talukdar et al. [5] extended the IntServ architecture to support seamless mobility and quality of service. This is an unrealistic assumption and is not feasible in practice. Their approach requires resource reservations to be made along the paths to other locations the mobile user may visit. An extended version of RSVP [6] protocol, called mobile RSVP [7], is

[1]In this paper, terms ‘flow’ and ‘connection’ are used interchangeably.

proposed to handle this type of reservation in wireless Integrated Services networks. In a series of papers Mahadevan et al. [8] studied whether DiffServ, as defined for wired networks, is suitable for wireless networks. According to their study, several enhancements including signaling and mobility considerations are needed. This signaling protocol should consider the low bandwidth and mobility characteristics of the wireless network.

Previous wireless IntServ and wireless DiffServ approaches inherit the drawbacks of their respective underlying architectures. Maintaining a per-flow state in every router does not scale and flow aggregation does not allow for quantitative services to be offered to flows. Furthermore, assuming that the exact mobility of a user is known beforehand is unrealistic. Also, reserving bandwidth in all the cells that the mobile user will visit is too conservative and will lead to poor network utilization. Both approaches require significant changes in the network infrastructure either by changing the behavior of the routers inside the network or by introducing new signaling protocols.

Building from our previous work [9], [10], we introduce, in this paper, two architectures *IntServ-Based Architecture* (ISB) and *DiffServ-Based Architecture* (DSB) for all-IP wireless networks applying IntServ and DiffServ service models. In both architectures, the core network is DiffServ-capable. However, in the first architecture (ISB), the wireless access network is based on IntServ while in the second architecture (DSB), the access network is based on DiffServ. The pros and cons of each architecture are analyzed. We also present an architecture called *Integrated All-IP Architecture* (AIP) which combines the advantage of both architectures. For the proposed architecture, a simple yet efficient resource management scheme is proposed and evaluated through simulations. The scheme is based on probabilistic behavior of mobile users and does not require precise knowledge of user mobility. It uses local information to predict the state of the network.

The rest of the paper is organized as follows. Section II presents the proposed architectures for all-IP wireless networks, i.e. ISB, DSB and AIP. Section III describes the proposed resource management mechanism for the proposed architecture. The high-level operation of an admission control algorithm is presented in section IV. Then, section V is dedicated to the analysis of the proposed algorithm. Simulation results are presented in section VI and section VII concludes this paper.

II. All-IP Wireless Network Architectures

A wireless IP network consist of two components: a wireless access network and a fixed core network. There are important issues that should be addressed in wireless IP networks in order to provide a seamless service in both fixed and mobile environments. Perhaps the most challenging issue is resource management and quality of service provisioning. This is even more difficult considering that there is no native resource management or quality of service control function in traditional IP networks. Current wired IP only offers the best effort service model which treats all packets from all users equally. In wireless environments due to specific characteristics of wireless channel, QoS provisioning is even more challenging. Furthermore, each wireless access network, has potentially its own wireless technology and administrative policies which makes it more difficult to have a uniform resource management mechanism.

The goal of this section is to discuss resource management architectures for integrating different wireless networks in order to provide seamless connectivity. The proposed architectures adopt IP as the common network layer protocol. To minimize the amount of change inside the network, we use the existing resource management mechanisms proposed for IP networks as much as possible. In all proposed architectures, the core network is based on the DiffServ model.

A. IntServ-Based Architecture

As mentioned earlier, scalability concerns of the IntServ model in wired Internet led to the development of the DiffServ model which is simpler and more scalable. However, some disadvantages related to the static nature of the DiffServ model have been pointed out such as static nature of the model. Therefore, a combination of the two models has been proposed [11], [12] to develop a dynamic and scalable architecture for wired Internet supporting end-to-end quality of service. In the same time, while QoS provisioning mechanisms based on resource reservation are quite popular in wireless networks, the DiffServ model does not support explicit reservation.

In a wireless environment, resource utilization is particularly important. The static nature of the DiffServ model can degrade the network performance. Fig. 1 depicts our proposed architecture for all-IP wireless networks based on a combination of IntServ and DiffServ models. In this architecture IntServ operates at the wireless access network and DiffServ at the core IP network.

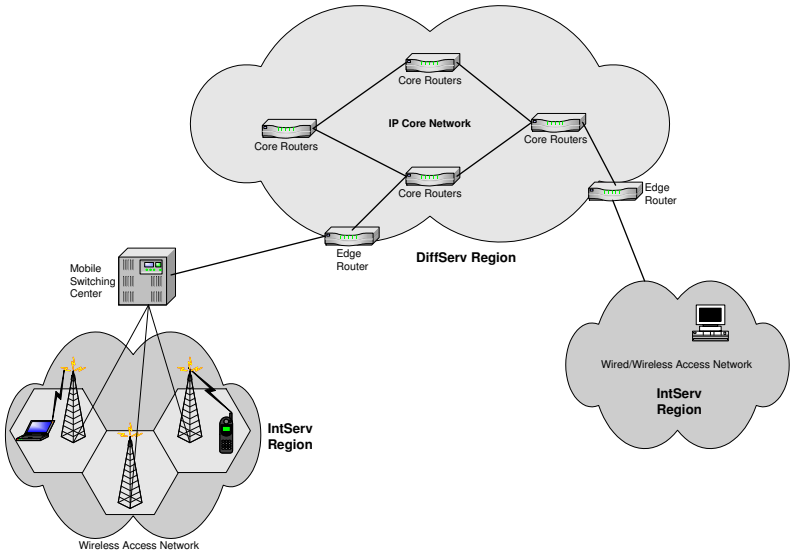

Fig. 1. IntServ-based all-IP architecture.

Since a wireless link can not accommodate a large number of flows as opposed to backbone links, scalability will not be a problem here. By using IntServ and RSVP in the wireless access network, quantitative services can be offered to mobile users. When IntServ is used to provide access to DiffServ network, a critical node in the network is the edge router at the border of network regions. The edge router must implement two interfaces, one for IntServ/RSVP and the other for DiffServ. Once a flow is admitted by the IntServ interface, its traffic is mapped to an appropriate PHB and packets will be marked accordingly. In the simplest case, guaranteed service is mapped to the EF PHB [13] and predictive service is mapped to an appropriate AF class [14].

An important feature of the proposed architecture is that it operates with RSVP without any required change. Therefore, the normal operation of RSVP is sufficient in this combined architecture as addressed by Bernet et al. in [11]. Only the local resource management at the base stations must be changed to support then appropriate scheme. An efficient admission control algorithm for the ISB architecture has been proposed in [9], all the processing involved in the connection setup are the standard IntServ/DiffServ operations except the admission control process at the base stations.

A unified scheduling algorithm is used in [15] for the IntServ region. In this scheme guaranteed

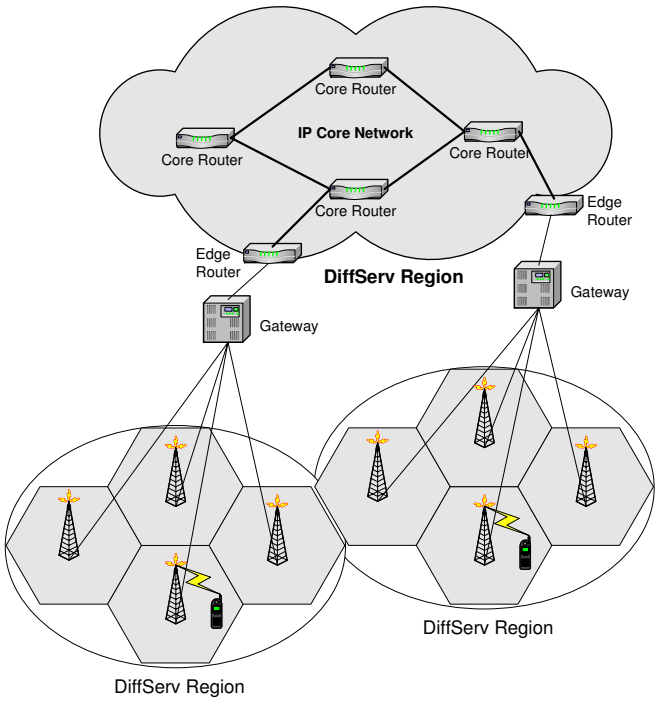

Fig. 2. DiffServ-based all-IP architecture.

service is provided by a weighted fair queuing algorithm [16]. Weighted fair queuing assigns a portion of link capacity to each active flow. The scheduling discipline for predictive service is a priority queue. The flows within each predictive class are scheduled according to a FIFO policy. Best effort flows have the lowest priority in the scheduling. The satisfaction of end-to-end QoS requirements is the responsibility of end systems. An end system could use any QoS routing mechanism to select a route that satisfies its end-to-end requirements.

B. DiffServ-Based Architecture

In DiffServ-Based architecture, not only the core network is DiffServ-capable but also the access networks are DiffServ-capable. Fig. 2 shows the DSB architecture. In this architecture a cellular network overlaid by DiffServ domains operates as the radio access network.

In the previous subsection, we discussed how the static nature of DiffServ can degrade the radio resource utilization in wireless access networks. Therefore, a more fine-grained architecture based on IntServ was proposed. It was also mentioned that because of limited radio resources, the number of flows and consequently the amount of state information required for IntServ/RSVP

operation is quite reasonable with respect to the scalability requirement. These assumptions are reasonable for low-bandwidth systems, e.g. a 3G networks. However, when the access network operates on a high-bandwidth IP-based wireless technology, e.g. a wireless LAN, these assumptions do not stand for the following reasons:

- Typically, such technologies have high capacities in order of several Mbps. Therefore, it is possible to have a large number of flows simultaneously in the network. Considering that future cellular technologies such as 4G will expand the available radio resources to the same orders, then this will be problematic even in those environments.
- Due to the inherent IP-based architecture of these technologies, traffic flows have different characteristics and requirements than those in conventional cellular networks. The applications intended for such environments are delay-tolerant and do not require strict QoS guarantees (web browsing compared to voice calls for instance). Also, their generated traffic is bursty in nature and hence it is difficult to describe their bandwidth requirements accurately a priori. The types of applications supported by conventional cellular networks are limited which facilitate the classification of their requirements. This is not true in wireless LAN environments.
- Mobility patterns are different in WLAN-based hot spot environments compared to those in conventional cellular networks. Hot spot traffic is more chaotic and hence more difficult to predict. As a result, it is not possible in practice to reserve appropriate amount of resources beforehand for each individual connection which may handoff to the hot spot. In contrast, traffic aggregates are usually more smooth and predictable thanks to the law of large numbers. This suggest that class-based resource management is more feasible in wireless environments.
- The wireless environment is rapidly changing. Wireless channel capacity fluctuates over time with interferences. So, it is difficult to achieve strict QoS guarantees similar to those in wireline networks with fairly stable channel quality. In this case coarse grained QoS guarantees like those offered by DiffServ are sufficient and in fact more appropriate for the target application types.

For all above reasons we believe that the DSB architecture is a more appropriate candidate for future all-IP wireless networks than the ISB architecture. At least for wireless LAN envi-

ronments it is a more feasible solution for providing end-to-end quality of service and resource management.

A class-based queueing (CBQ) adapted to wireless environments [17] can be used in this architecture. CBQ associates a hierarchical structure with a link, thus aggregating the packet streams belonging to different connections into classes consisting of one or more connections. CBQ associates quantitative bandwidth commitments with the hierarchical class organization to provide controlled link sharing by ensuring fairness in the sense that each interior and leaf class gets its allocated bandwidth over a relevant time interval.

C. Integrated All-IP Architecture

So far, we have described two candidate architectures, namely, ISB and DSB, for future all-IP wireless networks. It was concluded that the former architecture is more suitable for cellular networks based on 3G technologies while the latter architecture is more appropriate for wireless LAN environments. As discussed earlier, the primary goal of the all-IP paradigm is to integrate different wireless networks including 3G and wireless LANs. In this perspective, a combination of the ISB and DSB architectures seems natural. It is then up to the access network operator to decide which architecture is appropriate for the offered services.

Fig. 3 depicts an instance of the AIP architecture, where wide area coverage is provided by a 3G cellular network and hot spot coverage is provided by several wireless LAN networks. The wireless LANs covering a single hot spot form a DiffServ region where each region is connected to the Internet through a gateway (not represented in the figure). The 3G network, in turn, forms an IntServ region which is also connected to the Internet via its own gateway.

There are several issues that need to be addressed in order to realize such an architecture. Perhaps the most challenging one is the interoperation between ISB-based and DSB-based regions. The same mechanism discussed in subsection II-A can be adopted in this architecture for service mapping between ISB and DSB domains. Since access networks, in this architecture, have potentially different QoS models, i.e. IntServ or DiffServ, resource management is more challenging than in the ISB and DSB architectures. In the next section, we will present a resource management scheme which is suitable for the AIP architecture.

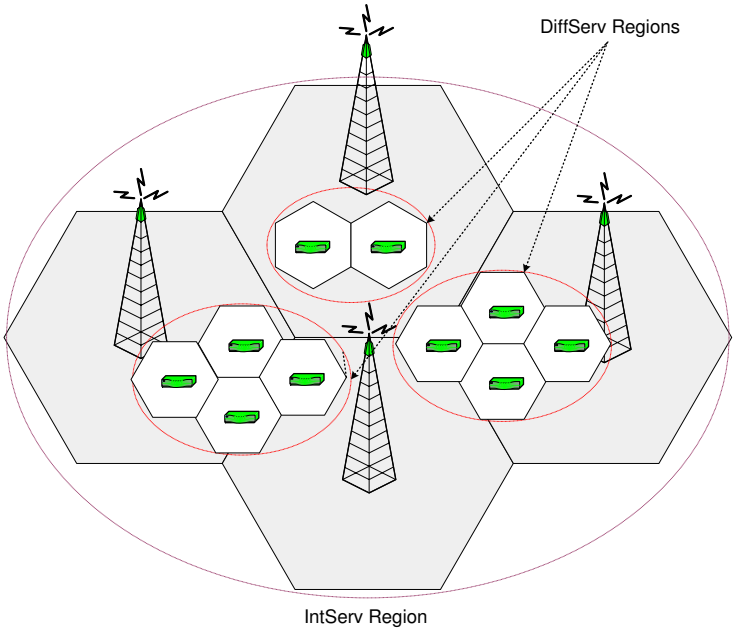

Fig. 3. Integrated all-IP architecture.

III. Resource Management Protocol

In wireless systems, efficient resource management is a critical issue. Perhaps the most important component in resource management in such environments is the connection admission control (CAC). In this section we present a hierarchical admission control mechanism for the AIP architecture. Without the loss of generality, we only consider the DiffServ regions when representing the proposed admission control scheme. Extending the results to IntServ regions is similar and is not presented in this paper.

Users in wireless networks are free to roam in the network. Mobility complicates the resource management problem. At connection-level, two important parameters which specify the quality of service are the *connection blocking probability* (P_b) and the *connection dropping probability* (P_d). To provide seamless mobility the admission control mechanism must be aware of the handoff traffic arriving in the future to reserve the resources required to accommodate this traffic. Otherwise the excessive traffic must be dropped (connection dropping) which negatively affects the quality of service to mobile users. It is usually preferable to block a connection (connection blocking) request at the first place rather than dropping the connection before completion. Although reserving resources for handoffs can prevent handoff dropping, over-reservation will

degrade the radio resource utilization. A fundamental question is how much resources must be reserved to guarantee a target dropping probability? Typically, the goal of a CAC scheme is to maintain a prespecified target connection dropping probability while minimizing the connection blocking probability.

In the AIP architecture there are two different types of handoffs:

1) *Inter-domain handoff:* between different domains; when there are some service level agreements (SLAs) between neighboring domains and there are some service negotiation protocols [18], mobile users can move from one domain to another while keeping their connections alive.
2) *Intra-domain handoff:* in one domain; mobile users can move between neighboring cells inside each domain while receiving the same quality of service.

Recently, Cheng and Zhuang [19] have considered DiffServ resource allocation in a domain-based cellular network. Their work is based on the *cell-cluster* concept proposed by Naghshineh and Acampora [20] where each cell-cluster corresponds to a DiffServ domain. In the cell-cluster approach, cells are grouped into clusters and each cluster is associated with a controller. A threshold is set for the whole cluster, then the cluster controller admits a new connection as long as the number of occupied channels in the cluster is less than the threshold, and the cell where the new connection is generated has a free channel to accept this new connection. After admission to the cluster, no further communication is necessary with the cluster controller for handoffs between cells in that cluster. Cheng and Zhuang extend this basic scheme to include guard channels for each cell in the DiffServ domains, however their proposed scheme is still static in that it reserves a fixed number of guard channels for each cell and domain regardless of the traffic load. This can result in network resource underutilization.

In this section, we propose a *prediction-based admission control* (PrBAC) for the AIP architecture similar to the two-level scheme proposed in [21] and [22]. We extend their scheme to include DiffServ domains as follows:

1) relative priorities between different service classes. The proposed admission control drops low-priority connections to accommodate high-priority handoffs.
2) We consider both intra-domain and inter-domain handoffs when dynamically adjusting the reservation thresholds.

3) Instead of using simple traffic patterns (Poisson arrivals and exponentially distributed connection durations), PrBAC uses a *minimum mean square error* predictor (MMSE) [23] to predict the bandwidth requirements in each cell and in each domain. Because we directly predict bandwidth requirements independent of the underlying traffic characteristics, this scheme is very suitable for IP networks where traffic patterns are not Poisson [24].

Although using traffic prediction for admission control has been used in other papers [25], [26], the novelty of our approach is that the MMSE predictor is on-line and does not rely on a specific traffic model. For example the FARIMA predictor used in [25] is very complex and can not be estimated using on-line traffic measurements. On the other hand, the ARIMA predictor of [26] is simpler but not suitable for *self-similar* Internet traffic [27] prediction as stated in [25] and similar papers. The key idea behind our approach is to predict traffic directly from on-line measurements without involving any traffic modeling. To use a MMSE predictor we do not need to specify any traffic model.

IV. Hierarchical Admission Control

Without loss of generality, we assume that there is a one-to-one correspondence between administrative domains and DiffServ domains. In order to support inter-domain handoffs we use an admission control which is local to the domain, i.e, it does not need any information exchange with neighboring domains. The idea is that regardless of the complexity and overhead associated with distributed CAC schemes [28]–[30], there is currently no standard protocol for exchanging information needed by distributed schemes between neighboring domains.

We extend the local algorithm for inter-domain handoffs to handle intra-domain handoffs as well, which leads to a simple and effective CAC scheme. In the gateway (GW), the bandwidth broker enforces DiffServ constraints while interacting with the CAC component. In the base station (BS), the CAC component makes the admission decision based on the bandwidth requirements of new connections (can be extracted from their SLAs) and handoffs from both neighboring cells and neighboring domains by a prediction method based on the minimum mean square error predictor.

The proposed scheme, PrBAC, is a measurement-based admission control. It is commonly believed that measurement-based admission control is a more feasible and realistic candidate for IP networks [31], [32].

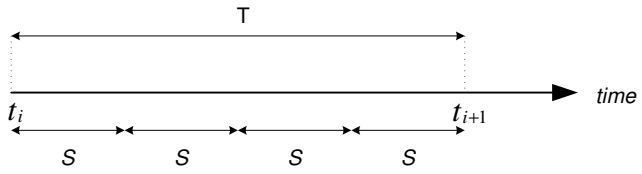

Fig. 4. Sampling mechanism in each control interval.

The proposed scheme, PrBAC, has a periodical control structure. At the beginning of each control interval of length T, each cell predicts the amount of bandwidth required to accept incoming handoffs during the current control interval. Then it reserves this amount of bandwidth to be used exclusively for handoffs until the end of this period. Fig. 4 shows the sampling mechanism used by the control algorithm where each control interval contains sampling points at distance s. The maximal sample taken in each interval is kept as the bandwidth usage for that interval. Below is the notation which will be used throughout the rest of this paper.

- B: the available bandwidth in the cell under consideration
- $B_T^u(t)$: the bandwidth allocated to all connections at time t
- $B_H^u(t)$: the bandwidth allocated to handoffs at time t
- B_T^i: the bandwidth usage during the control interval i
- $\widehat{B}_T^i$: the predicted value of B_T^i
- B_H^i: the bandwidth required for handoffs that will arrive during the control interval i
- $\widehat{B}_H^i$: the predicted value of B_H^i

The PrBAC scheme only takes care of handoffs belonging to expedited forwarding (EF) and assured forwarding (AF) classes. When it is necessary, PrBAC drops best effort (BE) flows in order to accommodate higher priority handoff flows. The only difference between EF and AF treatment is that for AF flows, PrBAC considers only their minimum bandwidth requirements.

A. Minimum Mean Square Error Predictor

To forecast the bandwidth usage for the current control interval, a MMSE predictor of order m is used. Let B denote the random variable to be predicted and $\widehat{B}$ the predicted value of B. A MMSE predictor for B is given by

$$\widehat{B} = \mathbf{WB} + \varepsilon, \tag{1}$$

where ε is the white noise error with mean 0 and variance σ_ε^2, and $\mathbf{B}$ is a vector of size m of the previous observations of B. In this equation, $\mathbf{W}$ is a weighting vector obtained as follows:

$$\mathbf{W} = \mathbf{\Gamma}\mathbf{G}^{-1}, \tag{2}$$

where $\mathbf{G}$ is the autocovariance matrix and $\mathbf{\Gamma}$ is an autocovariance vector starting at lag m,

$$\mathbf{G} = \begin{bmatrix} \rho_0 & \rho_1 & \cdots & \rho_{m-1} \\ \rho_1 & \rho_0 & \cdots & \rho_{m-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{m-1} & \rho_{m-2} & \cdots & \rho_0 \end{bmatrix}, \tag{3}$$

and

$$\mathbf{\Gamma} = [\ \rho_m \quad \ldots \quad \rho_1\ ]. \tag{4}$$

The autocovariance function ρ_k can be computed by

$$\rho_k = \frac{1}{m} \sum_{i=k+1}^{m} \mathbf{B}(i)\mathbf{B}(i-k), \tag{5}$$

where m is the order of the MMSE predictor. And finally, the mean squared error of the MMSE predictor is given by

$$\sigma_\varepsilon^2 = \sigma_B^2 - \mathbf{\Gamma}\mathbf{G}^{-1}\mathbf{\Gamma}'. \tag{6}$$

B. Admission Control at Base Stations

Assume that a new connection request arrives at time $t \in (0, T]$ during the control interval i. Let b denote the amount of bandwidth required by this connection which can be expressed using the effective bandwidth concept [33]. Let $B_H^r(t) = \widehat{B}_H^i - B_H^u(t)$ denote the residual amount of bandwidth that we have predicted to be used by the upcoming handoffs until the end of this interval, i.e., during interval $(t, T]$. Also, let $B_T^f(t) = B - B_T^u(t)$ denote the total amount of free bandwidth at time t. The admission control at a base station follows the pseudo-code presented in Fig. 5.

In this algorithm `N-DiffServ` is the standard DiffServ module which enforces DiffServ requirements at the ingress point to the domain. `H-DiffServ` is the same as `N-DiffServ` except that it may drop BE flows in order to accommodate EF and AF handoffs. Note that B_T^f and B_H^r include bandwidth usage of both EF and AF flows. The tuning parameter $\alpha \geq 1$ is an

```
1: if (handoff connection request) then
2:    if (H-DiffServ accepts) then
3:       grant admission;
4:    else
5:       reject;
6:    end if
7: else /* new connection request */
8:    if (N-DiffServ accepts) & (B_T^f − B_H^r) > αb then
9:       grant admission;
10:   else
11:      reject;
12:   end if
13: end if
```

Fig. 5. Admission control at base stations.

adaptable parameter that can be adjusted based on the difference between measured connection dropping probability and the target P_{QoS}.

The pseudo-code in Fig. 6 describes an additional admission condition which makes the algorithm presented in Fig. 5 more conservative. The additional condition in Fig. 6 is used when the algorithm in Fig. 5 accepts the new connection request. As mentioned earlier, PrBAC uses the maximum amount of bandwidth usage sampled in each control interval to represent the amount of bandwidth required for that interval. If at time t it is found that the bandwidth requirement for the current interval is underestimated, then PrBAC looks ahead at the next control interval. If the predicted bandwidth requirement for the next interval, after accepting this new connection, is greater than the total amount of available bandwidth then the new connection request will be rejected.

```
1: if (B_T^u(t) + b) > B̂_T^i then
2:    if B̂_T^{i+1} < B then
3:       grant admission;
4:    else
5:       reject;
6:    end if
7: end if
```

Fig. 6. The conservative condition.

C. Admission Control at Gateways

If a base station accepts a new connection request then it will send this request to the domain gateway (GW) for second level admission. At this level, the GW takes into account two factors: 1) DiffServ domain constraints; and 2) inter-domain handoffs.

The same algorithm we described for BSs can be applied in the GWs considering a domain as a virtual cell. For such virtual cell, B_H is the bandwidth required for handoffs from neighboring domains (neighboring virtual cells) and B is the total bandwidth available in the domain. While it is possible to use the predicted values from the domain boundary cells at this level of the PrBAC, a direct prediction is preferred. This method has the advantage of less communication overhead and more accurate predictions due to aggregation. The more traffic is aggregated and smoothed, the more accurate prediction is possible.

Each BS will contact its corresponding GW only for new connections and handoffs from other domains. After admission, no more communication with the GW is required for intra-domain handoffs. This reduced communication leads to fast handoff processing which is necessary to prevent QoS degradation at upper network layers (e.g., a delayed handoff process increases the packet loss and delay at network layer).

V. Connection Dropping Probability

The accuracy of PrBAC is determined by the accuracy of the predictor. For example, if MMSE could predict the exact bandwidth requirements, then PrBAC could guarantee zero percent connection dropping while achieving the optimal connection blocking probability. This is not possible in practice.

During the life of a connection, a mobile user may cross several cell boundaries and hence may require several successful handoffs. Failure to get a successful handoff at any cell in the path forces the network to drop the connection. While the handoff failure probability, P_f, is an important parameter for network management, the probability of dropping a connection, P_d, may be more relevant to the user and service provider. Nevertheless, connection dropping probability is a system dependent parameter which is particularly affected by user mobility. Let H denote the number of handoffs during the life of a connection, then $P_d = 1-(1-P_f)^H$ where H itself is a random variable that depends on several system parameters such as mobile velocity and cell

size. In particular, the average probability of connection dropping is given by [34]

$$P_d = \frac{hP_f}{(\mu + hP_f)}, \tag{7}$$

where μ and h denote the average connection completion and average handoff rate. Consequently, for any given target connection dropping probability P_{QoS}, a target handoff failure probability P^*_{QoS} can be computed as

$$P^*_{\mathrm{QoS}} = \frac{P_{\mathrm{QoS}}}{1 - P_{\mathrm{QoS}}} \left(\frac{\mu}{h}\right) . \tag{8}$$

In the worst case, a handoff will fail when the predicted value $\widehat{B}_H$ is less than the actual value B_H (it is possible to accept a handoff even in this situation due to the residual free bandwidth). Therefore, this is an upper bound for handoff failure. That is

$$\Pr\{\text{Handoff Failure}\} \leq \Pr\{\widehat{B}_H < B_H\}, \tag{9}$$

or, equivalently, to satisfy the target handoff failure probability P^*_{QoS}, it is obtained that

$$\Pr\{\widehat{B}_H < B_H\} \leq P^*_{\mathrm{QoS}} . \tag{10}$$

To guarantee that $\widehat{B}_H > B_H$, we compute an upper confidence interval δ for the predicted value $\widehat{B}_H$ as follows

$$\Pr\{(B_H - \widehat{B}_H) > \delta\} \leq P^*_{\mathrm{QoS}}, \tag{11}$$

therefore,

$$\Pr\{\varepsilon > \delta\} \leq P^*_{\mathrm{QoS}} . \tag{12}$$

We know that ε is white noise with normal distribution $\mathcal{N}(0, \sigma^2_\varepsilon)$. Therefore,

$$\delta = \sigma_\varepsilon \Phi^{-1}(1 - P^*_{\mathrm{QoS}}), \tag{13}$$

where $\Phi(.)$ is the integral over the tail of a normal distribution which can be expressed in terms of the error function [35].

VI. Simulation Results

For the sake of simplicity there is only one traffic class with fixed bandwidth requirements in the simulated system. This basic implementation is sufficient to show the performance of the PrBAC scheme compared to the traditional trunk reservation scheme. We have also implemented the scheme proposed in [19] which we refer to as the *cell-domain admission control* (CDAC) scheme.

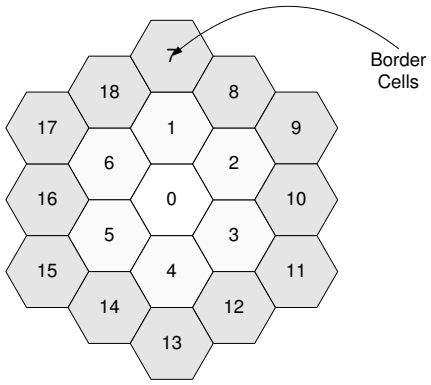

Fig. 7. 2D cellular system.

A. Simulation Parameters

Simulations were performed on a two-dimensional cellular system consisting of 19 hexagonal cells (see Fig. 7). Opposite sides wrap-around to eliminate the finite size effect. Each domain has 19 cells, each cell has 20 bandwidth units available and each new connection requires one bandwidth unit. For the sake of simplicity, we have assumed that 12 cells out of 19 cells are bordering cells and 50% of their handoff traffic is due to the inter-domain handoffs.

To predict handoffs, MMSE(10) which is an MMSE predictor with history of size 10 is used. Connection durations and cell residency times are exponentially distributed with mean 20 and 5 units of time respectively. We also extended the basic CDAC scheme to support inter-domain handoffs. This extended version treats inter-domain handoffs similar to PrBAC. The target connection dropping is set to $P_{\mathrm{QoS}} = 10^{-2}$ and the reservation thresholds for CDAC are 10% and 20% at cell-level and domain-level respectively.

B. MMSE Predictor Evaluation

The proposed PrBAC is intended for all-IP networks carrying traffic with self-similar characteristic. To evaluate the accuracy of MMSE, the MMSE predictor is compared with several self-similar predictors including fGn [23], FARIMA [23] and GARMA [36] for IP traffic. In this experiment we used an Ethernet traffic trace (`pAug89.TL`[2]) from Bellcore which is collected by Leland et al. [37]. This trace has information on the time-stamp and the packet size of traffic.

[2]Accessible at `http://ita.ee.lbl.gov`

TABLE I

THE ACCURACY OF PREDICTORS.

Predictor	SNR^{-1}
MMSE	0.27
fGn	0.32
FARIMA	0.22
GARMA	0.23

The data they collected is cumulative. To get a time series, we need a uniform time scale. We extracted the traffic data at 10^{-2} millisecond intervals. We dedicated the first 2×10^3 samples of this trace to estimate parameters of the fractional models (actually we used the reported parameters in [38] for FARIMA and [36] for GARMA predictors) and then we implemented and used predictors to forecast 20×10^3 samples into the future.

Although we have used traffic from wired Ethernet, these results should remain valid for any traffic with the same degree of self-similarity (the so-called Hurst parameter for this traffic trace is 0.8). Considering that future cellular networks will be able to carry IP traffic, it seems reasonable to have the same traffic characteristics for wired and wireless IP traffic. For example, Jiang et al. [39] showed that cellular digital packet data traffic exhibits long-range dependencies.

Finally, the reverse of *signal to noise ratio* (SNR^{-1}) defined as

$$\mathrm{SNR}^{-1} = \frac{\sum \varepsilon^2}{\sum B^2}, \tag{14}$$

is used as the accuracy measure to compare these predictors. The smaller the SNR^{-1}, the more accurate is the predictor. Table I summarizes the results of this comparison. In particular, it shows that the accuracy of MMSE is reasonably close to that of the best predictor (FARIMA).

C. Results and Analysis

Let $\rho = \lambda/\mu$ denote the arrival load in each cell, where λ is the new connection arrival rate and $1/\mu$ is the mean call duration. Simulations were done for a wide range of loads from 20 to 1000 Erlang load per cell. For each load, simulations were done by averaging over 4 samples, each for 10^5 new connection requests. In addition to connection blocking/dropping probability

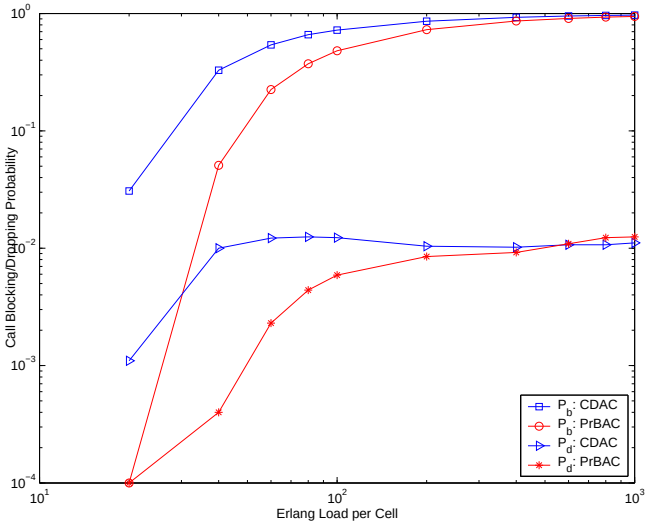

Fig. 8. Connection blocking/dropping with Poisson traffic.

as the QoS measures, connection completion probability, P_c, is also computed as the effective measure for network utilization. Connection completion probability is given by

$$P_c = (1 - P_b)(1 - P_d)\,. \tag{15}$$

Intuitively, P_c shows the percentage of new connection requests that were successfully completed.

Two scenarios were considered for the simulations. In the first scenario, Poisson generated traffic was used to evaluate the performance of CAC schemes. Poisson process is bursty and not a good candidate for linear prediction based on MMSE. In the second scenario, an autoregressive model was used to generate the arrival traffic. In particular, AR(1) model was used for the connections interarrival times. It is clear that AR(1) is more predictable than Poisson using our MMSE predictor due to its linear structure. Therefore, we expect to see a better performance in terms of handoff dropping and bandwidth utilization for the second scenario, i.e. non-Poisson traffic. The primary goal of considering these two scenarios is to show that PrBAC performance is not affected by the arrival traffic model in contrast to the existing CAC schemes that are designed only for Poisson traffic.

1) Poisson Traffic: Figs. 8 and 9 show the QoS and utilization measures for Poisson generated traffic, where inter-arrival times for new and handoff connections are exponentially distributed. Furthermore, each cell of the system experiences the same rate of new arrivals and handoffs.

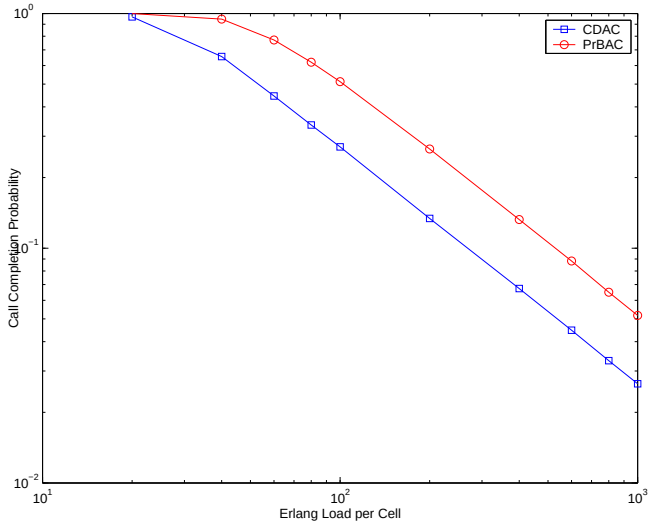

Fig. 9. Connection completion with Poisson traffic.

Both schemes can provide a limit for connection dropping probability while at the same time the connection blocking probability of PrBAC is lower than CDAC. This difference in the connection blocking probability can be explained with respect to the dynamic nature of PrBAC. CDAC is static and can not adapt to the changing traffic therefore results in degraded bandwidth utilization. In contrast, PrBAC is dynamic and tune its admission threshold every control period based on the measured traffic samples. However, as shown in Fig. 8, both schemes can provide a guaranteed dropping probability for a wide range of traffic loads. Fig. 9 shows that call completion probability under PrBAC scheme is always higher than CDAC. Also, by increasing the load, the call completion probability decreases.

2) Non-Poisson Traffic: As mentioned earlier, the performance of PrBAC is determined by the accuracy of MMSE. Although Poisson generated traffic is not a good test case for MMSE predictor, PrBAC performs better than static CDAC due to its dynamic nature. It is interesting to see the performance of both schemes under a different traffic pattern where traffic is more predictable than Poisson traffic.

Figs. 10 and 11 show the connection blocking/dropping and connection completion probability for non-Poisson generated traffic, where the inter-arrival times for new and handoff connections are derived from an autoregressive model of order one, AR(1), with coefficients 0.5 and 0.8. An

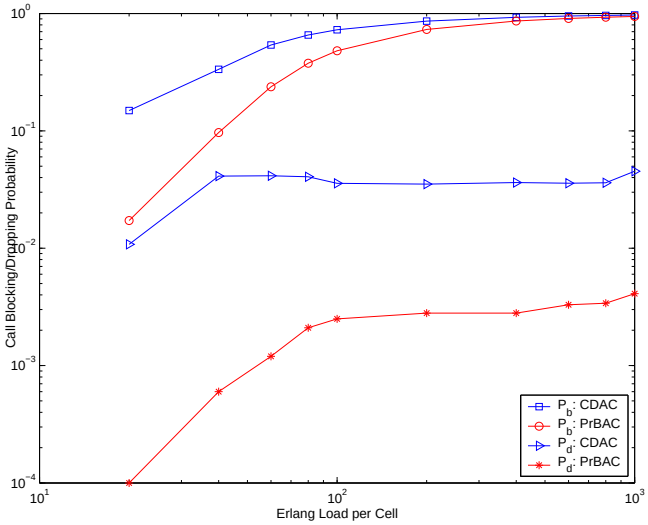

Fig. 10. Connection blocking/dropping with non-Poisson traffic.

AR(1) model with coefficient θ is defined by

$$(X_n - 1/\lambda) = \theta\,(X_{n-1} - 1/\lambda) + Z \tag{16}$$

where X is the stochastic process defined by the model showing the time to next arrival. Also, λ is the arrival rate and Z is the deriving normal variable [23]. In (16), X_n represents the interarrival time for connection n as a combination of interarrival time for connection $n-1$ and a normal variable with zero mean and unit variance.

Fig. 10 shows that connection blocking/dropping curves of PrBAC are below the ones for CDAC. However, as it can be seen, PrBAC achieves a dropping probability which is an order of magnitude less than the prespecified target dropping probability. More investigation is required to make PrBAC reactive to such discrepancies in order to achieve better performance results. Fig. 11, on the other hand, represents the completion probability for PrBAC and CDAC. As for Poisson generated traffic, PrBAC outperforms CDAC. Despite the fact that connection dropping probability of PrBAC is far less than that of CDAC in non-Poisson scenario, the final connection completion probabilities are not much different in both scenarios. The reason is that dropping probabilities in both cases are too small, i.e. close to zero, and hence do not have much influence on the completion probability. In other words, connection blocking probability is the major factor in final connection completion probability which is almost the same in both scenarios.

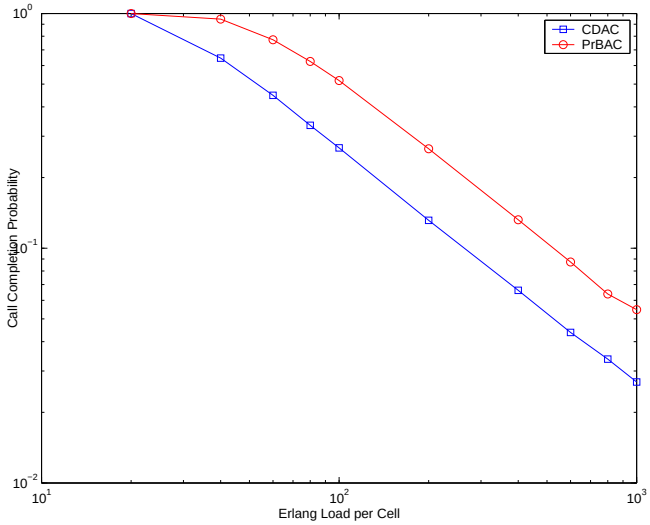

Fig. 11. Connection completion with non-Poisson traffic.

Although real traffic patterns are more complicated than any of the ones used, since the performance of PrBAC is better than CDAC in each tested case we can deduce that PrBAC will perform better than CDAC when presented with real traffic patterns. Of the tested cases the one using the simple AR(1) model shows the greatest performance difference between PrBAC and CDAC.

VII. Conclusion

In this paper, we studied the application of IP and its associated QoS service models for integrating wireless networks ranging from 3G cellular networks to wireless LAN hot spots. Two architectures, namely, ISB and DSB, were proposed for all-IP wireless networks applying IntServ and DiffServ respectively. In both architectures, the core network is DiffServ-capable but in the first proposal (ISB), the wireless access network is based on IntServ while in the second one (DSB) the access network is based on DiffServ too. We analyzed the advantages and disadvantages of each architecture and proposed AIP, an architecture which combines advantages of both architectures. We then proposed a hierarchical admission control scheme based on forecasting future handoff traffic for the AIP architecture. The key idea is to use online measurements to predict traffic directly without relying on any particular traffic model. Our scheme was validated using simulations for different types of traffic.

The results of this work can be used to build a simple and efficient admission control scheme for wireless IP networks, where traffic diversity prohibits conventional traffic modeling. However, there are several issues that require further research for a complete realization of the AIP architecture:

- *Adaptive applications:* Adaptive applications are critical for the success of integrating heterogeneous networks. First, wireless network transport capacity fluctuates over time, hence, applications must be able to adapt to the changing environment such as capacity degradations. Second, different wireless networks have different transport capabilities. A bandwidth-intensive application which hands off from a hot spot to a 3G network must be able to shrink its bandwidth requirement according to the new environment capabilities.
- *Service negotiation:* Adaptive applications and services require a standard protocol to communicate their resource demands when roaming among different wireless networks. Such service negotiation protocol should not incur heavy signaling overhead on the wireless network. At the same time it must be able to handle the service mapping between different resource management models, i.e. the IntServ vs. DiffServ models.
- *Vertical handoff:* In this paper we addressed hard vertical handoff where different networks operates in different regions. When there is significant overlapping, e.g. 3G coverage in a hot spot area, soft handoff schemes may be more appropriate. Deciding which network is more appropriate to handoff to is an important issue when there are more than one wireless network accessible in the coverage area. Such decision can be affected by the service charge, end-to-end delay and resource availability in overlapping networks.

In addition to the above, other issues such as security, pricing and billing must be addressed as well.

REFERENCES

[1] I. F. Akyildiz, J. Xie, and S. Mohanty, "A survey of mobility management in next-generation all-IP-based wireless systems," *IEEE Wireless Communications Magazine*, vol. 11, no. 4, pp. 16–28, Aug. 2004.

[2] P. Agrawal, T. Zhang, C. J. Sreenan, and J.-C. Chen, "All-IP wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 2, no. 4, pp. 613–616, May 2004.

[3] B. Braden *et al.*, "Integrated services in the Internet architecture: An overview," RFC 1633, IETF, June 1994.

[4] S. Blake *et al.*, "An architecture for differentiated services," RFC 2475, IETF, Dec. 1998.

[5] A. K. Talukdar, B. Badrinath, and A. Acharya, "Integrated services packet networks with mobile hosts: Architecture and performance," *ACM/Kluwer Wireless Networks*, vol. 5, no. 2, pp. 111–124, 1999.

[6] B. Braden *et al.*, "Resource reservation protocol (RSVP): Version 1 functional specification," RFC 2205, IETF, Sept. 1997.

[7] A. K. Talukdar, B. Badrinath, and A. Acharya, "MRSVP: A resource reservation protocol for an integrated services with mobile hosts," *ACM/Kluwer Wireless Networks*, vol. 7, pp. 5–19, 2001.

[8] I. Mahadevan and K. M. Sivalingam, "Architecture and experimental framework for supporting QoS in wireless networks using differentiated services," *Mobile Networks and Applications*, vol. 6, pp. 385–395, 2001.

[9] Y. Iraqi, M. Ghaderi, and R. Boutaba, "Enabling real-time all-IP wireless networks," in *Proc. IEEE WCNC'04*, vol. 6, Atlanta, USA, Mar. 2004, pp. 1500–1505.

[10] M. Ghaderi, J. Capka, and R. Boutaba, "Prediction-based admission control for DiffServ wireless Internet," in *Proc. IEEE VTC'03*, vol. 3, Orlando, USA, Oct. 2003, pp. 1974–1978.

[11] Y. Bernet *et al.*, "A framework for integrated services operation over DiffServ networks," RFC 2998, IETF, Nov. 2000.

[12] J. Harju and P. Kivimaki, "Co-operation and comparison of DiffServ and IntServ: Performance measurements," in *Proc. IEEE LCN 2000*, Tampa, USA, Nov. 2000, pp. 177–186.

[13] B. Budiardjo, B. Nazief, and D. Hartanto, "Integrated sservices to differentiated services packet forwarding: Guaranteed service to expedited forwarding PHB," in *Proc. IEEE LCN 2000*, Tampa, USA, Nov. 2000, pp. 324–325.

[14] T. Chahed, C. Fayet, and G. Hebute, "On mapping of QoS between integrated services and differentiated services," in *Proc. IEEE/IFIP IWQoS'00*, Pittsburgh, USA, June 2000, pp. 173–175.

[15] D. D. Clark, S. Shenker, and L. Zhang, "Supporting real-time applications in an integrated services packet network: Architecture and mechanism," in *Proc. ACM SIGCOMM'92*, Baltimore, USA, Oct. 1992, pp. 14–26.

[16] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single-node case," *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 344–357, June 1993.

[17] C. Fragouli, V. Sivaraman, and M. B. Srivastava, "Controlled multimedia wireless link sharing via enhanced class-based queuing with channel state-dependent packet scheduling," in *Proc. IEEE INFOCOM'98*, vol. 2, San Francisco, USA, Apr. 1998, pp. 572–580.

[18] J.-C. Chen *et al.*, "Dynamic service negotiation protocol (DSNP) and wireless DiffServ," in *Proc. IEEE ICC'02*, vol. 2, New York, USA, April 2002, pp. 1033–1038.

[19] Y. Cheng and W. Zhuang, "DiffServ resource allocation for fast handoff in wireless mobile Internet," *IEEE Communications Magazine*, vol. 40, no. 5, pp. 130–136, May 2002.

[20] M. Naghshineh and A. S. Acampora, "Design and control of micro-cellular networks with QoS provisioning for real-time traffic," in *Proc. IEEE Universal Personal Communications*, San Diego, USA, Sept. 1994, pp. 376–381.

[21] J.-H. Lee, K. Lee, M.-R. Choi, and S.-H. Kim, "Two-level threshold for RCAC (region-based call admission control) in multimedia wireless networks," in *Proc. IEEE International Conference on Networks*, Oct. 2001, pp. 208–213.

[22] H.-C. Lin and S.-S. Tzeng, "Double-threshold admission control in cluster-based micro/picocellular wireless networks," in *Proc. IEEE VTC'00*, vol. 2, Tokyo, Japan, May 2000, pp. 1440–1444.

[23] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed. New York, USA: Springer-Verlag, 1991.

[24] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226–244, June 1995.

[25] Y. Shu, Z. Jin, J. Wang, and O. W. Yang, "Prediction-based admission control using FARIMA models," in *Proc. IEEE ICC'00*, vol. 3, New Orleans, USA, June 2000, pp. 1325–1329.

[26] T. Zhang, E. Berg, J. Chennikara, P. Agrawal, J. C. Chen, and T. Kodama, "Local predictive resource reservation for

handoff in multimedia wireless IP networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 10, pp. 1931–1941, Oct. 2001.

[27] M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: Evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, Dec. 1997.

[28] M. Naghshineh and M. Schwartz, "Distributed call admission control in mobile/wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 4, pp. 711–717, May 1996.

[29] B. M. Epstein and M. Schwartz, "Predictive QoS-based admission control for multiclass traffic in cellular wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 523–534, Mar. 2000.

[30] S. Wu, K. Y. M. Wong, and B. Li, "A dynamic call admission policy with precision QoS guarantee using stochastic control for mobile wireless networks," *IEEE/ACM Transactions on Networking*, vol. 10, no. 2, pp. 257–271, Apr. 2002.

[31] S. Jamin, P. B. Danzig, S. J. Shenker, and L. Zhang, "A measurement-based admission control algorithm for integrated services packet networks," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 524–540, Feb. 1997.

[32] A. Jamalipour and J. Kim, "Measurement-based admission control scheme with priority and service classes for application in wireless IP networks," *J. Communication Systems*, vol. 16, no. 6, pp. 535–551, May 2003.

[33] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1091–1100, 1995.

[34] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Transactions on Vehicular Technology*, vol. 35, no. 3, pp. 77–92, Aug. 1986, see also: CEAS Tech. Rep. No. 773, College of Engineering and Applied Sciences, State University of New York, June 1999.

[35] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge, UK: Cambridge University Press, 1992.

[36] R. Ramachandran and V. R. Bhethanabotla, "Generalized autoregressive moving average modeling of the Bellcore data," in *Proc. IEEE LCN'00*, Tampa, USA, Nov. 2000, pp. 654–661.

[37] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 71–86, Feb. 1997.

[38] Y. Shu, Z. Jin, L. Zhang, and L. Wang, "Traffic prediction using FARIMA models," in *Proc. IEEE ICC'99*, vol. 2, Vancouver, Canada, June 1999, pp. 891–895.

[39] M. Jiang, M. Nikolic, S. Hardly, and L. Trajkovic, "Impact of self-similarity on wireless data network performance," in *Proc. IEEE ICC'01*, Helsinki, Finland, June 2001, pp. 477–481.