Predictive Call Admission Control for All-IP Wireless and Mobile Networks

Kelvin L. Dias, Stênio F. L. Fernandes and Djamel F. H. Sadok
Computer Science Center, Federal University of Pernambuco
CP 7851, Cidade Universitária, Recife-PE 50732-970, Brazil
Phone: +55 81 3271.8430 Fax: +55 81 3271.8438
{kld,sflf,jamel}@cin.ufpe.br

ABSTRACT
This paper proposes a novel call admission control (CAC) scheme for wireless and mobile networks. Our proposal avoids per-user reservation signaling overhead and takes into account the expected bandwidth to be used by calls handed off from neighboring cells based only on local information stored into the current cell where user is seeking admission. To this end, we propose the use of two time series-based models for predicting handoff load: the Trigg and Leach (TL), which is an adaptive exponential smoothing technique, and ARIMA (Autoregressive Integrated Moving Average) that uses the Box & Jenkins methodology. These methods are executed locally by each base-station or access router and forecast how much bandwidth should be reserved on a periodic time window basis. The two prediction methods are compared through simulations in terms of new call blocking probability and handoff dropping probability. Despite the TL method simplicity, it can achieve similar levels of call blocking probability and handoff dropping probability than those of the computational demanding ARIMA models. In addition, depending on the schemes settings, the prediction methods can grant an upper bound on handoff dropping probability even under very high load scenarios.

Categories and Subject Descriptors
C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design, Wireless Communication.

General Terms
Algorithms, Performance, Design.

Keywords
Call Admission Control, All-IP Wireless and Mobile Networks, Quality of Service, Scalability, Time Series Analysis

1. INTRODUCTION
All-IP wireless and mobile networks represent the convergence of two key technologies: Internet and wireless cellular systems. The combination of both technologies suggests that a coming trend will be an increasing demand for IP based wireless/mobile access to traditional and multimedia applications with varying quality of service (QoS) requirements. Figure 1 illustrates an envisioned scenario with heterogeneous wireless technologies integrated through IP mobility aware protocols (Mobile IPv4/IPv6, Cellular IP, Hawaii, etc.) that will seamlessly interwork with the global Internet [13], [14].

The research effort is especially challenging when dealing with provisioning of quality of service (QoS) guarantees. Users applications may experience performance degradation due to the properties of wireless channels and due to user mobility from handoffs. Handoff in wireless/mobile networks is the mechanism that transfers an ongoing call from the current cell as the mobile station (MS) moves through the coverage area of the system. If the target cell does not have sufficient available bandwidth, the call will be dropped. From the user's point of view handoff dropping is less desirable than the blocking of a new call.

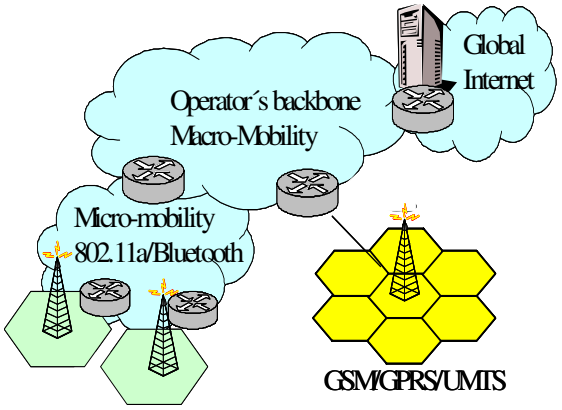

Figure 1. A scenario for all-IP mobile and wireless networks.

An important component for mobile/wireless networks is the Call Admission Control (CAC) mechanism. It must be used to address the mobility effects, accepting or rejecting new users in the network. CAC schemes not only have to ensure that the

network meets the QoS of newly arriving calls if accepted, but should also guarantee that QoS of existing calls does not deteriorate.

On the other hand, Internet frameworks for QoS provisioning rely, basically, on two architectures: Integrated Services (IntServ) [15] and Differentiated Services (DiffServ) [16]. While the IntServ architecture provides strict QoS guarantees through per-user explicit signaling for CAC and reservation using RSVP (Resource Reservation Protocol), it fails in providing the scalability objectives due its reservation-based approach. The DiffServ proposal aims at providing less strict QoS guarantees through packet classification at network ingress and differentiation of the treatment according to a set of classes named PHB (Per Hop Behavior), hence offering better network scalability. The Bandwidth Broker (BB) is a network entity proposed for implementing resource management policies in the DiffServ architecture, including the CAC mechanism [17].

In wireless and mobile networks, reservation of resources is more challenging than in wired networks due the scarcity of bandwidth in wireless links. In our opinion, a scalable QoS architecture for wireless/mobile networks should provide CAC schemes that avoid excessive per-user signaling for wireless link reservation purposes.

This paper proposes a novel call admission control (CAC) scheme for wireless and mobile networks. Our proposal avoid per-user reservation signaling overhead and take into account the expected bandwidth to be used by calls handed off from neighboring cells based only on local information stored into the current cell where user is seeking admission. To this end, we propose the use of two time series-based models for predicting handoff load: the Trigg and Leach (TL), that is an adaptive exponential smoothing technique[9], and the Autoregressive Integrated Moving Average (ARIMA) in conjunction with the Box & Jenkins methodology [10][11]. These models indicate how much bandwidth should be reserved on a periodic time window basis. The two proposals are compared through simulations in terms of new Call Blocking Probability (CBP), Handoff Dropping Probability (HDP) and Bandwidth Utilization. Furthermore, an analysis regarding the quality of the predictions depicts that the time window prediction interval should be set carefully to avoid overestimation and so the waste of the scarce wireless bandwidth. Despite the TL method simplicity, it can achieve similar levels of call blocking probability and handoff dropping probability than those of the computational demanding ARIMA models. In addition, depending on the schemes settings, the prediction methods can grant an upper bound on handoff dropping probability even under very high load scenarios.

The remainder of this paper is organized as follows. In section 2, we describe the related research work. Section 3 gives an overview of the Trigg and Leach and ARIMA techniques for forecasting. We then present the novel CAC scheme in section 4. Performance results are presented in section 5. Finally, concluding remarks are given in section 6.

2. RELATED WORK

Proposals for CAC in wireless/mobile networks present in the literature can be divided into two categories: fixed and dynamic strategies. Fixed strategies, such as the guard channel (GC) [1] scheme, give preferential treatment to handoff calls reserving a fixed number of channels exclusively for them. The advantage of this strategy is its simplicity because there is no need for the exchange of control information between base stations. However, this scheme is not flexible to handle varying traffic loads, since there is no information about current and neighboring cell's load.

Proposed dynamic reservation strategies [2], [3], [4], [5], [6], [7] extend the basic guard channel scheme according to the estimated handoff call rate derived from the number of calls in the neighboring cells and the mobility pattern of these calls to reserve bandwidth in advance in the next cell or in a group of cells. Resource reservation can be problematic, in general, due to the possibility of poor network utilization due to unnecessary blocking of new users and can get even worse if the reservation are made in several adjacent cells. Furthermore, these schemes imply a large amount of signaling overhead.

The scheme proposed in [2] uses the aggregate history of handoffs in each cell to predict the probability a call will be handed off to a certain neighboring cell. Based on handoff prediction, the number of channels is reserved in advance. Each base station records the number of handoff failures and adjusts the reservation by changing the estimation window size. One problem with history-based schemes is the overhead to develop, store and update traffic histories for the different cells. Furthermore, due to short-term changes (e.g., diversion of traffic due to accidents) and medium-term changes (e.g., traffic re-routing during road constructions), these estimates cannot be fully reliable.

The call admission control proposed in [3] takes into consideration the number of calls in adjacent cells, in addition to the number of calls in the admission cell. The authors developed a theoretical model to compute the requirements for handoff requests in order to maintain a target handoff dropping probability. The proposed model assumes that all bandwidth requests are identical, which is not valid if multimedia services with varying bandwidth requirements are to be supported by the network.

Next, we will describe some existing research that aims at optimizing bandwidth utilization (decreasing call blocking probability), but keeping low levels of dropping probability for handoffs.

In [4] a predictive channel reservation (PCR) scheme based on mobile positioning systems (GPS -Global Positioning System) is proposed. This scheme makes predictive channel reservation for each MS based on its current position and orientation. The reservation is triggered if the MS reaches a certain threshold distance from the next cell. A reservation may be deemed invalid (false reservation) if the MS changes its moving direction. In this case, the cancellation of the reservation must be sent to de-allocate the reserved channel. Furthermore, rather than strictly mapping each reserved bandwidth portion to the MS that made the reservation, all reserved bandwidth is used as a generic pool to serve handoff requests but not new calls. When a MS arrives from a neighboring cell after a handoff, it may use bandwidth from the reserved portion if there is any available. Otherwise, the handoff connection will compete in the free bandwidth portion with other new call attempts. The HPCR (Hybrid PCR) scheme is a PCR variant, which integrates the threshold distance with GC, reserving a very small fixed portion of the bandwidth for

handoffs. It was shown in [4], that this hybrid approach improves the handoff dropping probability without jeopardizing the bandwidth utilization.

The ACR (Adaptive Channel Reservation) scheme was proposed in [5] and it is based on the PCR proposal, but it uses a threshold time instead of a threshold distance to trigger the bandwidth reservation in the next predicted cell. The authors argue that using a threshold time permits a better control of the different degrees of mobility to trigger the reservation in the next cell, avoiding waste of bandwidth due to unused reservations. For example, considering a MS located in the overlapping area of two adjacent cells with a very slow moving speed of this MS (close to 0) and requiring a channel for its call. If the PCR scheme is used, two channels (each cell has one channel occupied) will be occupied by this call, one channel is used for communication in the current cell and the other is reserved for this call in the adjacent cell because the threshold distance was reached. Since the MS of this call is almost stationary, the reserved channel may not be used for the lifetime of this call. Consequently, PCR can lead to under-utilization of wireless channels.

The PCR as well as the ACR schemes introduce a lot of signaling messages for reservation and cancellation of false reservations. Moreover, the reservations can decrease the dropping probability at the expense of increasing the blocking probability, what may give rise to poor network utilization. The use of GPS for predicting user mobility is also advocated in proposals [6], [7]. While such dynamic reservation-based schemes have demonstrated significant performance advantages over well engineered guard channels, the per-user dynamic reservation approach place computation and communication burdens on the network's infrastructure which increases with the numbers of users and handoffs. Hence, the scalability and applicability of such solutions to future micro and pico-cellular networks is not well established.

A similar approach to ours is proposed in [8]. The authors proposed a local predictive resource reservation for handoff based on the Wiener process (a Markov process where only the present value is relevant for predicting the future) and a methodology for granting an upper bound on HDP. To grant an upper bound on HDP, the amount of resource that must be reserved for future handoff demands should be set to the upper limit of the confidence interval for the predicted handoff load. In addition, the authors also use an ARIMA prediction method and show that the Wiener prediction obtained quite similar results for predicting the handoff demand based on traces collected from a single cell simulation scenario. The limited results obtained for the CBP and HDP metrics were depicted only for the Wiener-based proposal. The lack of performance results in terms of CBP and HDP for their ARIMA-based prediction seems to be justified by the very similar results obtained from the comparative trace analysis with the Wiener-based method conducted in that paper. As it will be shown in our paper, applying the methodology suggested in [8] could lead to bandwidth overestimation for handoffs. Furthermore, our ARIMA-based proposal differs significantly from that in [8] because we did not adopt the upper limit of the predicted handoff confidence interval to reserve bandwidth. Instead, we suggest directly the use of the predicted value by choosing an appropriate prediction time window size to avoid unnecessary reservations of the scarce wireless bandwidth.

3. FORECASTING PROCEDURES

In this section we present a short description of the forecasting procedures used to evaluate the traffic load arriving at each cell.

A time series can be defined as a realization of a stochastic process. Time series may enfold features such as trends and seasonality and one of the purposes of its analysis is the generation of forecast of future values. This procedure normally requires that time series present some kind of regularity in its behavior. Usually, future values are predicted based on past values, because a steadiness is assumed. This regularity in time series can be expressed through the concept of stationary time series[10]. Therefore, forecasting techniques are based on the idea that future can be predicted by discovering specific patterns of events in the past. Using time series modeling and analysis to predict bandwidth requirement in a computer network environment has lately become a useful and widespread tool. Researchers in the networking field are increasingly adopting modeling techniques widely used by econometricians and statisticians [12].

3.1 Exponential Smoothing and Variants

Exponential smoothing techniques have long been the methods of choice for univariate forecasting due to its accuracy and ease of use. They have become increasingly accepted because of their effortlessness and overall performance. It is highly recommended for short-term prediction. Among the simplest methods is the ordinary (simple) exponential smoothing, which assumes no trend and no seasonality whereas Trigg and Leach procedure could be seen as its adaptive approach.

3.1.1 Simple Exponential Smoothing

Let Y_t denote a univariate time series. Simple exponential smoothing assumes that the forecast $\hat{Y}$ for period $t+h$ is given by a variable level $\hat{a}$ at period t

$$Y_{t+h} = \hat{a}_t,$$

which is recursively estimated by a weighted average of the observed and the predicted value for Y_t.

$$\hat{a}_t = \alpha Y_t + (1-\alpha)\hat{Y}_t,$$

$$\hat{a}_t = \alpha Y_t + (1-\alpha)\hat{a}_{t-1}$$

where $0 < \alpha < 1$ is known as the smoothing parameter (constant). The main drawback of this technique is the choice of the smoothing parameter since setting it close to 1 could give rise to a highly reactive model. On the contrary, choosing the smoothing constant close to 0 could lead to an insensitive model.

3.1.2 Adaptive Exponential Smoothing: Trigg and Leach

In order to assist the selection of α and to improve awareness capability of the predictor, a number of adaptive methods have been recommended in the literature. The most representative and

widely used is the Trigg and Leach [9] technique. Its mainly advantage rely on the fact that there is no need to specify the smoothing parameter previously. Trigg and Leach procedure can regulate the smoothing constant α whenever a change occurs in the time series basic structure. Let α_{t+1} be the one-step ahead smoothing parameter. So, the prediction in $t+1$ for the level is

$$\hat{a}_{t+1} = \alpha_t Y_t + (1-\alpha_t)\hat{Y}_t,$$

$$\alpha_{t+1} = \frac{E_t}{M_t}$$

where

$$E_t = \beta\varepsilon_t + (1-\beta)E_{t-1},$$

$$M_t = \beta|\varepsilon_t| + (1-\beta)M_{t-1}$$

and $\varepsilon_t = Y_t - \hat{Y}_t$ (prediction error at t).

Values close to zero point out a well-controlled prediction system (smaller prediction errors) whereas values near to the unity indicate an out of control prediction system (huge prediction errors). It is important to emphasize that α_{t+1} allow the system to reconcile by not being too reactive to changes. But most importantly, α_t will vary based on variations in the data pattern.

3.1.2.1 Trigg and Leach Upper Confidence Bound

In order to offer statistical guarantees regarding the worst-case handoff dropping probability (HDP) for the next time interval, we may use predicted value as the upper confidence bounds for that predicted value as suggested in [8] for the Wiener process. For example, if the network operator has to guarantee a maximum target handoff dropping probability of 5%, the reserved bandwidth ψ will be set to the 95% upper confidence bounds of the forecasted bandwidth requirements for handoff calls $E(\Omega)$. This way, we can determine a level L such that $Prob(\Omega \le L) = 1 - HDP$. This level L is called $(1-HDP)*100\%$ upper confidence bound for Ω. This value is given by:

$$\psi = E(\Omega) + Z\gamma\sqrt{\left(\frac{\alpha}{2-\alpha}\sigma^2\right)},$$

where $Z\gamma$ is the q-quartile of the standard Normal distribution of $N(0,1)$, α is the smoothing parameter, and σ^2 the sample variance.

3.2 ARIMA Models and the Box & Jenkins Methodology

There are some classical approaches for modeling stationary time series. Models for stationary processes are the Autoregressive (AR), the Moving Average (MA) and the Autoregressive Moving Average (ARMA). Taking a time series $\{X_t\}$, which is stationary and with nonseasonal patterns, if it follows an autoregressive process of order p, denoted by $X_t \sim AR(p)$, then $\{X_t\}$ is given by

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varepsilon_t, \quad (1)$$

where c, ϕ_1, ϕ_2, ..., ϕ_p are unknown parameters, the ϕ_i being called autoregressive parameters, and ε_t is a white noise process [10]. The term Moving Average comes from the fact that $\{X_t\}$ is built from a weighted sum, similar to an average, of the most recent values of ε, and then it can be expressed as

$$X_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q}. \quad (2)$$

μ, θ_1, ..., θ_q are unknown parameters, the θ_i being called moving average parameters, and ε_t is a white noise. If X_t follows a moving average process of order q, it is denoted by $X_t \sim MA(q)$. It is possible to build models that pursue simultaneously autoregressive and moving average expressions. One example is a time series $\{X_t\}$ that follows an autoregressive process with moving average terms, denoted $X_t \sim ARMA(p,q)$, given by

$$X_t = c + \phi_p X_{t-p} + \varepsilon_t + \theta_q\varepsilon_{t-q}, \quad (3)$$

where c, ϕ_i and θ_i are unknown parameters, the ϕ_i being the autoregressive parameters and the θ_i being the moving average parameters. This is an autoregressive moving average process of order (p,q).

It is possible that the traffic load presents some non-stationary patterns, which induces the use of classical approaches for modeling them, such as the Autoregressive Integrated-Moving Average (ARIMA) and the Seasonal Autoregressive Integrated-Moving Average (SARIMA). Another approach is to use some kind of conversion in order to make it stationary. For example, one can take differences, logarithms or squared roots of the observations. A traditional procedure is to use a class of transformations called the Box-Cox transformation [10].

Particularly, processes that, after the application of a finite number d of differences, reduce to ARMA models are called

$ARIMA(p,d,q)$ models. The application of difference to the time series is a method to transform a non-stationary time series to a stationary one. An $ARIMA(p,d,q)$ model can be represented by

$$\Delta^d X_t = \mu + \phi_p \Delta^d X_{t-p} + \varepsilon_t + \theta_q \varepsilon_{t-q}, \quad (4)$$

where the ϕ_i are the autoregressive parameters, the θ_i are the moving average parameters and Δ^d indicates that the order of differentiation is d.

Taking a close look at the equation 3 it is necessary to find a way to estimate the values of $\theta \equiv (c, \phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q)$, known as the vector of population parameters, on the basis of observations on $\{X_t\}$. A usual inference technique on which estimation could be based is Maximum Likelihood (ML). Given the sample of size T, the first step is to calculate the likelihood function (LF), $L(\theta; x)$. This function can be found by calculating a probability density $f_{X_T, X_{T-1}, \cdots, X_1}(x_T, x_{T-1}, \cdots, x_1; \theta)$ that is strong related to the assumption that the particular distribution for the white noise process ε_t assumes a Gaussian white noise form, i.e., $\varepsilon_t \sim i.i.dN(0, \sigma^2)$. So, the maximum likelihood estimate of θ is the value for which this sample is most likely to have been observed, that is $\theta = \arg\max L(\theta; x)$, $\theta \in \Theta \subset \Re$. It is a common sense to use the reduced and conditional log-likelihood $l(\theta; x) \propto \ln L(\theta; x)$, where the LF has the form $L(\theta; x) = f_{X_1}(x_1; \theta) \prod_{t=1}^{T} f_{X_T | X_{T-1}}(x_t \mid x_{t-1}; \theta)$. For example, it is easy to show that the conditional log-likelihood function for a Gaussian ARIMA(p,0,q) process is

$$l(\theta; x) = \frac{-T}{2} \log(2\pi) - \frac{T}{2} \log(2\sigma^2) - \sum \frac{\varepsilon_t^2}{2\sigma^2}, \quad (5)$$

where

$$\varepsilon_t = x_t - c - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p} - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$$

.

An alternative solution for (5) could be performed by solving the system of equations given by $\nabla l(\hat{\theta}) = 0$, usually referred to as likelihood equations. In both cases, there is no closed-form or explicit solution and therefore numerical maximization must be used. The idea would be to make a number of distinct guesses for θ, and try to infer the value of $\hat{\theta}$ for which $l(\theta; x)$ is largest. There are several algorithms for numerical maximization or optimization procedures. For instance, on could use Grid Search, Steepest Ascent, Newton-Raphson, Davidon-Fletcher-Powell or Broyden-Fletcher-Goldfarb-Shanon (BFGS) methods. In this work, the BFGS algorithm was used.

Needless to say that is indispensable a formal procedure to estimate the best model given a number of observations. This leads to a discussion of stochastic model building where Box & Jenkins methodology is widely used to discover models from the series, estimate their parameters and then evaluate the adequacy of the model's fit to the experimental data. The Box & Jenkins methodology tries to provide a flexible procedure so that one may obtain high-quality and suitable models. The methodology consists of three basic stages: Identification, Parameter Estimation and Diagnostic Checking. We refer the reader to Harvey [11] for a more complete explanation related to the Box & Jenkins procedure.

In the first stage, a tentative model is normally selected based on the sample autocorrelation function or the correlogram, which tries to identify the p and q orders for the ARIMA process. Given a time series, the first stage may recommend a number of specifications (i.e., p and q orders), each of which satisfies some diagnostics checks. For that reason, some kind of measure of goodness of fit is required to decide on the best models presented. There are a number of model selection criteria where the decision rule is to select the model that minimizes some variable. The Akaike Information Criterion (AIC) has the following form:

$$AIC = -2 \log L(\theta; x) + 2n, \quad (6)$$

where n is the number of parameters (n=p+q). One should notice that AIC has a predisposition to pick models that are over-parameterized. Another possibility is the Bayes Information Criterion (BIC) where the 2n in equation 6 is replaced by nlogT. In this work we use both criteria. Hence, we decided to offer a portfolio of models and selected the best one based on the smallest AIC (BIC). The ML estimation previously explained was performed during the second stage.

In this work, we automate the B&J methodology in order to identify, estimate and perform the diagnostic check to the handoff load on every cell on a cellular network. We used a sample time interval of 30 or 60s and collected the first 30 samples (called the training period) before starting the automated B&J procedure. This quantity is a sufficient amount of samples to achieve convergence to the ML estimative. After the training period, for each new handoff load measured in each cell, during a chosen sample time interval, we performed the whole B&J procedure all over again.

4. THE PROPOSED CAC

Our novel CAC estimates the total amount of required bandwidth for future handoff calls using TL or ARIMA. The process for predicting the required bandwidth for handoff calls is local, that is, the base station uses only local information (collected bandwidth due to handoffs) that serves as the input for the prediction method without exchange of messages among neighboring cells to this end. Suppose that a base station knows the amount Ω of required resources for handoff calls at the current time t. The amount of resources required for handoffs $E(\Omega)$ at a future time $t + \Delta t$ can be predicted based on the

current Ω and its predicted value from the previous time interval $t - \Delta t$.

The novel CAC should determine whether the admission cell has sufficient bandwidth to support the user requirements and takes into account the predicted handoff load for that cell. Let ψ (the reserved bandwidth) be the upper confidence bound for the expected bandwidth due to handoff calls $E(\Omega)$ for the next prediction interval. The reserved bandwidth can also be the actual forecasted value from a chosen time series model. The following condition must be met:

$$\sum_{i=1}^{N} Bi + B + \psi \leq C$$

This equation verifies whether the admission cell has sufficient bandwidth to support the new request. *N* is the number of existing connections, *C* is the wireless link capacity and *Bi* is the bandwidth being used by the ith connection in that cell. *B* is the bandwidth required by the newly requested connection. At the start of each interval, a new ψ is used to control the admission decision. Upon each handoff arrival in a cell, during a prediction interval, the current ψ is decreased by the MS's bandwidth that has arrived until it reaches 0 or a new prediction interval is initiated

5. PERFORMANCE RESULTS

The simulated model consists of a cellular network with 19 hexagonal cells as depicted in Figure 2. In order to avoid the border effects, when a MS moves out the system this MS will be wrapped around to re-enter the system from the other side. Such a toroidal arrangement is an efficient way to approximately simulate very large systems [3], [5]. In this paper, the unit of bandwidth is called bandwidth unit (BU), which is assumed to be the required bandwidth to support a voice connection as in [2], [7]. Each cell is assumed to have a fixed link capacity of 100BUs. The traffic model used is similar to the one used in [2], [7]. Call requests are generated according to Poisson distribution with rate λ (call/cell/second) in each cell. The simulated traffic consists of users with bandwidth requirements of 1 BU (voice) and 4 BUs (video) with probabilities Rvo and 1-Rvo, respectively, where Rvo is also called the voice ratio as in [2]. In our simulations Rvo is set to 0.7, that is, 70% of voice traffic and 30% of video traffic. The lifetime of each call is exponentially distributed with mean 180s [4], [5], [7].

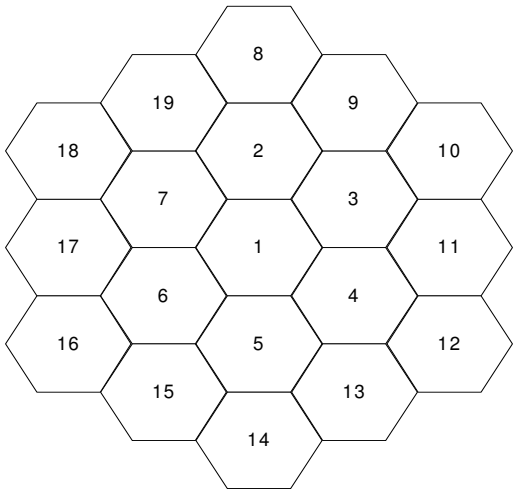

Figure 2. Simulated Cellular Topology.

Upon each new call request or handoff call, the user chooses a moving direction among six probable target cells. At any time, while crossing a cell the MS can change its moving direction with probability equal to 50%. If a MS changes its moving direction, a new target cell is randomly selected (uniformly distributed) as well as a new residence time is chosen. The time that a call spends in a cell prior to handoff to another cell (residence time) is exponentially distributed with mean 60s.

5.1 Simulation Results for the Metrics of Interest

The metrics of interest in this paper are: (1) handoff dropping probability (HDP) defined as the ratio of the number of handoff calls dropped to the total number of handoff call attempts; (2) call blocking probability (CBP), that is, the ratio of the number of new calls blocked by the network to the number of new call requests; and (3) bandwidth utilization.

The models are labeled "M-B-T" in graphs, where M represents the model adopted for prediction (TL or ARIMA), B is the reserved bandwidth type which may be based either on the predicted value (Pred) or on the upper confidence bound (CI) for that predicted value. T is prediction interval (30 or 60s).

Figure 3 and Figure 4 depict the CBP and HDP comparison for the models using a prediction interval of 30s. Both models achieved similar levels of CBP. The HDP comparison shows that TL achieved a slightly greater HDP than ARIMA's in higher loads. This scenario indicates that simplistic TL method can achieve satisfactory levels of prediction as compared to the ARIMA model.

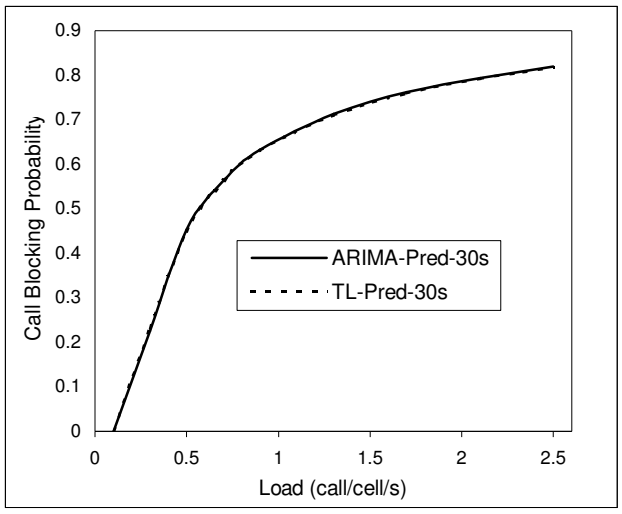

Figure 3 CBP - Prediction Interval: 30s; Reserved Bandwidth Type: Predicted Value.

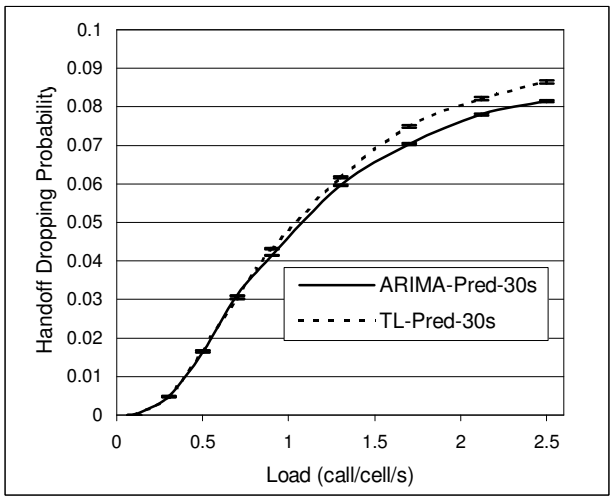

Figure 4 HDP - Prediction Interval: 30s; Reserved Bandwidth Type: Predicted Value.

In order to evaluate the proposal of using the upper confidence bound for the predicted value as the amount of bandwidth that should be reserved on each cell to guarantee a maximum target HDP during the cell overload, Figure 5 and Figure 6 show the results considering the upper confidence bound for a 95% confidence level (CI). Hence, it is expected that the worst case HDP be inferior to 5%. As can be seen, the ARIMA's HDP is better than TL's (Figure 6). However, TL's HDP was kept below the maximum target HDP of 5%. Moreover, the smallest HDP for ARIMA is achieved at the expense of a greater CBP, which generated more blockings of new calls than TL's (Figure 5) and, consequently, providing bandwidth under-utilization as it is depicted in Figure 7, where the bandwidth utilization for TL outperforms the one for ARIMA.

In order to verify our argument that the approach of using the upper confidence bound for the predicted load may cause overestimation for bandwidth reservations, Figure 8 and Figure 9 depict comparisons between the 95% upper confidence bounds for the forecasted value and the actual handoff demand collected during the simulations for TL and ARIMA, respectively. These graphs were based on traces obtained from the same simulation as that of Figure 5 and Figure 6, considering only the load 1.7 (call/cell/second). They refer to bandwidth due to handoffs into the central cell (cell 1 depicted by Figure 2) in our topology of 19 cells. It is easy to see that using the upper confidence bound of the predicted value, ARIMA models may overestimate the bandwidth needs for reservations. Hence, it is important to take into account the tradeoff between the HDP and the bandwidth utilization.

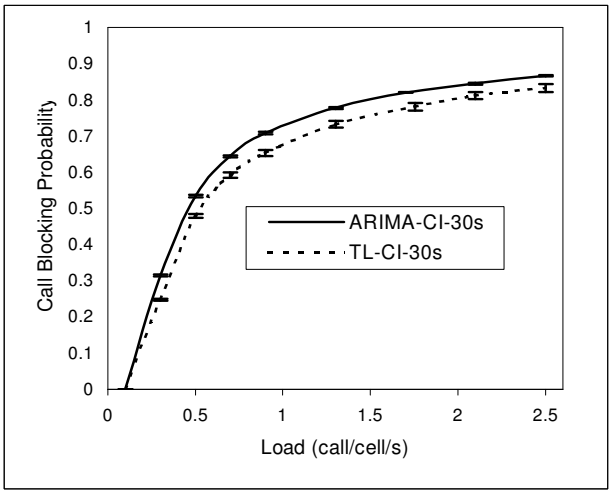

Figure 5 CBP - Prediction Interval: 30s; Reserved Bandwidth Type: Upper Confidence Bound Value

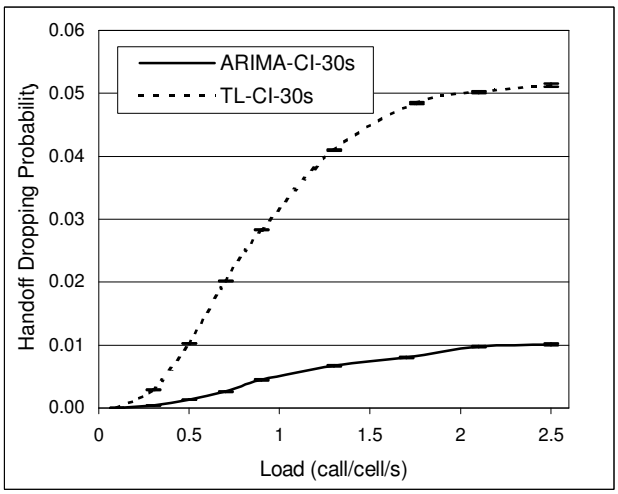

Figure 6 HDP - Prediction Interval: 30s; Reserved Bandwidth Type: Upper Confidence Bound Value.

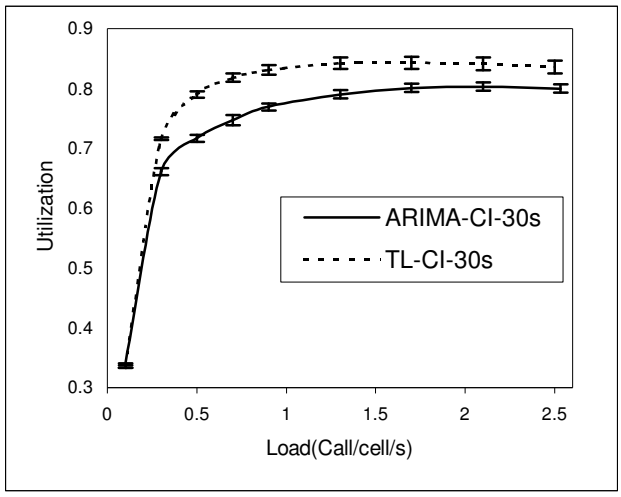

Figure 7. Utilization - Prediction Interval: 30s; Reserved Bandwidth Type: Upper Confidence Bound Value

We believe that a more interesting prediction approach is to adopt an adequate prediction interval and the predicted value forecasted by the method (TL or ARIMA). By regulating the prediction interval (i. e., the time window adopted for making forecasts), it may be possible to achieve the desirable level of HDP without jeopardizing the bandwidth utilization.

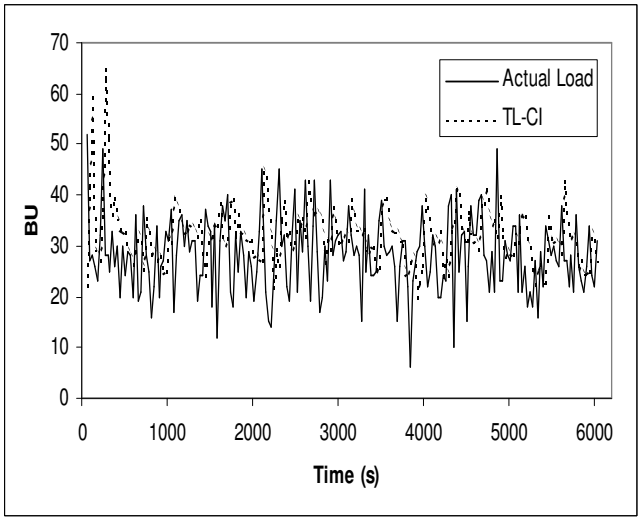

Figure 8. Actual and upper confidence bound for the predicted handoff load (TL).

In order to check if a different prediction interval could provide a smaller and controlled HDP Figure 10 and Figure 11 show the results for TL and ARIMA using different prediction intervals and the predicted value instead of the upper confidence bound for that value to reserve the wireless bandwidth. When the methods use 60s as the prediction interval, both methods kept HDP below 5%. The ARIMA model achieves the smallest HDP in both scenarios (i.e., with 30 and 60s). Again, choosing the appropriate prediction interval is a tradeoff between the desirable HDP and bandwidth utilization.

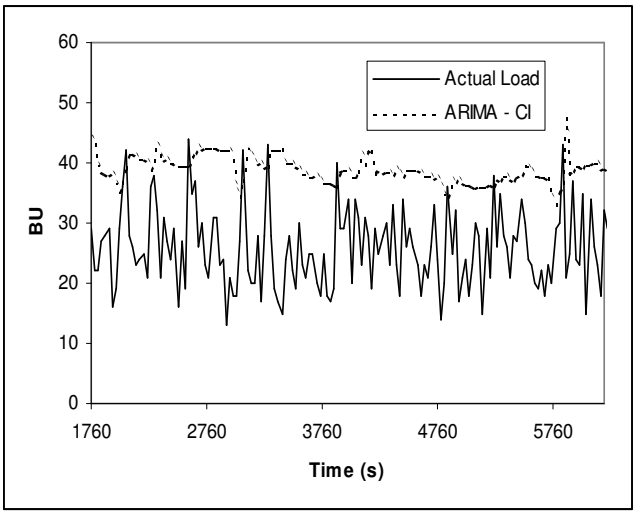

Figure 9 Actual and upper confidence bound for the predicted handoff load (ARIMA).

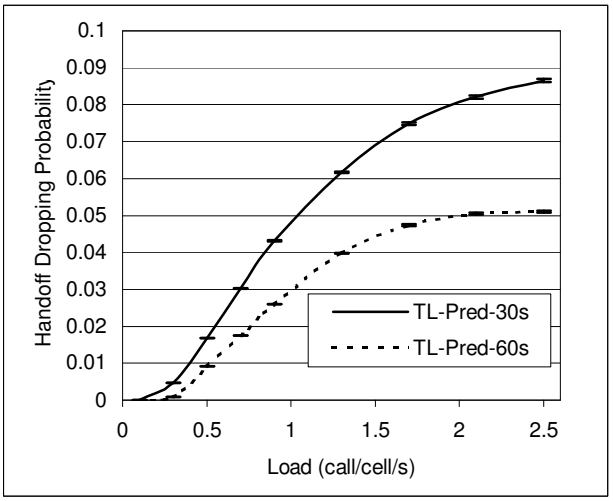

Figure 10. HDP- Prediction interval: 30s and 60s (TL).

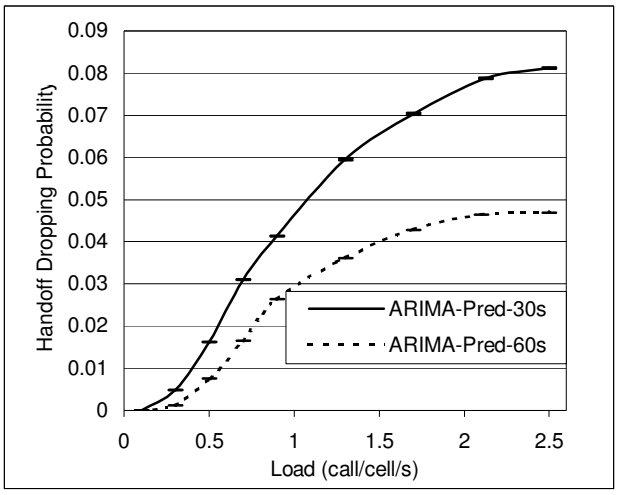

Figure 11 HDP – Prediction interval:30 and 60s (ARIMA).

6. CONCLUDING REMARKS

In this paper, we propose a novel CAC scheme for wireless and mobile networks that avoids per-user reservation signaling. In order to predict the expected bandwidth of future handoffs we utilized two time series-based methods: an adaptive exponential smoothing method, called Trigg and Leach (TL), which is effortless and does not impose computation overhead on the network elements and, the ARIMA-based method that requires a training period for model selection. In addition, TL method does not require a huge amount of saved data to perform forecasting, but ARIMA-based does. Our approach can also grant an upper bound on the handoff dropping probability even under higher loads based on the choice of an adequate prediction interval. Our future work is concerned with the proposal of an adaptive algorithm to dynamically adjust the TL's and ARIMA's prediction interval to optimize the bandwidth utilization depending on the current network load, mobility scenario and HDP objectives.

7. REFERENCES

[1] D. Hong, S. Rappaport, "Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritised and Nonprioritised Handoff Procedures," IEEE Tran. Vehic. Tech., Aug. 1986

[2] S. Choi and K. G. Shin, "Predictive and Adaptive Bandwidth Reservation for Handoffs in QoS Sensitive Cellular Networks," In Proc. ACM SIGCOMM´98.

[3] M. Naghshineh and M. Schwartz, "Distributed Call Admission Control in Mobile/Wireless Networks," IEEE JSAC, 14(4), May 1996. pp. 711-717.

[4] M.H. Chiu and M. A. Bassiouni, "Predictive Schemes for Handoff Prioritization in Cellular Networks based on Mobile Positioning," IEEE JSAC, 18(3), Mar. 2000

[5] Z. Xu et al. " A New Adaptive Channel Reservation Scheme for Handoff Calls in Wireless Cellular Networks," Proc. of IFIP Networking2002. pp 672-684, 2002.

[6] W.-S. Soh and H. S. Kim, "Dynamic Guard Bandwidth Scheme for Wireless Broadband Networks," Proc. IEEE Infocom'01, Anchorage, AK, Apr. 2001

[7] W.-S. Soh and H. S. Kim, "QoS Provisioning in Cellular Networks Based on Mobility Prediction Techniques," IEEE Comm. Mag., Jan. 2003, pp 86-92.

[8] T. Zhang et al. " Local Predictive Resource Reservation for Handoff in Multimedia Wireless IP Networks," IEEE JSAC, 19(10), Oct. 2001.

[9] D. W. Trigg, D. H. Leach, Exponential Smoothing with an Adaptive Response Rate, Operational Research Quarterly, vol. 18, 1967, pp. 53-59

[10] W. A. Fuller, Introduction to statistical time series (New York: John Wiley & Sons, Inc., 1996).

[11] Andrew C. Harvey, "Time Series Models", Cambridge: MIT Press, 2nd. Ed., 1993

[12] S F. L Fernandes et al. "Time Series Applied to Network Traffic Prediction: A Revisited Approach," In International conference on applied modelling and simulation-AMS 2002, MA, USA.

[13] L. Bow and W. Leroy, "Toward an all-IP based UMTS System Architecture," IEEE Network, 15(1), Jan-Feb 2001, pp.36-45.

[14] R. Berezdivin et al., "Next-Generation Wireless Communications Concepts and Technologies," IEEE Comm. Mag., 40(3), Mar 2002, pp. 108-116.

[15] R. Braden et al. " Integrated Services in the Internet Architecture: an Overview," IETF RFC 1633, Jun. 1994.

[16] S. Blake et al. "An architecture for Differentiated Services," IETF RFC 2475, Dec. 98.

[17] K. Nichols et al, "A Two-bit Differentiated Services Architecture for Internet. RFC 2638, Jul 1999.