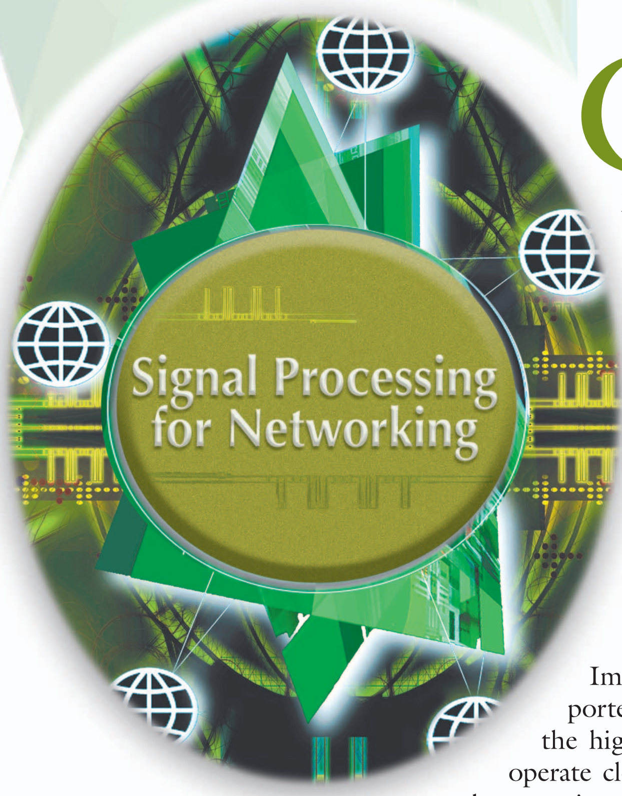
An overview of CAC principles in DS-CDMA networks

Call Admission Control in Wireless Multimedia Networks

Raghuveer M. Rao, Cristina Comaniciu, T.V. Lakshman, and H. Vincent Poor



Call admission control (CAC) is a mechanism used in networks to administer quality of service (QoS). Whereas the CAC problem in time-division multiple access (TDMA)-based cellular networks is simply related to the number of physical channels available in the network, it is strongly related to the physical layer performance in code-division multiple access (CDMA) networks since the multiaccess interference in them is a function of the number of users and is a limiting factor in ensuring QoS. The CAC mechanism will thus rely on the "soft capacity" of the CDMA network as determined by the level of multiaccess interference, often characterized by the signal-to-interference ratio. In such systems, the CAC design leads to significant interaction between the physical- and medium-access control layers. Multimedia wireless services stand to benefit from CAC schemes in providing differentiated service guarantees to different classes of users.

We begin with a problem stated in economic terms. Imagine a highway whose building and operation are supported by customers paying tolls. From a revenue standpoint, the highway operator is most interested in having the highway operate close to full capacity. However, customers are interested in the opposite: comfort and ease of passage. The more vehicles there are, the slower the passage and greater the likelihood of delays; that is, the quality of travel suffers. A similar situation prevails in the world of multimedia data transmission over communication networks. Any given network has finite resources; that is, the numbers of nodes, links, and buffers and the bandwidth are finite. Thus, there are a maximum number of packets that can be in a network at any given time. Although there are considerations related to the economics of networks that favor operating at or close to full capacity,

1053-5888/04/$20.00©2004IEEE

there are other considerations, related to QoS, that provide impetus to operating at less than full capacity.

The higher the packet traffic in a network or part of a network, the greater the average delay per packet due to the limited resources. Generally speaking, therefore, the more packets there are, the lower the QoS. Thus, maintaining QoS requires limiting the number of calls. However, rejection of calls not only lowers potential revenue but also creates a perception on the customer's part of the provider's inability to offer service when needed. Therefore, the service provider has an interest in reducing call-blocking probability. End-to-end delay is just one element of QoS. There are other QoS measures that must be considered as well. The overall problem faced by the network is one of measuring and forecasting QoS, minimizing call blocking probability, and maximizing throughput while maintaining QoS.

The network should provide various levels of service as a function of customer categories. The typical parameters that must be managed are latency, jitter, bandwidth, and packet-loss rate [1], [2]. Requirements vary as a function of multimedia data type and service (see Table 1). Packet loss is mainly due to buffer overflow. A buffer at a router simply drops incoming packets if the buffer is full. Packet corruption can occur as a result of transmission error, that is, erroneous reception of bits due to physical layer impairments.

A lightly loaded network affords less loss of packets due to a lower likelihood of overcrowded buffers. Light loading can also reduce end-to-end delay and, in wireless networks based on CDMA protocols, lower packet corruption caused by interference. QoS provisioning refers to a network's capability to provide different levels of service to different classes of traffic. To implement QoS provisioning, a desired QoS is negotiated between the customer and the network on each call, and the network QoS parameters are set accordingly.

Table 1. QoS requirements for multimedia service classes.

Multimedia traffic typically consists of streams of text, video, and voice. This traffic may result from applications that are streaming (one-way real-time transfer), interactive (two-way real-time) or nonreal time (Web browsing, file transfer). Interactive applications, such as voice over IP, have stringent QoS requirements needing rate, delay, and jitter (delay variation) guarantees. Carrying voice requires rate guarantees of a few kilobits to tens of kilobits per second depending on the compression used. Reductions in rate will result in poorer voice quality noticeable to the users. Round-trip delays have to be low (100–200 ms), and jitter has to be small as well while packet losses can be tolerated to an extent. Other applications, such as the recently popular gaming applications, are also delay sensitive and may need rate guarantees. For streaming applications, typically voice or video streaming, a guaranteed rate is important, but delay, delay jitter, and packet loss can be tolerated. Such applications use an initial permissible startup delay to fill a playout buffer that can be made large enough to absorb jitter and some short-term rate fluctuation. Rate guarantees are needed to prevent buffer overflow or underflow at longer time scales. Several nonreal-time applications, such as Web browsing and file transfers, are elastic in the sense that they can adapt to the rates that can currently be made available by the network.

Physical layer issues are an essential component of QoS management in wireless networks, especially with mobile platforms as varying channel conditions and numbers of users directly affect reliability of communication. Thus, QoS-assurance schemes must potentially integrate functions at the physical- and medium-access control (MAC) layers. CAC has emerged as one key component of such schemes. The reader can refer to other articles, in particular [3], for addressing other issues of relevance to QoS. The basic premise of CAC is to admit a new user into the network only when admitting the user satisfies the assured QoS of all existing users as well as that of the new user. Clearly, this requires quantification of QoS requirements. As the capacity of CDMA systems is interference limited, the signal-to-interference ratio (SIR), which has a direct bearing on the bit error rate (BER), is usually used as input to the CAC algorithms. The CAC design requires understanding and characterization of both the physical layer performance and the access control performance at the MAC layer.

In this article, we review CAC issues in multimedia CDMA systems by illustrating the basic principles underlying the various schemes that have been proposed progressively, from the simplest to the complex. We introduce SIR as a measure of QoS and describe relatively simple schemes to administer CAC. The expression for SIR resulting from linear minimum mean-squared error (LMMSE) processing is then presented. We next illustrate how CAC for multiple-class service can be cast into an optimality framework and then discuss recent work addressing self-similar multiple access interference. Finally, concluding remarks are provided. It must be noted that, although most of the material presented has been developed in the DS-CDMA context, several of the concepts are applicable to CAC in other systems as well.

Signal-to-Interference Ratio and Complete Sharing Schemes

Simple CAC schemes based on SIR measurements were first proposed by Liu and El Zarki [4]. Suppose there are M cells in a DS-CDMA network of interest with n_k calls in progress in cell k and we model the reception at the receiving antenna of a particular cell's base station (BS) taking into account path loss, log-normal shadowing, and multipath fading through the following expression for the received field strength [4]:

$$\Gamma(r) = 10^{\xi/10} r^{-\alpha} \tag{1}$$

where α is a constant typically ranging from two to four, r is the distance between the receiver and transmitter, and ξ is the transmit field strength in decibels and is normally distributed. Assuming the background Gaussian noise to be insignificant in comparison with the intrasystem interference (see [4]), the total power received by the BS in cell k is the sum of the power from all the mobiles in the system and is given by

$$I(k) = \sum_{h=1}^{M} \sum_{i=1}^{n_h} I_i(h, k) \quad (2)$$

where $I_i(h, k)$ is the power received by cell k's BS from mobile i of cell h. Suppose, with ideal power control (i.e., each mobile's signal is perceived with the same strength at its BS), S is the power level of a mobile's signal at its home cell BS and $r_{ik}^{(h)}$ is the distance between mobile i of cell h and the BS of the cell k. Then

$$I(k) = Sn_k + S \sum_{h \neq k} \sum_{i=1}^{n_h} \left(\frac{r_{ih}}{r_{ik}^{(h)}} \right)^{-\alpha} 10^{(\xi_{ik} - \xi_{ih})/10}. \quad (3)$$

Equation (3) assumes the controlled power per mobile at each home BS is the same. Thus, the SIR at BS k is given by

$$\mathrm{SIR}_k = \frac{S}{I(k) - S}. \quad (4)$$

The variable quantity in the expression for SIR in (4) is *I(k)*, the total power received at the kth BS. Equation (3) shows that this is a random variable because it depends on several other random variables, namely, the number of callers, their positions, and the transmitted power of interfering calls in neighboring cells. As the number of out-of-cell callers increases, the distribution of the out-of-cell interference tends to a Gaussian [5].

The capacity of CDMA systems is limited by the level of multiaccess interference in the system, which is measured by the SIR. In general, because SIR drops and the probability of packet error increases as the number of users increases, it appears reasonable to maintain SIR above set thresholds by limiting the number of incoming users. This ensures QoS when it is seen as maintaining BER below a certain limit. The *residual capacity* of a cell is defined as the maximum number of additional calls the BS can accept such that the system-wide outage probability (i.e., the probability that the transmission quality is less than acceptable) is maintained below a certain level. An appropriate criterion then for accepting a new call is to do so when the cell's residual capacity is greater than or equal to one. By improving the physical layer performance, through multiuser detection receivers [6], [7], for example, lower SIR thresholds can be set and more call requests may be admitted in the system for the same level of QoS.

Call requests can either be from new calls or handoffs. Demands are placed on the residual capacity principally by new calls because handoffs typically take place when the strength of the signal received from the mobile at the new BS is higher than that at the current BS. Therefore, power control with the new BS results in reduced power output from the mobile, at least for some time after completion of handoff.

Liu and El Zarki [4] propose two algorithms for CAC based on the SIR. From (3) and (4), the reciprocal of the SIR can be written as

$$\frac{1}{\mathrm{SIR}_k} = n_k + \sum_{h \neq k} \sum_{i=1}^{n_h} \left(\frac{r_{ih}}{r_{ik}^{(h)}} \right)^{-\alpha} 10^{(\xi_{ik} - \xi_{ih})/10} - 1. \quad (5)$$

For a given lower bound on the reciprocal SIR, an upper bound can be derived for n_k, the number of users in cell k. Suppose the actual SIR is required to be above a threshold $\mathrm{SIR}_{\mathrm{TH}}$. As the term to the right of the plus sign in (5) does not depend on n_k, the difference between the reciprocal SIRs for two different values of n_k would simply be equal to the difference between those values. Therefore, the residual capacity at cell k is given by

$$R_k = \max\left(0, \left\lfloor \frac{1}{\mathrm{SIR}_{\mathrm{TH}}} - \frac{1}{\mathrm{SIR}_k} \right\rfloor \right) \quad (6)$$

where the function $\lfloor\ \rfloor$ denotes the largest integer less than or equal to its argument. Thus, a call is admitted if the residual capacity is at least equal to one.

The CAC approach outlined above is very simple. One merely measures the uplink (i.e., reverse link) SIR at each BS and decides on the residual capacity with respect to the threshold $\mathrm{SIR}_{\mathrm{TH}}$. A call is admitted at a BS if there is any residual capacity left. The value of $\mathrm{SIR}_{\mathrm{TH}}$ itself is chosen such that the outage probability is below a specified limit. This requires determining, experimentally or through simulations, the variation of outage probability as a function of $\mathrm{SIR}_{\mathrm{TH}}$. The probability of a call being blocked at a BS is simply the probability of the station's residual capacity going to zero. As changes in $\mathrm{SIR}_{\mathrm{TH}}$ affect the outage and blocking probabilities in opposing directions, choosing the SIR threshold requires striking a compromise between the two probabilities.

The above algorithm has a key limitation. The reciprocal SIR is affected not only by the number of callers in the cell of interest but also by those in other cells along with their positions relative to the BS being considered, as shown in (5). For example, the admission of a few new callers in a short time span by BSs in a neighborhood based on their assessment of individual residual capacities might actually cause the expression for the second term in parentheses of (6) in one or more of the BSs to become negative. Liu and El Zarki [4] propose a second approach to address this problem, which results because the previously outlined approach lets each BS proceed on its own assuming a static view of traffic at other BSs.

Liu and El Zarki's second algorithm is a modification of the first. Suppose each BS, in addition to its own SIR computation, also receives the SIR measurements of its adjacent cells. Such exchange of information would be a peer-to-peer communication at the network layer. Then, the base in cell k evaluates the residual capacity in its cell as

$$\min\left(\left\lfloor \frac{1}{\mathrm{SIR}_{\mathrm{TH}}} - \frac{1}{\mathrm{SIR}_k} \right\rfloor, \left\lfloor \frac{1}{\beta}\left(\frac{1}{\mathrm{SIR}_{\mathrm{TH}}} - \frac{1}{\mathrm{SIR}_j}\right) \right\rfloor\right)$$

$$\text{for } j \text{ over adjacent cells.} \quad (7)$$

Here β is a parameter that permits taking into account the fact that the residual capacity is affected by activity in neighboring cells. The key contribution of the second term in (7) is in forcing each cell to contribute to the lowering of SIR in its neighboring cells. The parameter β becomes another quantity to be determined along with $\mathrm{SIR}_{\mathrm{TH}}$ to balance the blocking and outage probabilities. As the SIR in the cell of interest should be given more weight than that from neighboring cells, we require $1/\beta > 1$. In a simulation study of a cluster of hexagonal cells (arranged into seven rows and seven columns), it was found [4], as expected, that increasing β lowered the outage probability but increased the call-blocking probability. By conducting simulations at different loading conditions (call arrival rate) ranging from a nominal "normal" loading condition of 45 Erlangs per cell to an overloading of 80 Erlangs per cell, the authors determined that a $\mathrm{SIR}_{\mathrm{TH}} = -19$ dB and $\beta = 0.5$, satisfied the design criterion that the outage probability be at most 0.02 under all loading conditions.

Even with these simple schemes, it is clear that CAC in CDMA networks works across layers. Results of measurement at the physical layer are conveyed to the MAC layer where the call admission decision procedures, such as Liu and El-Zarki's Algorithms I or II, are implemented. The schemes described above fall under the category of complete sharing CAC schemes, that is, a new user is always allowed into a cell as long as there is nonzero residual capacity in that cell. The schemes do not differentiate between the various classes of service with different QoS requirements that one encounters in multimedia transmission. A few other complete-sharing approaches have been proposed (see [8]). Also, an alternative path-loss model to the one in (1) exists for microcell, multipath, and line of sight (LOS) propagation that is valid in inner cities [5], [9]. Methods have been proposed for real-time measurement of SIR (e.g., [10]) in DS-CDMA systems that can be used to provide input to the SIR computation.

Improving Physical Layer Performance

Although (4) looks at SIR purely as it is perceived at the BS antenna, it is obviously affected by the receiver processing. Specifically, various multiuser-detection and channel-compensation schemes have been proposed that lead to an enhancement of SIR (e.g., [11]–[13]). Here we present expressions for SIR resulting from LMMSE processing. Suppose we model the received signal at a BS as [7]

$$r(i) = \sqrt{P_1}b_1(i)h_1(i)s_1 + \sum_{k=2}^{K}\sqrt{P_k}b_k(i)h_k(i)s_k + \sigma w(i) \quad (8)$$

where P_k, $b_k(i) = \pm 1$, and $h_k(i)$ denote the transmit power, the ith transmitted bit, and channel gain respectively for the kth user and $s_k \in R^N$ is the kth user's signature sequence. The final additive term denotes white noise with variance σ^2 with $w(i)$ being a zero-mean, unit variance, circularly symmetric, complex Gaussian random variable. The sequence $h_k(i)$ is also complex valued and random with mean and variance denoted $\bar{h}_k$ and ξ_k^2, respectively. Thus, over a bit interval, $r(i)$ yields an N-dimensional vector with N being the CDMA processing gain.

Suppose we desire to detect data from user 1. An LMMSE receiver [14] estimates $b_1(i)$ as the sign of the real part of the inner product of a vector c_1 and $r(i)$ where

$$c_1 = \sqrt{P_1 h_1}\left(\mathbf{SDS}^T + \sigma^2\mathbf{I}_N\right)^{-1} s_1. \quad (9)$$

In (9), $\mathbf{D}$ is a $K \times K$ diagonal matrix whose kth diagonal entry is $P_k E\{h_k(i)h_k^*(i)\}$ and $\mathbf{S}$ is a $N \times K$ matrix whose kth column is s_k. The resulting SIR is given by [15]

$$SIR_1 = P_1|\bar{h}_1|^2 s_1^T\left(\mathbf{S}_1\mathbf{D}_1\mathbf{S}_1^T + \sigma^2\mathbf{I}_N + P_1\xi_1^2 s_1 s_1^T\right)^{-1} s_1 \quad (10)$$

where $\mathbf{S}_1$ is a $N \times (K-1)$ matrix that is a truncation of $\mathbf{S}$ with the first column removed, and $\mathbf{D}_1$ is a truncation of $\mathbf{D}$ with its first row and first column removed.

If the signature sequences of the K users are chosen randomly and independently, for large systems (large spreading gains and large numbers of users), the SIR expression of (9) can be well approximated by [7]

$$\widehat{\mathrm{SIR}}_1 = \frac{P_1|\bar{h}_1|^2\gamma}{1 + P_1\xi_1^2\gamma} \quad (11)$$

where γ is the fixed point of the expression

$$\gamma = \left[\sigma^2 + \frac{1}{N}\sum_{k=2}^{K}\frac{P_k\left(\xi_k^2 + |\bar{h}_k|^2\right)}{1 + P_k\left(\xi_k^2 + |\bar{h}_k|^2\right)\gamma}\right]^{-1}.$$

Being dependent only on the transmit power of all channels, channel means and variances, and independent of the signature sequences, the expression in (11) provides a simplification that makes numerical calculation easier in evaluating CAC procedures [7]. In the next section, we outline the optimal CAC procedure of [7], which also uses the LMMSE receiver processing just outlined.

Differentiated Services and Optimal CAC

With undifferentiated service, simple complete sharing would work fine. In such cases, there is not much more one can do beyond maintaining the SIR for each user. However, with multiple classes of service, the required minimum SIR depends on the user class (i.e., class of service specified by the user). The process of call requests and service can be cast into a Markov framework if the requests arrive in Poisson fashion and the calls have exponential holding times.

Various schemes have been proposed to address the issue of traffic type differentiation in the call admission process. A systematic way of doing it is to index the various classes as class 0, class 1, and so on. Suppose there are L such classes, that is, $0, 1, \ldots, L-1$. CAC schemes have been proposed in which the classes are first prioritized. For example, the assignment of the indices to the different classes can be done such that class i has priority over class j if $i < j$ [16]. This means that an incoming call for class i has a higher priority than that of class j if both calls arrive simultaneously. Furthermore, admission is granted only if the SIR resulting after the admission does not fall below a threshold. In threshold-based schemes, thresholds are set for admission of a call of a given class at each BS. To illustrate, let us suppose there is a call request from class i. Let

$$\mathrm{SIR}_{\ell,i};\ \ell = 0, \ldots, L-1 \tag{12}$$

be an estimate of class ℓ traffic's SIR resulting from accepting the class i call. Then, the call is accepted only if

$$\mathrm{SIR}_{\ell,i} \geq \Phi_{\ell,i};\ \ell = 0, \ldots, L-1 \tag{13}$$

where $\Phi_{\ell,i}$ are thresholds. The thresholds are set such that if two classes of traffic have the same data rate, the threshold is lower for the higher priority traffic. If the data rates are not the same, the thresholds are such that (13) is more likely to be satisfied for the higher priority class than the lower priority class. That is, with different data rates for different classes, one may end up having a higher threshold for a higher priority class than a lower priority class. However, with identical numbers of users, the left-hand side of the inequality in (13) will likely be greater for the higher priority traffic than the lower in such cases. In all threshold-based schemes, the thresholds are selected to achieve bounds on outage and blocking probabilities.

The threshold-setting process above for the different classes is somewhat ad hoc. It is actually possible to cast the CAC problem into an optimality framework when we assign varied costs to actions taken during the call admission process. The process of call requests and service fit into a Markov setting if the requests arrive in Poisson fashion and the calls have exponential holding times. With the mix of cost assignments and the Markov set up, it becomes possible to formulate an optimization problem.

We develop the optimization framework as follows. Suppose there are L user classes, of which some correspond to voice traffic and the others to data traffic. We assume that call admission decisions are made at time instants referred to as *decision epochs*. An epoch itself coincides with an arrival or departure, with the decision epochs corresponding to arrivals. Suppose there are $x_l(n)$ users of class l at epoch n. Furthermore, let there also be $r_l(n)$ service requests for that class. That is, $x_l(n)$ is the number of class-l calls in progress, which means that they were admitted at a previous decision epoch, and $r_l(n)$ is the number of class-l call requests that are waiting to be admitted. The CAC state at epoch n is given by the state vector

$$\mathbf{x}(n) = [x_1(n), r_1(n), \ldots, x_L(n), r_L(n)]^T. \tag{14}$$

Let the various possible states be denoted $\mathbf{x}_i$; $i = 1, \ldots, I$. The action that the CAC controller takes is either to admit or reject a call request. For example, the action might be to admit a_1 users of Class 1, a_2 users of Class 2, and so on, on a first-come/first-serve basis. Call requests of any class that arrive when the buffer of that class is full are automatically rejected. A set of actions prescribed as a function of state and epoch defines a CAC policy. Suppose we are interested in those policies that prescribe actions strictly as a function of state and not the epoch, that is, a particular value of the state vector always engenders the same action regardless of time instant. Such policies are known as stationary policies. Under a stationary CAC policy, if the call request arrival processes among the classes are independent and Poisson while the call holding times are also independent and exponentially distributed, then the state transitions conform to a first order Markov process. That is,

$$\begin{aligned} P(\mathbf{x}(n+1) = \mathbf{x}_i | \mathbf{x}(n), \mathbf{x}(n-1), \ldots) \\ = P(\mathbf{x}(n+1) = \mathbf{x}_i | \mathbf{x}(n)). \end{aligned} \tag{15}$$

One optimization criterion is to minimize the average number of requests rejected at any decision epoch subject to constraints on the SIR for various classes. This optimization framework falls under the general category of a semi-Markov decision process or SMDP (see Table 2). Suppose admission requests for Class l arrive in Poisson fashion at rate λ_l per second and last an average of $1/\mu_1$ seconds. Then, for any stationary policy, it will be possible to evaluate the probability of transition from one state to another. Let

$$P_{ij}(U) = P(\mathbf{x}(n+1) = \mathbf{x}_j | \mathbf{x}(n) = \mathbf{x}_i) \quad (16)$$

under a policy U. A stationary policy U is a mapping

$$U : X \to A \quad (17)$$

where X is the set of states and A is the set of actions.

Let $P_i(U)$ be the probability of being in state $\mathbf{x}_i$ at a given decision epoch. Given $P_{ij}(U)$, the $P_i(U)$s are found by solving the equations

$$\sum_{j \neq i} P_j(U) P_{ij}(U) = P_i(U)$$
$$\sum_i P_i(U) = 1. \quad (18)$$

Let R_m be the number of calls of Class m rejected between two successive epochs and $E_U\{R_m|\mathbf{x}_i\}$ be the expected number of such rejected calls given that the CAC system was in state $\mathbf{x}_j$ at the first of these two epochs and policy U was adopted. Then

$$E_U\{R_m\} = \sum_{i=1}^{I} P_i(U) E_U\{R_m|\mathbf{x}_i\} \quad (19)$$

is the expected number of Class m calls rejected under the policy. A weighted average cost function can be formed as

$$C(U) = \sum_{i=1}^{I} w_i E_U\{R_i\}. \quad (20)$$

The optimization problem at hand is one of finding U that minimizes the average cost in (20) subject to a SIR constraint for each class. Optimization of average cost in an SMDP context can be solved either through the value iteration algorithm or through linear programming [7], [17], [18]. Simulation results in [7] show substantial improvement in performance, as measured by blocking probability, of the optimal CAC approach used in conjunction with LMMSE receiver processing, compared to nonoptimal approaches.

Linear programming suffers from the twin "curses" of dimensionality (i.e., exponential rise in time required to solve the problem as the dimension increases, as is typical for any dynamic-programming-based approach [19]) and modeling (i.e., requirement of knowledge of transition probabilities). Recently developed reinforcement learning algorithms for constrained Markov decision processes to compute the optimal policy without explicit knowledge of the transition probabilities of the underlying Markov decision process [20], [21] provide some relief from the curse of modeling. The optimization framework explained in this section can be enlarged by considering joint optimization between the network and the physical layer, as the state space used to construct the optimal policy depends on the physical layer capacity (given the SIR constraints), which, in turn, can be optimized (optimal power allocation) based on the current number of users admitted into the system, which is, of course, influenced by the admission control) [6].

Table 2. Semi-Markov decision processes and average cost criterion.

Semi-Markov decision processes (SMDPs) occur widely in economics and operations research. Suppose we track events whose instants of initiation are Poisson distributed and whose durations are independent and exponentially distributed. Then the system state, by which we mean the total number of events in progress at any time, is a Markov process. If the event durations are not exponentially distributed but obey some other probability law, then the state is not a Markov process. However, it is semi-Markov in the sense that a Markov chain, or discrete-time Markov process, results if we make the timeline discrete to instants or epochs where there are state transitions (i.e., either a new event is initiated or an event in progress is completed). Let $\mathbf{X}$ be the set of states of a Markov chain and $\mathbf{A}$ be the set of actions that we can take. Let

$$U : (\mathbf{X}, t) \to \mathbf{A}$$

define a mapping from the set of states to the set of actions at epoch t. If the transition probability from one state to another is dependent on the action taken at that state, we get a *decision process* whose state transitions may not constitute a Markov chain. A *control policy* U, as defined above, is a prescription for taking an action at each epoch. If a unique action is associated with each state independent of epoch and transition history as $U : \mathbf{X} \to \mathbf{A}$ the state transitions form a Markov chain and U constitutes a stationary policy. Let $c_U(i)$ be the cost incurred in state $\mathbf{x}_i$ under a stationary policy U. Under certain assumptions, the long-term average cost, that is, the cost per epoch tends in the limit to the expected steady-state cost $\sum_i \pi_i c_U(i)$ where π_i is the steady-state probability of being in state $\mathbf{x}_i$. The minimization of this cost function over all stationary policies can be accomplished using the value iteration algorithm or linear programming.

Recent Developments: Bursty Data and CAC

A recent development in wireless CAC research is the investigation of heavy-tailed and long-range-dependent (LRD) phenomena. Unlike voice traffic, data traffic is bursty in nature. That is, it typically occurs in bursts interspersed with extended periods of quietness. A model that has been developed to describe such traffic is that of on/off processes with heavy-tailed distributions for the durations of the on and off periods. It has been shown that aggregation of a large number of such sources leads to processes that are self-similar and often LRD [22]–[25]. If the wireless traffic is dominated by such data transmission, then the multiple-access interference (MAI) will also be a bursty LRD process [26].

The long-range dependence can potentially be exploited in CAC through long-range prediction. If the SIR can be predicted accurately several steps ahead, then one can make call admission decisions beforehand that would contribute to a smoother process by minimizing periods of underutilization. Zhang et al. [9] offer a scheme that combines rate control with CAC in the presence of long-range dependent MAI. Rate-control schemes are used to control the bit rate of the mobile users as a function of current load and SIR. If MAI can be predicted, so can the SIR as it is merely the ratio of the signal power to that of the MAI.

The predictor in Zhang et al. [9] takes the form of

$$\hat{I}(i) = \zeta \hat{I}^P(i) + (1-\zeta)\hat{I}^L(i); 0 < \zeta < 1 \qquad (21)$$

where $\hat{I}^L(i) = (1/m)\sum_{n=i-m}^{i-1} I(n)$ is a long-range prediction of the MAI at instant i based on samples of the MAI at the previous m instants and $\hat{I}^P(i) = I(i-1)$ is a short term prediction using the previous sample. It should be noted that the quantity I above is different from the total signal level in (2). Studies of optimum prediction of LRD processes using long-range autoregressive models [27], [28] point to the domination of short-term effects even in LRD processes when they are Gaussian (in which case they are fractional Gaussian). Thus, if LRD prediction is to offer significant gain, it may have something to do with the heavy-tailed nature of the process at hand. Further studies are required to explore this matter.

Another dimension can be added to the CAC problem if we consider the integration of Web traffic (characterized as self-similar) with real-time traffic such as voice. This topic is addressed in [29], where an admission control policy is designed for World Wide Web (WWW) users, such that voice activity is exploited and an average delay requirement for Web packets can be met. The average delay requirement for data packet transmission is caused by the MAC, which schedules data transmission according to the residual available capacity (after the predicted voice contribution has been subtracted). The residual capacity is determined such that SIR requirements (different SIR requirements can be considered for voice and WWW traffic) for all users in the system can be met. Assuming that new WWW connection requests arrive according to a Poisson distribution, a maximum arrival rate can be determined such that the desired QoS (SIR and average delay) can be met, and an upper level admission control can be implemented to limit the arrival rate for the WWW users to this arrival rate target.

Conclusion

We have provided an overview of CAC principles as they apply to wireless DS-CDMA networks offering multiple classes of service. The key distinguishing feature for CAC in such networks that leads to cross-layer design issues is the role of SIR. CAC algorithms for multimedia CDMA systems play the role of balancing the system load such that both physical layer QoS requirements (BER targets) as well as network layer requirements (delays) can be met for each user in the system. Among the implications of this role is the influence of node-level signal processing on CAC issues. In this context, it is evident that optimal joint design of optimization algorithms at the physical layer and admission strategy is highly desirable. Although work has been done for joint receiver optimization and admission control [7], the analytical framework it is based on is an SMDP formulation, which will not hold true for some applications of interest (such as Web browsing), which are characterized by self-similar traffic. Some other open problems that may be considered are the implementation of distributed algorithms for ad hoc networks and cross optimizations of the transmitter design and CAC. In addition to the schemes outlined here, novel approaches, such as interaction between CAC and packet scheduling, are being explored. Methods for improving the prediction of long-range dependent processes may enable more reliable MAI prediction for CAC in bursty traffic. A general formalism and recent results in this area are reviewed in the monograph by Comaniciu et al. [30]. The impact of CAC for wireless systems should be realized in the near future with the anticipated widespread availability of wireless multimedia services.

Acknowledgments

This work was supported in part by NSF Grant CCR 00-86017 and Xerox-UAC Grant HE-1442-2003.

Raghuveer M. Rao received the Ph.D. degree in engineering from the University of Connecticut in 1984. After serving as member of technical staff at Advanced Micro Devices Inc. from 1985 to 1987, he joined the Rochester Institute of Technology, where he is a professor of electrical engineering and imaging science. He has held visiting appointments with the Indian Institute of Science, the Naval Surface Warfare Center, the Air Force Research Laboratories, and Princeton University. He was an associate editor for *IEEE Transactions on Signal Processing* and *IEEE Transactions on Circuits and Systems-II*. He is currently an associate editor for the *Journal for Electronic Imaging*. He received the IEEE Signal Processing Society's Best Young Author Paper Award and is a Fellow of SPIE.

Cristina Comaniciu received the M.S. degree in electronics from the Polytechnic University of Bucharest in 1993 and the Ph.D. degree in electrical and computer engineering from Rutgers University in 2001. From 1998 to 2001, she was with the Wireless Information Network Laboratory (WINLAB) at Rutgers University. From 2002 to 2003, she was a research associate with the Electrical Engineering Department at Princeton University. Since August 2003, she has been an assistant professor at Stevens Institute of Technology, Department of Electrical and Computer Engineering. Her research interests include cross-layer design for wireless networks, radio resource management for cellular and ad hoc networks, admission/access control for multimedia wireless systems, and multiuser detection. She is a coauthor of *Wireless Networks: Multiuser Detection in Cross-Layer Design*.

T.V. Lakshman received the master's degree in physics from the Indian Institute of Science, Bangalore, and the Ph.D. degree in computer science from the University of Maryland, College Park. He is currently the director of the High-Speed Networking Research Department at Bell Laboratories. His recent research has been in issues related to traffic characterization and provision of quality of service, energy-minimizing routing and scheduling in wireless ad hoc networks, architectures and packet-processing algorithms, and high-speed routers. He was a corecipient of the 1995 ACM Sigmetrics/ Performance Conference Outstanding Paper Award and the IEEE Communications Society 1999 Fred W. Ellersick Prize Paper Award. He was an editor of the *IEEE/ACM Transactions on Networking* . He is a Senior Member of the IEEE.

H. Vincent Poor received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 until 1990, he was with the University of Illinois. Since 1990, he has been with Princeton University, where he is the George Van Ness Lothrop Professor in Engineering. His research interests are in the areas of wireless networks, advanced signal processing, and related fields. Among his publications is *Wireless Communication Systems: Advanced Techniques for Signal Reception*. He is a member of the NAE and is a Fellow of the IMS, OSA, and the IEEE. In 1990, he was president of the IEEE Information Theory Society, and in 1991–1992 he was on the IEEE Board of Directors. Among his recent honors are the IEEE Graduate Teaching Award, the NSF Director's Award, and a Guggenheim Fellowship.

References

[1] T.V. Lakshman, A. Ortega, and A.R. Reibman, "VBR video: Tradeoffs and potentials," *Proc. IEEE*, vol. 86, pp. 952–973, May 1998.

[2] O. Yu and S. Khanvilkar, "QoS provisioning over GPRS wireless mobile links," in *Proc. IEEE Wireless Communications and Networking Conf.*, Orlando, FL, 2002, vol. 1, pp. 322–326.

[3] R.A. Berry and E.M. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Processing Mag.*, vol. 21, pp. 59–68, Sept. 2004.

[4] Z. Liu and M. El Zarki, "SIR-based call admission control for DS-CDMA cellular systems," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 638–644, May 1994.

[5] C.-J. Ho, J.A. Copeland, C.-T. Lea, and G.L. Stuber, "On call admission control in DS/CDMA cellular networks," *IEEE Trans. Veh. Technol.*, vol. 50, pp. 1328–1343, Nov. 2001.

[6] C. Comaniciu and H.V. Poor, "Jointly optimal power and admission control for delay sensitive traffic in CDMA networks with LMMSE receivers," *IEEE Trans. Signal Processing*, vol. 51, pp. 2031–2042, Aug. 2003.

[7] S. Singh, V. Krishnamurthy, and H.V. Poor, "Integrated voice/data call admission control for wireless DS-CDMA systems," *IEEE Trans. Signal Processing*, vol. 50, pp. 1483–1495, June 2002.

[8] C. Comaniciu, N.B. Mandayam, D. Famolari, and P. Agrawal, "QoS guarantees for 3G CDMA systems via admission and flow control," in *Proc. 52nd IEEE Vehicular Technology Conf.*, 2000, vol. 1, pp. 249–256.

[9] P. Harley, "Short distance attenuation measurements at 900 MHz and 1.8 GHz using low antenna heights for microcells," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 5–11, Jan. 1989.

[10] D. Ramakrishna, N. Mandayam, and R. Yates, "Subspace based estimation of the signal-to-interference ratio for CDMA cellular systems," *IEEE Trans. Veh. Technol.*, vol. 49, pp. 1732–1742, Sept. 2000.

[11] U. Mitra and H.V. Poor, "Adaptive receiver algorithm for near-far resistant CDMA," *IEEE Trans. Commun.*, vol. 43, pp. 1713–1724, Feb./Mar./Apr. 1995.

[12] U. Madhow and M.L. Honig, "MMSE interference suppression for direct-sequence spread-spectrum CDMA," *IEEE Trans. Commun.*, vol. 42, pp. 3178–3188, Dec. 1994.

[13] H. Liu and M.D. Zoltowski, "Blind equalization in antenna array CDMA systems," *IEEE Trans. Signal Processing*, vol. 45, pp. 161–172, Jan. 1997.

[14] X. Wang and H.V. Poor, *Wireless Communication Systems: Advanced Techniques for Signal Reception*. Upper Saddle River, NJ: Prentice-Hall, 2004.

[15] J. Evans and D.N.C. Tse, "Large system performance of linear multiuser receivers in multipath fading channels," *IEEE Trans. Inform. Theory*, vol. 46, pp. 2059–2078, Sept. 2000.

[16] W.S. Jeon and D.G. Jeong, "Call admission control for CDMA mobile communications systems supporting multimedia services," *IEEE Trans. Wireless Commun.*, vol. 1, pp. 649–659, Oct. 2002.

[17] H.C. Tijms, *Stochastic Modelling and Analysis*. Chichester, U.K.: Wiley, 1986.

[18] W.-B. Yang and E. Geraniotis, "Admission policies for integrated voice and data traffic in CDMA packet radio networks," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 654–664, May 1994.

[19] R. Bellman and S. Dreyfus, *Applied Dynamic Programming*. Princeton, NJ: Princeton University Press, 1962.

[20] F.J. Vazquez-Abad and V. Krishnamurthy, "Self learning call admission control for multimedia wireless DS-CDMA systems," in *Proc. 6th Int. Workshop Discrete Event Systems*, Zaragoza, Spain, Oct. 2002, pp. 399–404.

[21] F.J. Vazquez-Abad and V. Krishnamurthy, "Policy gradient stochastic approximation algorithms for adaptive control of constrained time varying Markov decision processes," in *Proc. 42nd IEEE Conf. Decision and Control*, Maui, HI, 2003.

[22] O. Cappe, E. Moulines, J.-C. Pesquet, A.P. Petropulu, and X. Yang, "Long-range dependence and heavy-tail modeling for teletraffic data," *IEEE Signal Processing Mag.*, vol. 19, pp. 14–27, May 2002.

[23] K. Park and W. Willinger, *Self-similar Network Traffic and Performance Evaluation*. New York: Wiley, 2000.

[24] S. Lee, W. Zhao, R. Narasimha, and R.M. Rao, "Discrete-time models for statistically self-similar signals," *IEEE Trans. Signal Processing*, vol. 51, pp. 1221–1230, May 2003.

[25] S. Lee, R. Rao, and R. Narasimha, "Characterization of linear scale invariant system input/output relationships and synthesis of network traffic traces," in *Proc. 35th Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, 2001, vol. 1, pp. 730–734.

[26] J. Zhang, M. Hu, and N.B. Shroff, "Bursty data over CDMA: MAI self-similarity, rate control and admission control," in *Proc. IEEE INFOCOM 2002, 21st Annu. Joint Conf. IEEE Computer and Communications Societies*, New York, vol. 1, pp. 391–399.

[27] S.A.M. Ostring and H. Sirisena, "The influence of long-range dependence on traffic prediction," in *Proc. IEEE Int. Conf. Commun.*, Helsinki, Finland, 2001, pp. 1092–1101.

[28] R.M. Rao, S. Lee, E. Bayraktar, and H.V. Poor, "Signal processing models for discrete-time self-similar and multifractal processes," in *Proc. 37th Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, 2003, pp. 8–12.

[29] C. Comaniciu, N.B. Mandayam, D. Famolari, and P. Agarwal, "Wireless access to the world wide web in an integrated CDMA system," *IEEE Trans. Wireless Commun.*, vol. 2, pp. 472–483, May 2003.

[30] C. Comaniciu, N.B. Mandayam, and H.V. Poor, *Wireless Networks: Multiuser Detection in Cross-Layer Design*. Boston, MA: Kluwer, 2004.