Call Admission Control for QoS Provisioning in 4G Wireless Networks: Issues and Approaches

Dusit Niyato and Ekram Hossain, University of Manitoba and TR*Labs*

Abstract

This article presents a survey on the issues and the approaches related to designing call admission control schemes for fourth-generation wireless systems. We review the state of the art of CAC algorithms used in the traditional wireless networks. The major challenges in designing the CAC schemes for 4G wireless networks are identified. These challenges are mainly due to heterogeneous wireless access environments, provisioning of quality of service to multiple types of applications with different requirements, provisioning for adaptive bandwidth allocation, consideration of both call-level and packet-level performance measures, and consideration of QoS at both the air interface and the wired Internet. To this end, architecture of a two-tier CAC scheme for a differentiated services cellular wireless network is presented. The proposed CAC architecture is based on the call-level and packet-level QoS considerations at both the wireless and wired parts of the network. A performance analysis model for an example CAC scheme based on this architecture is outlined, and typical numerical results are presented.

Supporting multimedia applications with different quality of service (QoS) requirements in the presence of diversified wireless access technologies (e.g., 3G cellular, IEEE 802.11 WLAN, Bluetooth) is one of the most challenging issues for fourth-generation (4G) wireless networks. In such a network, depending on the bandwidth, mobility, and application requirements, users will be able to switch among the different access technologies in a seamless manner. Efficient radio resource management and call admission control (CAC) strategies will be key components in such a heterogeneous wireless system supporting multiple types of applications with different QoS requirements.

A CAC scheme aims at maintaining the delivered QoS to different calls (or users) at the target level by limiting the number of ongoing calls in the system. In contrast to traditional voice-oriented circuit-switched cellular wireless networks, 4G networks will be based on packet switching at the wireless interface and will interwork with IP-based Internet. Therefore, while designing a CAC scheme for such a network, packet-level performance measures (e.g., packet dropping probability, packet transmission delay) at both the wireless and wired interfaces (e.g., at the IP-aware wireless router/base station) will need to be considered in addition to the call-level performance measures.

CAC: General Model and Classification

A Threshold-Based Mechanism

The concept of threshold-based CAC is applicable for both hard- and soft-capacity wireless systems. Threshold-based CAC is performance based on the availability of resource **I** [1]. The objective of a threshold-based CAC is to maintain every element in **I** less than that in the threshold vector $\mathbf{I}_{th}$. These thresholds are defined based on the congestion condition of the system. When a call arrives, the algorithm estimates the increase $\Delta\mathbf{I}$ that the incoming call would cause to the current value of **I**. Generally, the CAC policy is based on the following condition:

$$\mathbf{I} + \Delta\mathbf{I} < \mathbf{I}_{th}. \qquad (1)$$

If this condition is satisfied, the incoming call is accepted; otherwise, it is rejected or queued. With this policy, the CAC scheme needs to be developed by considering the elements of matrix **I**, which are the performance measures of the system, the way to estimate increase $\Delta\mathbf{I}$, and the optimal value of threshold $\mathbf{I}_{th}$. In this case the threshold can be set statically without considering the current status of the network (static scheme). However, adaptive CAC algorithms (e.g., [2]) can adjust the thresholds dynamically, resulting in superior performance over static schemes.

For systems with hard capacity, that is, time-division multiple access (TDMA) and frequency-division multiple access (FDMA) systems, the elements of matrix **I** can simply be the number of occupied channels, and the increase in resource usage $\Delta\mathbf{I}$ can be the number of channels required by an incoming call. In this case, the thresholds can be chosen so that the target QoS levels can be achieved.

On the other hand, for systems with soft capacity (e.g., code-division multiple access [CDMA] and orthogonal frequency-division multplexing [OFDM] systems), there is no evident relationship between the number of users and the available capacity for incoming calls. For example, in a CDMA network the elements of matrix **I** can be derived from

0890-8044/05/$20.00 © 2005 IEEE

the signal-to-interference ratio (SIR) at the receiver, and an incoming call is admitted if the SIRs of the ongoing calls and that of the new call can be maintained above the desired level.

Classification

Centralized and Distributed Approaches — A call admission control algorithm can operate in either a centralized or distributed fashion. In the former case, the CAC algorithm is executed in a central site (e.g., mobile switching center [MSC]). In this case, information from every cell needs to be transferred to the central site, and the CAC needs to be performed remotely from the local cell. In distributed CAC, the CAC algorithm is executed locally at the base station of each cell.

A distributed CAC algorithm can follow either a collaborative or local approach. In the former case, information is exchanged among the neighboring cells for resource reservation and admission control while the decision is made locally. In the latter case, information collection and decision making are done locally. Although the collaborative approach can provide more accurate information for CAC decision, it incurs more communication overhead.

Traffic-Descriptor-Based and Measurement-Based Approaches — A call admission control policy can use either a traffic descriptor-based or measurement-based approach. In the traffic descriptor-based approach, it is assumed that the knowledge about the traffic pattern of the incoming calls is available. Therefore, a CAC algorithm can simply determine the expected amount of resource usage by summing together the resources used by all ongoing calls and the resource to be used by an incoming call. If the sum is less than some predefined threshold, the incoming call is accepted; otherwise, the call is rejected. Although traffic-descriptor-based CAC schemes are simple, they are relatively conservative since the ongoing calls may not always use the maximum amount of resources as specified in the descriptor.

Instead of using explicit traffic descriptors, the information on the traffic pattern can be obtained by measuring the characteristics of the call. In that case, a CAC decision can be made dynamically based on the actual state of the network (e.g., packet arrival rate). Measurement-based CAC schemes are based on this principle.

Most of the CAC algorithms in the hard-capacity cellular networks are traffic-descriptor-based. On the other hand, most CAC algorithms in CDMA systems are measurement-based in which SIR information is measured and used to ensure the QoS of the ongoing calls. The CAC algorithms used in WLANs mostly adopt a measurement-based approach.

Classification Based on the Granularity of Resource Control — A taxonomy of CAC algorithms based on the granularity of resource control was presented in [3]. Three different criteria were used to categorize a CAC algorithm. The first criterion is the type of information used by the decision making process for CAC. Generally, CAC algorithms consider resource usage of mobiles, but some consider mobility patterns of users. In the latter case, accuracy of resource reservation can be improved by taking the direction and speed of users into account.

The second design criterion is spatial distribution (position and movement) of mobiles, which can be either uniform or nonuniform. The third criterion is based on how the information is organized and manipulated by the CAC algorithms. The information can be on an aggregate of flows or per-flow basis in which the CAC algorithms use the information on the group of mobiles or an individual mobile, respectively. Based on these criteria, the granularities of CAC algorithms are different. For example, an algorithm that considers the resource usage of all ongoing calls assuming uniform spatial distribution (e.g., the guard channel scheme) and one that considers the mobility of an individual user assuming nonuniform spatial distribution have the largest and smallest granularity of resource control, respectively.

Traditional CAC Approaches

The Guard Channel Approach

To prioritize handoff calls over new calls, some channels (referred to as guard channels) are reserved for handoff calls [4]. Specifically, if the total number of available channels is C and the number of guard channels is $C - K$, a new call is accepted if the total number of channels used by ongoing calls (i.e., busy channels) is less than the threshold K, while a handoff call is always accepted if there is an available channel. According to this channel reservation, the threshold must be chosen such that the handoff call dropping probability is minimized, while the system can admit as many incoming new calls as possible.

Although the guard channel scheme with a static threshold is easy to implement, it may not be efficient. Higher channel utilization could be achieved through adaptation of the threshold according to the state of the network.

A more general scheme, the *fractional guard channel scheme*, was introduced in [5]. In this case an incoming call is accepted with a certain probability that depends on the number of busy channels. In other words, when the number of busy channels becomes larger, the acceptance probability for a new call becomes smaller, and vice versa. This helps keep the handoff call dropping probability smaller and also avoid congestion.

A Collaborative Approach Based on Estimation

This is a distributed approach to CAC. Information is exchanged among neighboring cells for resource reservation and admission control, while the admission control decision is made locally. CAC algorithms of this type were proposed in [6] that use estimates of call dropping and call blocking probabilities. The maximum number of ongoing calls N is estimated from

$$P_{hd} = \frac{1}{2} erfc\left(\frac{N - \overline{m}}{\sigma}\right), \tag{2}$$

where P_{hd} is the target call dropping probability, and $\overline{m}$ and σ denote the mean and variance of the number of calls in the home cell, respectively. The mean and variances are approximated from the number of users in the home cell and neighboring cells. The call blocking probability $P_{nb}(t)$ during time interval $t - 1$ to t is estimated locally as follows:

$$P_{nb}(t) = (1-\omega)P_{nb}(t-1) + \omega\frac{s(t)}{r(t)}, \tag{3}$$

where $s(t)$ and $r(t)$ are the number of blocked calls and the number of calls that arrived during time interval $t - 1$ to t, respectively, and ω is the weight used to calculate the exponential weighted moving average. The decision on whether an incoming call is accepted or rejected is made based on Eqs. 2 and 3.

A Noncollaborative Approach Based on Prediction

In a picocellular wireless network with high user mobility, exchanging information among cells for resource reservation and admission control might incur significant control overhead. Therefore, CAC algorithms designed based on local information (e.g., history of bandwidth usage) would be desirable. In such a case, resource reservation is based only on local information in the home cell, which is used to predict the resource needed in the future [2].

In [2] two prediction techniques were used: *Wiener filtering* and *time series analysis* (e.g., the autoregressive moving average [ARMA] model). In the former case the prediction can be done directly from the historic data, whereas in the latter case the time series model needs to be constructed and the corresponding parameters need to be estimated so that the prediction can be performed based on this model afterward. Such a local predictive approach to CAC was shown to perform as well as a collaborative approach when traffic fluctuation is moderate.

A Mobility-Based Approach

Mobility-based approaches exploit user mobility information for efficient CAC. For example, the *shadow clustering* concept was introduced in [7] based on user mobility information to estimate future resource requirements in a microcellular wireless network. The idea here is that every mobile terminal with an active wireless connection exerts an influence on the cells in the vicinity of its current location and along its direction of travel. To calculate the shadow cluster and the corresponding levels of intensity, the information on call holding time, current direction, velocity, and position of the active mobile terminal need to be considered.

Although mobility-information-based CAC schemes can improve the efficiency of resource reservation and admission control, calculating the amount of incoming traffic for a particular cell would be nontrivial; also, real-time exchange of control messages among cells would incur large communication overhead.

A Pricing-Based Approach

A pricing-based approach to CAC was proposed in [8], where the objective is to maximize the utility of wireless resources. Utility is generally defined as the users' level of satisfaction with perceived QoS. For example, utility is a decreasing function of new call blocking and handoff call dropping probabilities (i.e., P_{nb} and P_{hd}, respectively). However, maximizing the utility of the network might not maximize the revenue of the service provider. Specifically, for higher user satisfaction, more resources should be allocated to each user. In contrast, to maximize revenue under flat rate pricing, the allocations need to be degraded to accommodate more users. Therefore, a CAC scheme can be designed such that the optimal operating point can be obtained.

In [8] the optimal point between utility and revenue was determined in terms of the new call arrival rate, and a pricing scheme was developed to achieve this optimal effective arrival rate in the network. In this case the QoS metric P_b referred to as the grade of service (GoS) is defined as

$$P_b = \alpha P_{nb} + \beta P_{hd}, \tag{4}$$

where α and β are the weights corresponding to the new call blocking and handoff call dropping probabilities, respectively, and $\alpha + \beta = 1$.

The metric P_b can be defined as a function of new call arrival rate λ_n (i.e., $P_b = g(\lambda_n)$). Then the utility function becomes $U = h(P_b)$. Assuming flat rate pricing, the revenue depends on the admissible number of users, which again depends on the new call arrival rate $f(\lambda_n)$. The optimal value of the new call arrival rate λ_n^* that maximizes the total utility $U(\lambda_n) = f(\lambda_n) \times h[g(\lambda_n)]$ can be calculated by differentiating $U(\lambda_n)$, and finding the point at which the slope equals zero.

Based on this optimal new call arrival rate, the pricing scheme is developed by changing the cost of a call. The pricing scheme adjusts the fee dynamically by taking the state of the network into account. If the network is congested, it will charge *peak hour price* $p(t)$, which is higher than normal hour price p_0. According to this pricing scheme, the demand function, which describes the reaction of users to the change in price, is expressed by

$$D[p(t)] = \exp\left(-\left(\frac{p(t)}{p_0} - 1\right)^2\right), p(t) \geq p_0, \tag{5}$$

and the peak hour price is calculated by considering the state of the network. With this pricing scheme, a user has an incentive not to initiate a call during peak hours, so network congestion can be avoided.

Call Admission Control in CDMA Systems

Due to the soft-capacity feature, the admission control decision in a CDMA network should be based on the state of ongoing calls (e.g., interference level). The CAC approaches used in hard-capacity systems based on the assumption of time-invariant cell capacity may degrade the system utilization in a CDMA system. Also, due to the soft handoff feature, the length of a handoff process becomes longer than that of hard handoff, and the CAC algorithm must take this duration into account.

CAC schemes based on the estimation/measurement of current state of interference and SIR were proposed in the literature [9]. While in interference-based approaches the objective is to keep interference from all sources (i.e., other mobiles in the same cell and other cells) below the acceptable level, SIR-based approaches emphasize the SIR requirement for each call considering the statistical factors such as voice activity, fading, and shadowing in the channel.

The main ideas of all the above CAC approaches are summarized in Table 1.

Call Admission Control in 4G Wireless Networks

The diverse QoS requirements for multimedia applications and the presence of different wireless access technologies pose significant challenges in designing efficient CAC algorithms for 4G wireless networks. Table 2 summarizes some of these major challenges.

Heterogeneous Networking

Due to the seamless connection and global mobility requirements in a 4G system, a call in one particular network must be able to roam and be handed over to another network transparently. This is called vertical handoff, and several related issues need to be addressed. The usual signal-strength-based handoff initiation may not be enough, and other system parameters such as the congestion level at the network must be considered as well. For instance, mobile users with non-real-time applications can be handed over to WLANs in order to mitigate congestion in cellular networks. All these factors will impact the call holding time distribution in each network.

From the CAC point of view, a vertical handoff results in a new subtype of handoff call. A CAC algorithm must determine the priority of this type of call over new calls. A new performance metric, *vertical handoff call dropping probability*, should be determined and should be kept below the acceptable threshold. Also, issues related to call dropping and/or

Approach	Main Idea
Guard channel	Some portion of the wireless resources is reserved for handoff calls so that handoff call dropping probability can be maintained below the target level.
Fractional guard channel	New calls are gradually blocked according to the current status (i.e., the number of ongoing calls) of the network.
Collaborative	The neighboring cells exchange information about the network status so that resource reservation can be made in advance accurately.
Noncollaborative	Using prediction techniques (e.g., ARMA model, Wiener filtering) to project the amount of the resources required locally so that the resources can be reserved in advance without the need for information exchange among neighboring cells
Mobility-based	Mobility information (i.e., position and direction of movement) of mobiles can be used to enhance the accuracy of the resource reservation.
Pricing-based	Dynamic pricing is used to limit the call arrival rate so that the maximum utility and revenue of the system is achieved.

■ Table 1. *Different approaches to CAC design in cellular wireless networks.*

Requirements	Description
Heterogeneous environment	4G systems will consist of several types of wireless access technologies, so CAC schemes must be able to handle vertical handoff and special modes of connection such as ad hoc on cellular.
Multiple types of services	4G systems will need to accommodate different types of users and applications with different QoS requirements.
Adaptive bandwidth allocation	With multimedia applications, system utilization and QoS performance can be improved by adjusting the bandwidth allocation depending on the state of the network and users' QoS requirements.
Cross-layer design	Both call- and packet-level QoSs need to be considered to design CAC algorithms so that not only the call dropping and call blocking probabilities, but also the packet delay and packet dropping probabilities can be maintained at the target level.

■ Table 2. *Challenges in CAC design for 4G wireless networks.*

queuing need to be addressed. For example, if a cellular network cannot accept a call vertically handed over from a WLAN, the call may be dropped, or stay connected with the WLAN and wait until the cellular network is able to accommodate the call.

Multiple Classes of Services and Interoperability with DiffServ-Based IP Networks

CAC algorithms should be designed to support multiple classes of services, each with specific QoS requirements. The service classes can be chosen to be similar to those used in differentiated services (DiffServ) [10] IP networks. This would enable seamless integration of wireless networks with the IP-based Internet.

DiffServ is one of the key technologies for providing QoS in the Internet. Instead of providing QoS on a per-flow basis as in the integrated services (IntServ) model, DiffServ operates on groups of flows by aggregating several IP-level flows with the same QoS requirements into the same group. By using the type of service (TOS) field in the IP packet header, DiffServ routers can identify the group to which an IP packet belongs and then use appropriate traffic management schemes so that the QoS requirements can be satisfied at the aggregate level.

In a DiffServ domain three groups of traffic services are provided: premium, assured forwarding, and best effort. When interworking 4G wireless systems with DiffServ-based IP networks, conversational and streaming calls can be mapped into premium and assured forwarding service classes, respectively, while non-real-time services can be mapped into the best effort service class. With this architecture, the CAC module at a DiffServ edge router should be able to negotiate an appropriate service level agreement (SLA) with the DiffServ domain, and the decision to accept or reject a call should be made based on both the availability of wireless resources and the negotiated SLA.

Adaptive Bandwidth Allocation

Due to the diversity of applications and QoS requirements for mobile users and the dynamic nature of wireless channel quality, adaptive bandwidth allocation (ABA) would be necessary to improve utilization of wireless network resources. Therefore, CAC strategies should be designed taking this into account. With ABA, when the network conditions are favorable, the quality of a call can be upgraded by assigning more resources. However, when the network becomes congested, the amount of bandwidth allocated to some of the ongoing calls will be revoked to accommodate more incoming calls so that the call dropping and blocking probabilities can be maintained at the target level. In [11] such an ABA algorithm was proposed that allocates a target level of bandwidth to a connection as much time as possible.

Again, ABA is needed during vertical handoff. The acceptable bandwidth should be negotiated, and the CAC strategy should be based on the result of negotiation. For example, when a call is handed over to a cellular network from a WLAN, bandwidth adaptation will be required for that call.

Cross-Layer Design

For a wireless network, cross-layer optimization can lead to significant improvement in the transmission protocol stack performance [12]. In the context of CAC, cross-layer design should be applied to capture both call-level and packet-level QoS performance.

Traditionally, CAC schemes have been based on call-level QoS measures only, although some CAC schemes take physical layer parameters such as SIR into account. However, the radio link level performance (e.g., resulting from different scheduling and error control schemes) have not been considered in designing a CAC scheme. In contrast to a traditional voice-oriented circuit-switched network, in a purely packet-switched wireless network the QoS needs to be described in terms of both call-level (e.g., call blocking and dropping probabilities) and packet-level performance metrics (e.g., packet transmission delay and packet dropping probability). Therefore, a new call should be admitted only if the "quality" of all ongoing calls including the incoming call in terms of packet-level performance can be maintained at the desired level.

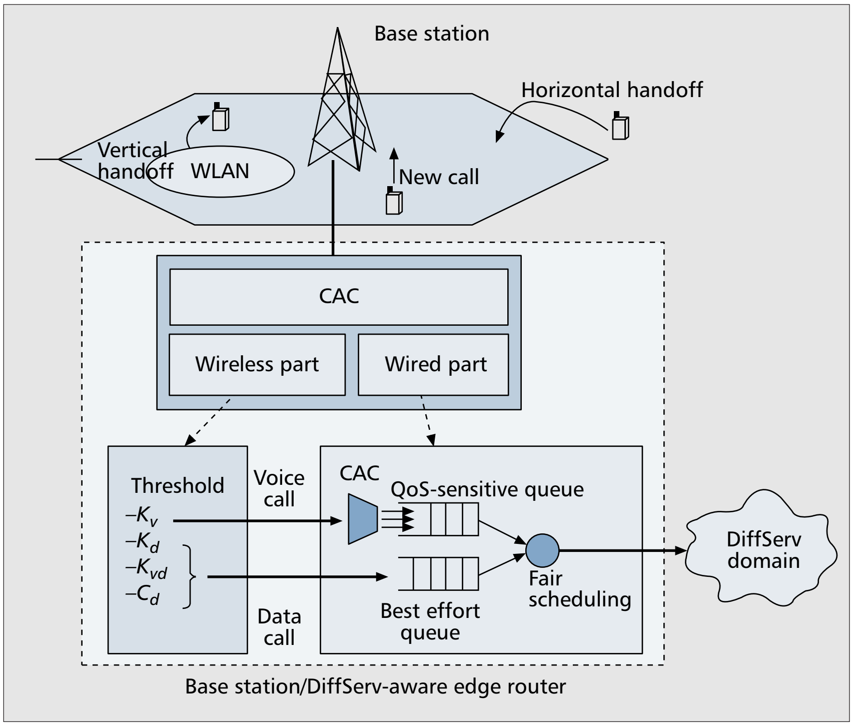

Figure 1. *The system model for the proposed CAC scheme.*

CAC for 4G Networks: Architecture and Example

A Novel CAC Architecture

We introduce a novel CAC architecture for 4G wireless networks. The CAC module is divided into two submodules (i.e., two-tier CAC): one for the wireless part and the other for the wired part (Fig. 1).

In the wireless part the CAC needs to handle multiple classes of calls as well as calls due to vertical handoff from other types of networks. If the call is used for data transfer, ABA can be applied to increase resource utilization. Moreover, CAC in the wireless part must consider the nature of capacity of the systems (i.e., soft or hard) so that resource reservation and admission control can be performed optimally.

The CAC submodule for the wired part, which internetworks with the DiffServ domain, is important in the sense that the dropping probability for the packets already transmitted over the air interface should be made as small as possible (to minimize waste of wireless resources), and the packet delay should not violate the SLA. Since the wireless resources are the scarcest resources in the system, the CAC submodule in the wired part must ensure that the wired network can maintain the QoS of traffic from wireless users (already transmitted across the wireless links) at the desired level.

Both the call- and packet-level performance requirements need to be satisfied in the wireless part. Packet-level QoS performance in the wireless part can be maintained through ABA and proper scheduling mechanisms. Call-level performance depends on the resource reservation and admission control strategy in the wireless part. However, in the wired part, only packet-level QoS requirements need to be satisfied.

Example: A Two-Tier CAC Algorithm

Using the above architecture, we show an example of a two-tier CAC algorithm that considers CAC at both the wireless and wired parts of the system. Threshold-based call admissions are employed to maintain call- and packet-level QoS. *A call will be admitted only if it can be accepted by the CAC components for both the air interface and wired part of the system*. For performance analysis, design, and engineering of the system, corresponding analytical models are developed. Typical numerical results based on the analytical modeling are also presented.

Tier I: The CAC Scheme in the Wireless Part — We assume that there are multiple types of users, and ABA is used to increase utilization of the wireless network resources. Vertical handoff from/to other types of networks is also taken into account. Threshold-based resource reservation and admission control is used.

The base station serves two types of calls, voice and data, and both these types of calls share a common pool of channels. The number of channels in the cell is fixed at C (i.e., hard-capacity), and one voice call requires only one channel. For a data call ABA is used to adjust channel allocation according to the state of the network. Under light load conditions, a data call is allocated as many channels as the user requests. Under heavy load conditions, each data call will receive at least one channel to maintain the connection.

To minimize handoff call dropping probability, the thresholds for new calls (i.e., both voice and data calls) are set at K_v and K_d, respectively. However, since data calls can be vertically handed over from other networks, the CAC mechanism prioritizes these vertical handoff calls by using the threshold K_{vd}, where $K_d \leq K_{vd}$. Therefore, $K_{vd} - K_d$ channels are reserved for both horizontal and vertical handoff calls, and $C - K_v$ channels are reserved for horizontal handoff calls. With these thresholds, the priority of a horizontal handoff call is the highest, and the priority of a new call is the lowest. Again, voice

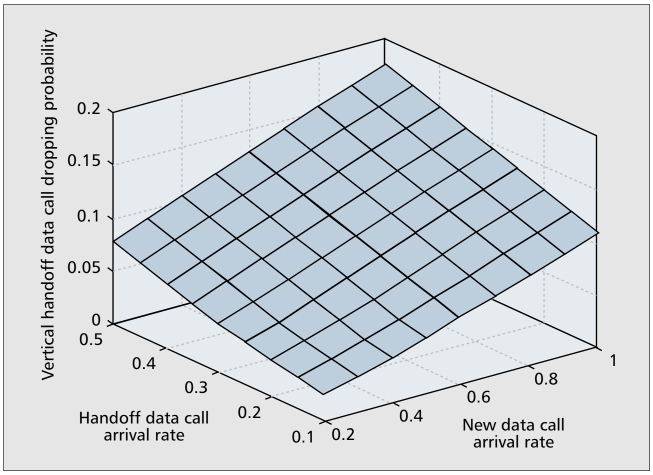

Figure 2. *Vertical handoff data call dropping probability under various new and handoff data call arrival rates.*

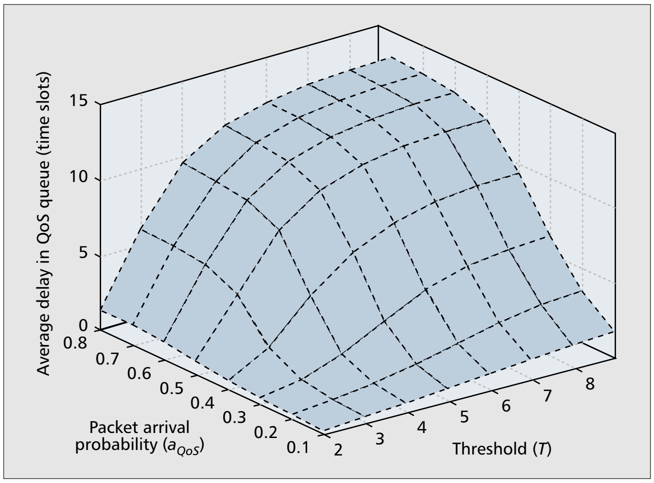

Figure 3. *Average delay for packets in the QoS queue.*

calls are prioritized over data calls by limiting the number of accepted data calls when the number of ongoing calls is equal to or greater than threshold C_d ($K_d \leq K_{vd} \leq C_d$).

The ABA algorithm includes mechanisms to increase and decrease the amount of bandwidth allocated to data calls. These mechanisms work as follows: When either a voice or data call arrives, if sufficient resources are not available, some of the ongoing data calls (randomly chosen) are downgraded so that the required amount of resources can be allocated to the incoming call. Similarly, when a call departs, the bandwidth freed will be randomly assigned to upgrade ongoing data calls. However, if there is no ongoing data call that can be downgraded, the incoming call is rejected.

Under the above assumptions on multiple types of calls, ABA, and vertical handoff, the system can be modeled by using a continuous-time Markov chain. The arrivals of new voice calls, handoff voice calls, new data calls, and horizontal and vertical handoff data calls are assumed to follow a Poisson process. We assume that the channel holding times for voice and data calls are exponentially distributed. The state space of the system is $\Phi = \{(V_n, D_n), 0 \leq V_n \leq C, 0 \leq D_n \leq C_d\}$, $n > 0$, where V_n and D_n are the number of ongoing voice and data calls, respectively.

The call-level performances of the system can be determined from the steady state probabilities. From this model we can obtain new call blocking and handoff call dropping probabilities for both voice and data calls as well as the vertical handoff call dropping probabilities for data calls.

For performance evaluation we assume that there are 40 channels in a cell, and the average channel holding times for voice and data calls are 5 and 10 min, respectively. We set the values for the different thresholds as follows: $K_v = 36$, $C_d = 36$, $K_{vd} = 34$, and $K_v = 32$. Figure 2 shows typical variations in vertical handoff call dropping probability (for data calls) under different data call arrival rates. This dropping probability also increases with increasing arrival rate, but at a slower rate than that for a voice call since the data call arrival rate is smaller than that for voice calls.

Tier II: The CAC Scheme in the Wired Part — As shown in Fig. 1, the DiffServ-aware edge router has two transmission queues: QoS and best effort (BE) with size U and V packets, respectively. The QoS queue is used for voice packets, while the BE queue is used for data packets. A CAC mechanism is applied at the QoS queue to guarantee packet-level QoS.

We assume that the router serves the queues in a time-division multiplexing fashion using fixed-size time slots, and only one packet is transmitted during one time slot. The router uses a fair scheduling mechanism based on the packetized version of generalized processor sharing (GPS). With this type of traffic scheduling, fairness is maintained in the sense that the packets in one queue will not affect the performance of those in the other. The amount of service for queue i, $S_i(t_1, t_2)$ ($i \in \{QoS, BE\}$), in time $[t_1, t_2)$ is governed by the weight ϕ_i, and for the two queues in Fig. 1, if both the queues are backlogged during period $[t_1, t_2)$, the following property is maintained:

$$\frac{S_{QoS}(t_1 t_2)}{S_{BE}(t_1 t_2)} = \frac{\phi_{QoS}}{\phi_{BE}}, \qquad (6)$$

where ϕ_{QoS} and ϕ_{BE} are the weights for the QoS and BE queues, respectively. With the work conserving property of fair scheduling, if one queue is not backlogged, the available service will be allocated to the other queue.

With this system model, the state space of the system can be defined as $\Xi = \{(X_n, Y_n, Z_n), 0 \leq X_n \leq U, 0 \leq Y_n \leq V, 0 \leq Z_n \leq T\}$, where X_n, Y_n, and Z_n represent the number of packets in the QoS queue, the number of packets in the BE queue, and the number of calls admitted into the QoS queue, respectively.

A threshold-based CAC mechanism is used to ensure that for QoS traffic the packet-level performance measures (e.g., packet dropping probability and average delay) are maintained at an acceptable level. Threshold T limits the number of calls admitted into the QoS queue. Specifically, when a call arrives, the CAC algorithm checks whether the number of ongoing calls is less than the threshold. If so, the new call is admitted; otherwise, the call is rejected. *This threshold can be set according to the congestion condition in the DiffServ domain*.

We assume that call arrival follows a Poisson process, and the channel holding time is exponentially distributed. For each flow the packet arrival follows a Bernoulli process with the probability of arrival of one packet in a time slot being a_{QoS}. We assume that probability a_{QoS} is the same for all QoS-sensitive flows. For the BE queue, we consider a batch Bernoulli process in which the probability that i packets arrive in one time slot is given by a_{BE}^{i}.

We obtain the packet-level QoS measures from the model. The length of a time slot is assumed to be 1 ms, and the weights for the queues are set to $\phi_{QoS} = 0.7$ and $\phi_{BE} = 0.3$. We vary the probability of arrival of one packet in the BE queue (a_{BE}^{1}) and obtain the performance results. Figure 3 shows typical variations in average delay D_{QoS} when the buffer size for each of the QoS and BE queues is 20 packets, and packet arrival probability at the BE queue is 0.2.

Conclusions

We have presented a comprehensive survey on the issues related to and solution approaches for the call admission control problem in the next-generation (e.g., 4G) wireless networks. Starting with the general model and classifications of the CAC strategies, different CAC schemes proposed in the literature have been reviewed, and the challenges in designing efficient CAC schemes for 4G systems have been outlined.

We have introduced a two-tier CAC architecture for 4G networks to ensure QoS in both the wireless and wired parts. In the general architecture, the CAC decision is based on both call-level and packet-level performance metrics. Based on this architecture, we have given an example of a two-tier CAC scheme considering two types of services (voice and data). Analytical models for performance evaluation of the proposed CAC scheme have been outlined, and typical numerical results have been presented.

Acknowledgment

This work was supported by a scholarship from TR*Labs*, Winnipeg, Canada. We would like to thank the reviewers for their valuable comments.

References

[1] L. Badia, M. Zorzi, and A. Gazzini, "A Model for Threshold Comparison Call Admission Control in Third Generation Cellular Systems," *Proc. IEEE ICC '03*, vol. 3, May 2003, pp. 1664–68.

[2] T. Zhang *et al.*, "Local Predictive Resource Reservation for Handoff in Multimedia Wireless IP Networks," *IEEE JSAC*, vol. 19, no. 10, Oct. 2001, pp. 1931–41.

[3] R. Jain and E. W. Knightly, "A Framework for Design and Evaluation of Admission Control Algorithms in Multi-Service Mobile Networks," *Proc. IEEE INFOCOM '99*, Mar. 1999, pp. 1027–35.

[4] D. Hong and S. S. Rappaport, "Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures," *IEEE Trans. Vehic. Tech.*, vol. VT-35, no. 3, Aug. 1986, pp. 77–92.

[5] R. Ramjee, R. Nagarajan, and D. Towsley, "On Optimal Call Admission Control in Cellular Networks," *Proc. IEEE INFOCOM '96*, vol. 1, Mar. 1996, pp. 43–50.

[6] B. Epstein and M. Schwartz, "Predictive QoS-based Admission Control for Multiclass Traffic in Cellular Wireless Networks," *IEEE JSAC*, vol. 18, Mar. 2000, pp. 523–34.

[7] D. A. Levine, I. F. Akyildiz, and M. Naghshineh, "A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks Using the Shadow Cluster Concept," *IEEE/ACM Trans. Net.*, vol. 5, Feb. 1997, pp. 1–12.

[8] J. Hou, J. Yang, and S. Papavassiliou, "Integration of Pricing with Call Admission Control to Meet QoS Requirements in Cellular Networks," *IEEE Trans. Parallel and Distrib. Sys.*, vol. 19, no. 9, Sept. 2002, pp. 898–910.

[9] J. Zhang *et al.*, "Resource Management in the Next-Generation DS-CDMA Cellular Networks," *IEEE Wireless Commun.*, vol. 11, no. 4, Aug. 2004, pp. 52–58.

[10] S. Blake *et al.*, "An Architecture for Differentiated Services," IETF RFC 2475, 1998.

[11] C. T. Chou and K. G. Shin, "Analysis of Adaptive Bandwidth Allocation in Wireless Networks with Multilevel Degradable Quality of Service," *IEEE Trans. Mobile Comp.*, vol. 3, no. 1, Jan.–Mar. 2004, pp. 5–17.

[12] G. Carneiro, J. Ruela, and M. Ricordo, "Cross-layer Design in 4G Wireless Terminals," *IEEE Wireless Commun.*, vol. 11, no. 2, Apr. 2004, pp. 7–13.

Biographies

DUSIT NIYATO [S'05] is an M.Sc. student in the Department of Electrical and Computer Engineering at the University of Manitoba and a researcher at TR*Labs*, Winnipeg, Canada. He received his B.Sc. in electrical engineering from King Mongkut's Institute of Technology Ladkrabang, Thailand, in 1999. From 1999 to 2003, he worked as a researcher in Embedded Systems Labs, Thailand. His main research interests are in the area of call admission control and performance modeling for 4G and broadband wireless networks.

EKRAM HOSSAIN [S'98, M'01] is an assistant professor in the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Canada. He received his Ph.D. in electrical engineering from the University of Victoria, Canada, in 2000. His research interests include wireless mobile communications and networking, distributed systems, and computer communication. Currently he serves as an Editor for *IEEE Transactions on Wireless Communications*, *IEEE/KICS Journal of Communications and Networks*, and *Wireless Communications and Mobile Computing Journal* (Wiley InterScience).