WIRELESS COMMUNICATIONS AND MOBILE COMPUTING
Wirel. Commun. Mob. Comput. 2003; **3**:895–906 (DOI: 10.1002/wcm.179)

A QoS-based charging and resource allocation framework for next generation wireless networks

Walid Ibrahim*,†, John W. Chinneck and Shalini Periyalwar
Systems and Computer Engineering, Carleton University, Ottawa, Ontario, K1S 5B6, Canada

SUMMARY

Wireless networks are evolving to include Internet access to interactive multimedia and video conferencing as well as traditional services such as voice, email and web access. These new applications can demand large amounts of network resources, such as bandwidth, to achieve the highest levels of quality (e.g. picture quality). In conjunction with this trend, charging and resource allocation systems must evolve to explicitly consider the trade-off between resource consumption and the Quality of Service (QoS) provided. This paper proposes a novel QoS-based charging and resource allocation framework. The framework allocates resources to customers based on their QoS perceptions and requirements, thereby charging fairly while improving resource allocation efficiency. It also allows the network operators to pursue a wide variety of policy options, including maximizing revenue or using auction or utility-based pricing. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: wireless; QoS; customer's satisfaction; charging; resource allocation; call admission control

1. Introduction

Designing the proper charging and resource allocation policy is crucial to the successful deployment of a telecommunication network. Such policies allocate network resources and recover costs fairly and competitively from the diverse population of customers. By tuning the charging and resource allocation policy, a service provider can attract new customers, compete with other service providers, and introduce new services and promotions.

Wireless networks are evolving to include Internet access to interactive multimedia and video conferencing as well as traditional services such as voice, email and web access. These new multimedia applications are very sensitive to the Quality of Service (QoS) provided. However, no QoS-based charging and resource allocation policies have yet been proposed for this wide range of heterogeneous applications.

Wireless service providers currently charge voice customers on a per-minute basis that gives customers a limited amount of access time for a predefined fixed price. Data customers are charged based on the amount of traffic they transmit/receive per month, regardless of their application type or the amount of resources they consume. Neither of these policies is suitable for the new generation of QoS-sensitive applications. Existing wire-line multi-service charging techniques [4,8,15,16,18,21] are also not suitable for wireless networks due to the fundamental differences between the wireline and wireless environments (e.g. limited resources, mobility, handoff etc.).

As a further complication, different customers are believed to have different QoS perception levels for

*Correspondence to: Walid Ibrahim, Systems and Computer Engineering, Carleton University, Ottawa, Ontario, K1S 5B6, Canada.
†E-mail: wibrahim@sce.carleton.ca

Copyright © 2003 John Wiley & Sons, Ltd.

the same multimedia application. For instance, a black and white video session might be very satisfactory for one customer and unacceptable for another. The value of information is also variable across and within the applications. An audio file download might be more valuable than an email while a business email might not be as valuable as a personal message.

Unfortunately, proposed multi-service wireless resource allocation policies [1,3,5,13,14,17,20], treat customers similarly despite the expected variance in their satisfaction levels. Treating customers equally wastes scarce network resources by allocating unneeded resources to satisfied customers while some customers receive unacceptable levels of network service. This can lead to the loss of business opportunities for the service provider since some customers would have paid more in order to receive a higher QoS satisfaction level. In order to support different types of applications with different QoS requirements, new charging and resource allocation policies should be able to allocate resources to customers based on their QoS requirements and charge them accordingly.

Although allocating resources based on customer QoS perception can theoretically improve the resource allocation efficiency, there is a danger that they will overbook network resources in order to guarantee their maximum QoS satisfaction. To prevent this, charging and resource allocation policies should give customers a strong incentive to reveal their true QoS requirements.

A new QoS-based integrated charging and resource allocation framework is proposed in Section 2. To demonstrate its flexibility, the framework is used to develop three specific new resource allocation policies in Section 3, each having different price allocation schemes, resource allocation schemes and admission control schemes. These new policies are compared with other currently proposed resource allocation policies in Section 4.

2. Integrated Charging and Resource Allocation Framework

This paper introduces a new integrated charging and resource allocation (ICRA) framework that takes into consideration the customer's QoS perception. Customers define their QoS requirements and are charged accordingly, as described in Sections 2.1 and 2.2. The proposed framework allows service providers to introduce a wide new range of fair and competitive services while allowing customers to select the level of service that best fits their needs. The framework uses a distributed approach: it operates at the individual cell level, allocating bandwidth and other limited resources within the cell. This reduces process complexity and inter-cell communication. The distributed approach also helps the service providers to deploy and upgrade the framework gradually without affecting the rest of the network.

The 'Bandwidth Market Price' (BMP) is defined as the current price to transmit 1 Gbit of traffic using 1 kbps of network bandwidth. In the new framework, customers are charged based on the BMP and the amount of bandwidth they need to reach a specific QoS satisfaction level. 'Customer' may imply an individual or an organization when the QoS satisfaction profile is built for a group of people within the organization.

The framework allows service providers to gain access to the customer's different QoS satisfaction levels and the maximum price they are willing to pay in order to reach each level. It thereby gives them a high degree of network management flexibility. As demonstrated later in Section 3, the framework allows service providers to create a custom ICRA policy that best fits their marketing strategy.

To ensure a minimal QoS level for low budget calls, the framework provides network operators with a set of tuning parameters such as maximum allowable call blocking ratio (CBR) and call dropping ratio (CDR). For example, the service provider could use the maximum CDR and/or CBR to ensure that high bid customers do not force the CDR and/or CBR beyond a certain threshold. The framework also allows the service provider to specify a minimum BMP to prevent customers from dragging the BMP below a certain limit.

There are two main types of resources in wireless networks: signal power and network bandwidth. The signal power and the customer's signal quality requirements and location determine the maximum data rate the customer may receive. In 3G wireless systems such as 1xEVDO [2] and HSDPA [6], the forward link is a 'fat pipe' with transmission at full power to every user. In this case power control does not apply, and the user data rate is determined by matching the modulation and code rate to the perceived signal quality, which may vary with distance, speed etc. The data rate a user can receive in a given instance may be greater than, equal to or less than the minimum rate needed by the application. The system exploits favorable signal quality conditions and transmits only to users who can achieve the desired rate. In this paper, it is implicitly assumed that users can be given their desired rate by scheduling them at instances when they can achieve it.

Copyright © 2003 John Wiley & Sons, Ltd.

Power control is not considered since a 'fat-pipe' transmission is assumed. A similar simplifying assumption is made for the reverse link as well.

2.1. QoS-Based Charging Scheme

To charge customers fairly based on the amount of network resource they consume, we introduce a new QoS-based charging scheme that takes into consideration the QoS parameters specified by the customers during call setup. It is a generic charging scheme that has a tariff component for each QoS parameter supported by the network. The charging scheme supports two grades of services. The basic service supports only the customer bitrate (BR). More advanced services could support the BR plus a guaranteed maximum end-to-end delay (ETED) and/or frame error rate (FER), as shown in Equation (1).

$$\mathrm{T} = \mathrm{T_B} + \mathrm{T_E} + \mathrm{T_F}\ \text{¢/Gbit of transmitted traffic} \quad (1)$$

T is the total tariff in ¢/Gbit of transmitted traffic, $\mathrm{T_B}$ is the BR tariff, $\mathrm{T_E}$ is the ETED tariff and $\mathrm{T_F}$ is the FER tariff. The BR tariff component is directly proportional to the BR assigned to the customer (Kbps) and the BMP (¢/Gbit/Kbps), as shown in Equation (2).

$$\mathrm{T_B} = \mathrm{BR} \cdot \mathrm{BMP} \quad (2)$$

In addition to the bitrate, real-time applications with tight delivery requirements may ask for a guaranteed ETED. For the forward link traffic, we are only interested in the delay portion from the time a data packet arrives at the base station until it is delivered to the mobile terminal. This delay is dominated by the scheduling delay at the MAC layer while a packet waits for a time slot(s) to be transmitted. For real-time applications, the MAC layer scheduler may reserve a periodic time slot on the access channel or grant time slots on a priority packet-by-packet basis. The charging scheme charges customers with bounded delay time an extra premium ($\mathrm{T_E}$) to offset the social cost they impose on the other active calls by reserving a transmission time slot or asking for a higher scheduling priority.

Similar to $\mathrm{T_B}$, the $\mathrm{T_E}$ is directly proportional to the bitrate. An exact calculation of $\mathrm{T_E}$ requires pricing the MAC layer time slots and calculating the number of time slots consumed by each call. For simplicity, $\mathrm{T_E}$ is calculated as a fraction of the basic service, as shown in Equation (3).

$$\mathrm{T_E} = \alpha \cdot \mathrm{T_B} = \alpha \cdot \mathrm{BR} \cdot \mathrm{BMP} \quad (3)$$

Table I. Example ETED tariff parameters.

ETED	ETED < 150	ETED < 200	ETED < 250	ETED < 300
α	0.4	0.3	0.25	0.2

α is a service provider configurable parameter that is inversely proportional to the ETED value specified by the customer. Service providers may adjust it based on their marketing strategy, e.g. as shown in Table I.

Like ETED, applications with bounded FER requirements may ask for a guaranteed maximum FER. In the wireless environment, a forward error correction (FEC), automatic repeat request (ARQ) or a hybrid FEC/ARQ policy can be used to control the call FER. The customer signal quality and hence the amount of resources consumed to guarantee a maximum FER depends to a great extent on the customer location and the surrounding environment (e.g. fading and interference). Charging customers based on the amount of resources consumed to guarantee a maximum FER is unfair. It is unfair to charge customers more because they are far from the base station or to overcome the interference they get from other customers. As a solution, the service provider should charge calls with guaranteed FER requirements an extra percentage of the basic service tariff that depends only on the customer's maximum allowable FER.

$$\mathrm{T_F} = \gamma \cdot \mathrm{T_B} = \gamma \cdot \mathrm{BR} \cdot \mathrm{BMP} \quad (4)$$

γ is a service provider configurable parameter that is inversely proportional to the FER parameter specified by the customer. Table II shows example settings for γ. The generic call tariff can now be defined as:

$$\mathrm{T} = \mathrm{BR} \cdot \mathrm{BMP} + \alpha \cdot \mathrm{BR} \cdot \mathrm{BMP} + \gamma \cdot \mathrm{BR} \cdot \mathrm{BMP}$$
$$\mathrm{T} = \mathrm{BR} \cdot \mathrm{BMP}(1 + \alpha + \gamma) \quad (5)$$

2.2. QoS Profile

The concept of the QoS profile is introduced here as a way to capture the customer's QoS perception and the price that they are willing to pay to achieve various levels of QoS satisfaction. It provides customers with

Table II. Example FER tariff parameters.

FER	FER = 0%	FER < 1%	FER < 2%	FER < 3%
γ	0.4	0.3	0.5	0.15

Copyright © 2003 John Wiley & Sons, Ltd.

Table III. A QoS profile example.

Satisfaction level	Customer QoS profile				
	QoS parameters				
	Bitrate	Delay	FER	Budget	Bid
Excellent	12.2 Kbps	150 msec	1%	1000	48.22
Good	10.2 Kbps	150 msec	1%	900	51.90
Fair	10.2 Kbps	250 msec	2%	800	54.09
Poor	7.4 Kbps	300 msec	3%	600	60.06

a way to initialize call QoS parameters such as bitrate (BR), maximum allowed end-to-end delay (ETED) and frame error rate (FER) based on their QoS perception. The QoS profile allows customers to define multiple satisfaction level records. Defining multiple records allows customers to compete for a lower satisfaction level if they cannot afford the price required to maintain the current satisfaction level. In this paper, customers are limited to four entries per QoS profile corresponding to their 'excellent', 'good', 'fair' and 'poor' satisfaction levels (Table III).

Since customers are expected to value the QoS parameters differently, and to assign different budgets, the framework compares customers based on the *bid* per kbit of allocated bandwidth. Equation (6) shows how the bid is derived from the budget and customer's QoS parameters.

$$bid = \frac{bud}{BR \cdot (1 + \alpha + \gamma)} \quad \frac{¢}{\text{Gbit} \times \text{Kbps}} \tag{6}$$

Customers are expected to assign lower budgets for lower satisfaction levels. However, they should adjust their budgets such that their bids increase as the satisfaction level decreases. Increasing the bids as the satisfaction level falls is essential so the call can compete effectively as network congestion increases (i.e. the budget and QoS parameters should provide a higher ¢/Gbit/kbps rate as the satisfaction level decreases). See Table III for an example of QoS profile.

Although it seems difficult for the average customer to construct an efficient QoS profile, providing customers with default settings for different types of applications and user-friendly interfaces for tuning settings will simplify the process. The user interface will help customers tune their settings within the acceptable limits. Settings outside the acceptable limits will be rejected by the system. Customers may have to tune the default settings only once before they run their applications for the first time. A customer can also store multiple QoS profiles to run the same application with different QoS settings on different occasions (e.g. weekdays, weeknights, weekends).

The QoS profile has several other benefits in addition to gathering customer QoS preferences. It reduces control and negotiation messaging overhead by transmitting the customer's QoS profile only once during call setup and allows customers to define the preferred way to degrade service in case of a resource crisis. Finally, it allows customers to guarantee a certain satisfaction level by assigning an infinite budget to that level. In this case, the customer commits to pay the current BMP in order to maintain the specified satisfaction level.

3. Proposed ICRA Policies

To demonstrate the flexibility of the framework, we developed three different ICRA policies: revenue-based auction-based and utility-based policies.

3.1. Revenue-Based ICRA Policy

The objective of the revenue-based policy is to maximize the service provider revenue by dynamically adjusting the BMP and the amount of bandwidth allocated to each customer. This objective can be stated as a mixed integer non-linear programming problem as shown in Equation (7), where Z is the service provider revenue and Φ is the set of existing calls.

$$\text{Max } Z = \sum_i BMP \cdot BR_i \cdot (1 + \alpha + \gamma) \quad \forall i \in \Phi \tag{7}$$

Due to the non-linearity of the objective function, finding the globally optimal revenue is extremely difficult. We use a heuristic price adjustment scheme to reduce the complexity of the problem to a mixed integer linear programming problem and then use another heuristic scheme to solve the resource allocation problem.

The aim of the price adjustment heuristic (Fig. 1) is to find a good (hopefully optimal) BMP that maximizes the network revenue by trading off network utilization and BMP. Increasing the BMP is not always the way to increase network revenue. A high BMP generates more revenue per unit of allocated bandwidth, but it might also lead to lower network utilization and hence, lower overall revenue.

At a given level of bandwidth usage, the network revenue is directly proportional to the BMP

Copyright © 2003 John Wiley & Sons, Ltd.

Inputs:

- QoS profile for each customer.
- bw^{max}: base-station bandwidth capacity.
- BMP^{min}: The service provider minimum acceptable BMP.

1. Set $R = 0$, bid list = ϕ.
2. For each customer i:
   - 2.1. For each satisfaction level j:
     - 2.1.1. Calculate the associated bid_{ij}.
     - 2.1.2. Discard bid_{ij} if it will force any existing call to terminate.
     - 2.1.3. If $bid_{ij} \geq BMP^{min}$ then add bid_{ij} to the bid list
3. Sort the bid list in ascending order.
4. For each bid_{ij} in the list
   - 4.1. Calculate the total customer bandwidth demand (bw)
   - 4.2. If $bw > bw^{max}$ then set $bw = bw^{max}$
   - 4.3. Calculate the network revenue based on Equation 7
   - 4.4. If revenue > R then R = revenue and BMP = bid_{ij}

Fig. 1. Revenue-based price adjustment scheme.

(Equation (7)), so network revenue increases with BMP until one of the customer's budget constraints is activated. Any increase in the BMP beyond this price forces that customer into a lower QoS satisfaction level. Consequently, the network utilization will drop along with the network revenue. If the BMP continues to increase, the revenue will again increase until the next budget constraint is activated. Therefore, the points at which the BMP equals one of the customer's bids are local maximum points. Any infinitesimal increase in the BMP above such a local maximum point leads to a local minimum point as one of the calls becomes unable to afford its current satisfaction level.

Since the local maximum points are well defined, the heuristic scheme identifies a good BMP by selecting the local maximum point with the highest generated revenue. In the worst case, the heuristic sets the BMP in a polynomial time of order 4N, where N is the number of active calls, simply by checking all of the satisfaction levels for all of the current users. The number of points searched can be reduced by neglecting all local maximum points with BMP lower than the service provider minimum acceptable BMP or which will force any existing call to termination.

During low traffic periods, the BMP will never go higher than the lowest bid associated with an excellent satisfaction level. During high traffic periods, it will never go higher than the lowest bids associated with a poor satisfaction level. Since customers are expected to assign higher budgets to higher satisfaction levels, the price adjustment scheme will continuously try to increase the existing call's satisfaction level to maximize the network revenue.

After setting the BMP, the resource allocation scheme calculates the amount of bandwidth for each customer based on the new BMP and the customer's QoS profiles. This can be modeled as a mixed integer linear programming problem and solved using a third party mathematical programming solver. While this approach has the advantage of truly maximizing the network revenue given the BMP, it can be very time-consuming and is therefore impractical in a real-time system.

We develop an efficient heuristic to solve the bandwidth allocation problem (Fig. 2). The allocation decision depends heavily on the amount of bandwidth available compared to the new total bandwidth demand. Given the BMP, the heuristic calculates the new total bandwidth demand by summing the maximum affordable bandwidth for all of the customers. If the available bandwidth at the base station is higher than the total bandwidth demand, all customers are allocated their maximum affordable bandwidth. Otherwise, we determine how much bandwidth to allocate to each customer using a sorted best-fit heuristic that tries to maximize the network revenue.

The heuristic operates recursively. First, identify the call that needs the most bandwidth to reach the next affordable satisfaction level. Then, update the amount of bandwidth allocated to the call, the call satisfaction level and the amount of resource available at the base station. Begin a new iteration. The scheme stops when all customers reach the excellent satisfaction level or when the amount of free bandwidth left is insufficient to upgrade any of the existing calls to the next higher affordable satisfaction level. To minimize the CDR, the resource allocation scheme can be modified to initialize each active call with enough resource to reach the poor satisfaction level. The scheme can also minimize the CBR by initializing an admitting call at the poor satisfaction level as well.

Inputs:

- Current BMP.
- QoS profile for each customer i.
- bw^{max}: Base station bandwidth capacity.

1. Set call list = ϕ.
2. Based on the current BMP, calculate the total bandwidth demand (bw).
3. If $bw < bw^{max}$, then allocate the maximum affordable bandwidth to each call and exit.
4. For each call i:
   - 4.1. Get the call next higher satisfaction level (j)
   - 4.2. If $bid_{ij} <$ BMP, then discard the call.
   - 4.3. Calculate the call's bandwidth demand to reach satisfaction level j.
   - 4.4. If there is not enough bandwidth to satisfy the call demand, then discard the call.
   - 4.5. Add the call to the call list
5. If the call list is empty, then exit
6. Select the call with the highest bandwidth demand.
7. Update the selected call allocated bw, satisfaction level, the amount of free bandwidth
8. Go to step 4

Fig. 2. Revenue-based best-fit resource allocation scheme.

Copyright © 2003 John Wiley & Sons, Ltd.

When a new/handoff call admission request arrives, an admission control scheme decides whether to accept or reject it. Such a scheme is traditionally used to limit the number of calls admitted to the network in order to guarantee the QoS offered to active calls and/or to minimize the handoff dropping probability [3,5,13,17,20]. In contrast, the revenue-based admission control scheme (Fig. 3) accepts a call only if the optimal projected revenue after the call is admitted is higher than the current revenue by a margin set by the service provider. This minimum revenue increase percentage parameter limits the effect of admitting a new call on the QoS offered to existing calls.

Using the price adjustment and resource allocation schemes, the admission control scheme compares the current revenue with the projected revenue if the admission request is accepted. If the projected revenue does not exceed the minimal revenue increase threshold, then the admission request is rejected. In order to control the QoS offered to existing calls, the service provider can configure the scheme to check the admitting call's highest affordable satisfaction level. If it is lower than a threshold set by the service provider, the admission request is also rejected. The admission request is also rejected if it forces the CDR above the maximum allowable threshold. Finally, before blocking a new call the admission scheme checks whether the new CBR exceeds the maximum allowable threshold. If so, the call is admitted unless admitting the call will force the CDR above the maximum threshold.

Inputs:

- QoS profile for each customer.
- R: current network revenue
- bw^{max}: base-station bandwidth capacity.
- *X*: service provider minimum revenue increase %.
- CDR^{max} : maximum call dropping ratio.
- CBR^{max}: maximum call blocking ratio.

1. Sum the poor satisfaction level bandwidth demand over all calls (bw^{pr}).
2. If $bw^{pr} > bw^{max}$, then reject the admission request and exit.
3. Get the projected BMP and resource allocation.
4. Calculate the projected CDR if the admission request is accepted.
5. If the projected CDR > CDR^{max}, then reject the admission request and exit.
6. Calculate the projected CBR if the admission request is rejected.
7. If the projected CBR > CBR^{max}, then accept the admission request and exit.
8. Calculate the network projected revenue (R^*) based on Equation 7.
9. If $R^* < R\ (1 + X/100)$, then reject the admission request and exit.
10. If the admitting call maximum affordable satisfaction level < the service provider threshold, then reject the admission request and exit.
11. Accept the admission request.

Fig. 3. Revenue-based admission control scheme.

Since dropping an ongoing call is more serious than blocking a new one, the admission control scheme treats handoff calls differently. No resources are reserved for handoff calls. Instead, it readjusts the resources allocated to active calls in order to free enough resources to reach the handoff call's poor satisfaction threshold at least. A handoff call admission request is accepted if the target cell has enough bandwidth to guarantee at least the poor satisfaction threshold to the existing active calls in addition to the handed-off one. The admission control scheme will try to maintain the CDR and the CBR below the maximum thresholds as long as there is enough bandwidth to support the call's poor satisfaction level bandwidth demand.

3.2. Auction-Based ICRA Policy

The auction-based policy allocates network bandwidth using an auction-based competitive market model in which customers compete for the available bandwidth. We think that an auction is a good way to introduce new services in a competitive environment where buyers frequently know more than the seller about the value of the service. The seller is reluctant to suggest a price first, out of fear that his ignorance will prove costly and so holds an auction to extract payment he might not otherwise realize. An auction is a simple way to determine market-based prices. It is efficient in the sense that an auction usually ensures that resources are allocated to those who value them most highly and ensures also that sellers receive the collective assessment of the value.

We have selected a general form of uniform second-price auction in which the M highest bidders win and pay a uniform price equal to the $M+1$st highest bid [12]. In a uniform second-price auction, no one is discouraged out of fear that they will pay too high a price. Aggressive bidders receive sure and certain awards but pay a price closer to market consensus. The price that the winning bidder pays is determined by competitor's bids alone and does not depend upon any action the bidder undertakes.

The resource allocation scheme (Fig. 4) implements an $M+1$st price auction to sell the available bandwidth to competing customers. The scheme first creates an array of records to collect the customer bids. Each customer has four records corresponding to their four satisfaction levels and each record has the following fields: the satisfaction level, the bandwidth demand and the associated bid. The records are sorted in a descending order based on the bid. Records with

Copyright © 2003 John Wiley & Sons, Ltd.

Inputs:

- For each call i, and satisfaction level j:
  - i. bid_{ij}: the satisfaction level bid.
  - ii. bw_{ij}: the satisfaction level bandwidth demand
- bw^{max}: base-station bandwidth capacity.
- BMP^{min}: service provider minimum acceptable BMP.

1. Set $X = 0$, record list = ϕ, BMP = ∞, call's current bandwidth $bw_i = 0$.
2. For each call i:
   2.1. For each satisfaction level j:
      2.1.1. If $bid_{ij} \geq BMP^{min}$ then add bid_{ij} and bw_{ij} to the record list
3. Sort the record list by descending order of bid value, sub-ordered on ascending value of satisfaction, sub-ordered on ascending value of bandwidth.
4. For each record in the list:
   4.1. If $X + bw_{ij} - bw_i \leq bw^{max}$ then:
      4.1.1. $X = X + (bw_{ij} - bw_i)$.
      4.1.2. $bw_i = bw_{ij}$.
      4.1.3. $BMP = bid_{ij}$

Fig. 4. Auction-based resource allocation scheme.

Inputs:

- QoS profile for each customer.
- bw^{max}: base-station bandwidth capacity.
- CDR^{max}: maximum call dropping ratio.
- CBR^{max}: maximum call blocking ratio.

1. Sum the call's poor satisfaction level bandwidth demand (bw^{pr}).
2. If $bw^{pr} > bw^{max}$, then reject the admission request and exit
3. Get the projected BMP and resource allocation.
4. Calculate the projected CDR if the admission request is accepted.
5. If the projected CDR > CDR^{max}, then reject the admission request and exit.
6. Calculate the projected CBR if the admission request is rejected.
7. If the projected CBR > CBR^{max}, then accept the admission request and exit.
8. If the admitting call maximum affordable satisfaction level < the service provider threshold, then reject the admission request and exit.
9. Accept the admission request

Fig. 5. Auction-based admission control scheme.

identical bids are sub-ordered in ascending order by satisfaction level (i.e. records with lower QoS satisfaction levels are higher in the list). Records with the same bid and satisfaction level are sub-ordered in ascending order based on the bandwidth demand.

The list of records is processed in order. For each record, check whether there is sufficient bandwidth available to satisfy the demand specified in the record. If so, the call's projected satisfaction level and the amount of available bandwidth are updated. The scheme terminates if the current record has a bid lower than the minimum BMP, all the active calls reach the excellent satisfaction level or insufficient bandwidth is available to improve the satisfaction level of any of the active calls.

Since customer bids increase as satisfaction levels decrease, the resource allocation scheme allocates bandwidth to satisfy each call's poor satisfaction level first. However, if a customer specifies an excellent satisfaction level bid higher than the bid specified by another customer for the poor satisfaction level, the first customer will reach the excellent satisfaction level before the second customer even reaches the poor satisfaction level. To guarantee call continuation (and hence minimize the CDR), the resource allocation scheme may allocate enough resource to active calls such that they all achieve their poor satisfaction level before beginning the resource auction. A second price auction is then held for the remaining bandwidth.

In order to recover the BMP, the policy makes a small modification to the $M + 1$st price auction rules. Instead of using the first unused bid as the auction price, the scheme uses the bid associated to the last used record as the projected BMP. This modification is required since some intermediate records might not be used because the amount of free bandwidth is insufficient to satisfy the record demand.

The admission control scheme (Fig. 5) uses the projected resource allocation and BMP to decide whether to accept or reject a call admission request. In order to guarantee the call continuation, a handoff call admission request is accepted if the resource allocation scheme can guarantee at least the poor satisfaction level to all active calls along with the handed-off call. A new call admission request is rejected if the call's bid and bandwidth demand are high enough to force one or more active calls to terminate. The admission request might also be rejected in the following two cases if the CBR is lower than the maximum threshold:

1. The call's highest bid is lower than the service provider minimum acceptable BMP.
2. The call's projected satisfaction level is lower than a minimum threshold set by the service provider.

3.3. Utility-Based ICRA Policy

The utility-based policy allocates network resources with the objective of maximizing the customer's *utility*, the value corresponding to the customer's satisfaction level as shown in Table IV. The numbers in Table IV are selected to show the increase in the customer's utility due to an increase in the satisfaction level, and are easily reset as desired by the network operators.

$$\text{Max} \quad Z = \sum_i u_i \quad \forall i \in \Phi,\ u_i \in (u^{xl},\ u^{gd},\ u^{fr},\ u^{pt}) \tag{8}$$

Where u_i denotes the customer's utility.

Copyright © 2003 John Wiley & Sons, Ltd.

Table IV. Customer's utility.

Satisfaction level	Excellent	Good	Fair	Poor
Utility	$u^{xl} = 4$	$u^{gd} = 3$	$u^{fr} = 2$	$u^{pr} = 1$

On arrival of a call admission request, the projected optimal bandwidth allocation, assuming that the call admission request is accepted, is calculated. The bandwidth allocation scheme supports two types of welfare fairness that are common in the literature: Utilitarian Criterion and Equality Criterion [7]. The utilitarian criterion, sometimes referred to as utility maximization, is a Pareto-Optimal allocation that results in the greatest sum of the customer utilities. The equality criterion is a Pareto-Optimal allocation that results in an equal level of utility for all customers.

The utility criterion algorithm operates recursively. First identify the call that needs the least bandwidth to reach the next affordable satisfaction level. The satisfaction level is affordable if its associated bid is not less than the service provider minimum acceptable BMP. The algorithm then updates the amount of bandwidth allocated to the call, the call satisfaction level and the amount of resource available at the base station, and begins a new iteration. The scheme stops when all customers reach the excellent satisfaction level or when the amount of free bandwidth left is insufficient to upgrade any of the existing calls to the next higher affordable satisfaction level. To minimize the CDR, the resource allocation scheme can be modified to initialize each active call with enough resource to reach the poor satisfaction level. The scheme can also minimize the CBR by initializing an admitting call at the poor satisfaction level as well.

Instead of increasing the average customer's satisfaction level, the equality criterion tries to allocate bandwidth to customers such that all of them have the same QoS satisfaction level. The bandwidth allocation algorithm operates in steps. Starting with the poor satisfaction level, it sets a target satisfaction level at each step and allocates bandwidth to calls so that each call reaches the target satisfaction level. A call is considered only if the bid associated with the target satisfaction level is not lower than the service provider minimum acceptable BMP. If enough bandwidth is available, the network updates the amount of resource assigned to each call, the call satisfaction levels and the amount of free bandwidth, and then advances the target satisfaction level.

If there is insufficient bandwidth, the scheme selects a subset of calls that will reach the target level. Two approaches can be used to select the subset. In the first approach, calls are prioritized based on the bid associated with the target level: bandwidth is allocated first to the call with the highest bid to insure that resources go to the call that most appreciates the QoS satisfaction level received. The second approach allocates the available resource such that the number of calls enjoying the target satisfaction level is maximized. Resources are allocated first to the call that requires the smallest amount of bandwidth to reach the target satisfaction level.

During high-traffic periods, the resource allocation is Pareto optimal since no call can improve its QoS satisfaction level without lowering the QoS satisfaction level received by another active call. After defining the projected Pareto optimal resource allocation, the projected BMP is set equal to the lowest bid associated with an active call at its current QoS satisfaction level.

To ensure call continuation, a handoff call is admitted if there is enough resource to guarantee at least poor QoS satisfaction level for all of the active calls in addition to the handed-off call. The admission control scheme uses the projected resource allocation and BMP to decide whether to accept or reject a new call admission request. The admission request is rejected if admitting the new call will force one or more of the active calls to terminate. The admission request might also be rejected in any of the following cases if the CBR is lower than the maximum threshold:

1. The projected BMP is lower than the network minimum BMP set by the service provider.
2. The call satisfaction level is lower than the minimum satisfaction level for admission.
3. The new value of the network objective function is lower than the current value (i.e. the new total utility is lower than the current value).

4. Experimental Results

The proposed polices are compared with the rate-based borrowing policy (rbbp) [5] via simulation. The rate-based borrowing policy is used for comparison for two reasons. First, it is one of the most recent policies proposed for third generation wireless networks. Second, it is similar to the proposed policy in various aspects such as making the resource allocation and admission control decisions based exclusively on local base station information, it handles several types

Copyright © 2003 John Wiley & Sons, Ltd.

of calls, differentiates between new and handoff calls, and controls both the CBR and CDR via service degradation.

4.1. Simulation Parameters

We model call arrivals as a Poisson process with a mean arrival rate λ that ranges from 1 call/min to 5 calls/min. Assuming medium call mobility, the probability that an arriving call is a handoff call is 0.5 [17]. There are five representative application types: narrowband audio, wideband audio, narrowband video, wideband video and data transfer. In References [5] and [17], the traffic types of new calls are assumed to have equal probabilities, i.e. 0.2 for each of the five types. However, we assume that voice calls will have a larger fraction of the market in the near future and so voice calls are assigned a probability of 0.3 while wideband video calls are assigned a probability of 0.1, with the other three types having probabilities of 0.2.

In addition to generating call arrivals and holding times, the traffic module assigns a random QoS profile to each call. The bitrate associated with each QoS satisfaction level is selected randomly based on the call type and the IEEE recommended QoS parameters for wireless broadband applications [9]. For realism, the selection of the audio QoS satisfaction levels is guided by the QoS Mean Opinion Score (MOS) in Reference [10], while the selection of the video QoS satisfaction levels is guided by the customer's MOS and the video profiles and levels in Reference [11]. Since the Lognormal distribution has been widely used to model personal income [19], we model the customer's excellent QoS budget as a lognormal distribution with a mean of 5 and a standard deviation of 1.

4.2. Performance Metrics

In addition to the widely used average CBR and CDR, performance is assessed using the following new performance metrics: (1) average revenue per unit of resource, (2) admitted calls average satisfaction level, (3) comprehensive grade of service and (4) comprehensive network satisfaction.

The network revenue is calculated by multiplying the BMP by the amount of resource consumed by the customers. The average revenue per unit of time ($\mathrm{Rev}_{\mathrm{avg}}$) is then calculated by dividing the total calculated revenue by the total simulation time. The CBR and CDR are calculated as the average ratio of calls blocked and dropped in the last 1000 calls. To calculate the admitted calls average satisfaction level, the total time the calls spent at each satisfaction level is measured and weighted based on the utility scale in Table IV. The admitted calls average satisfaction level is then found by dividing the total weighted time by the total call time. The comprehensive grade of service (GoS_{comp}) is calculated by including the weights associated with dropped (weight $= -2$) and blocked (weight $= -1$) calls. The negative weight is assigned to each dropped call, regardless of the satisfaction level it enjoyed prior to being dropped. The simulation results show that the proposed framework is insensitive to the assigned weights. For improved accuracy, a customer survey could be conducted to adjust the weight assigned to each satisfaction level.

In wireless networks, finding the optimal cell size is one of the critical network design issues. Adjusting the cell size affects the number of calls the cell will serve (i.e. the cell traffic load). Since the amount of resource per cell is fixed, increasing the traffic load may increase the cell's generated revenue, but it may also increase the CBR and CDR, thereby reducing the customers' satisfaction level. To help find the proper traffic load and hence cell size, we introduce Comprehensive network satisfaction (S_{comp}) as a new performance metric. The S_{comp} penalizes the network-generated revenue (service provider satisfaction) for any reduction in customer satisfaction levels. Since the maximum grade of service that the system can achieve is 4, we define the normalized comprehensive grade of service as $GoS'_{\mathrm{comp}} = GoS_{\mathrm{comp}}/4$. Comprehensive network satisfaction is then calculated as shown in Equation (9).

$$S_{\mathrm{comp}} = \mathrm{Rev}_{\mathrm{avg}} \times GoS_{\mathrm{comp}} = \mathrm{Rev}_{\mathrm{avg}} \times GoS_{\mathrm{comp}}/4 \tag{9}$$

4.3. Results

As shown in Figure 6, all of the proposed policies generate more revenue than the rbbp, with the revenue-based policy generating the most. It outperforms the rbbp by 48% at a traffic load of 1 call/min and by up to 68% at a traffic load of 5 calls/min. The proposed policies achieve a much lower CBR than the rbbp (Fig. 7). Instead of blocking an incoming call if the customer cannot afford the excellent satisfaction level, the proposed policies attract more customers by admitting low budget calls at a lower satisfaction level. They adjust the amount of resource allocated to existing customers in order to control the CBR.

While maximizing the network revenue, the revenue-based policy guides the BMP higher, and

Copyright © 2003 John Wiley & Sons, Ltd.

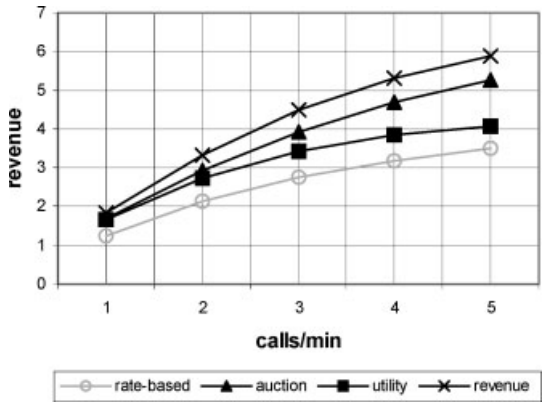

Fig. 6. Network revenue/unit of resource.

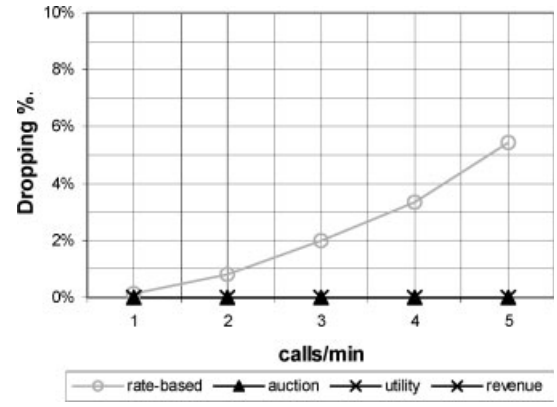

Fig. 8. Call dropping percentage.

thus blocks more calls than the utility and auction-based policies. The utility-based policy blocks more calls than the auction-based policy in order to maintain the satisfaction levels of existing calls. Figure 8 shows that the framework also significantly reduces the CDR. It does so by degrading the QoS provided to existing active calls and admitting handoff calls, if there are enough resources to satisfy their minimum QoS requirements.

The simulation results show that the utility-based policy provides customers with the highest average satisfaction level (Fig. 9). By reducing the CBR, the auction-based policy provides admitted calls with a lower average satisfaction level than the rbbp. Reducing the CBR increases the number of calls sharing the limited network resources and hence reduces the admitted calls average satisfaction level. While maximizing the network revenue, the revenue-based policy provides admitted calls with the worst average satisfaction level. When the dissatisfaction due to blocking and dropping calls is considered in the comprehensive grade of service, the auction-based

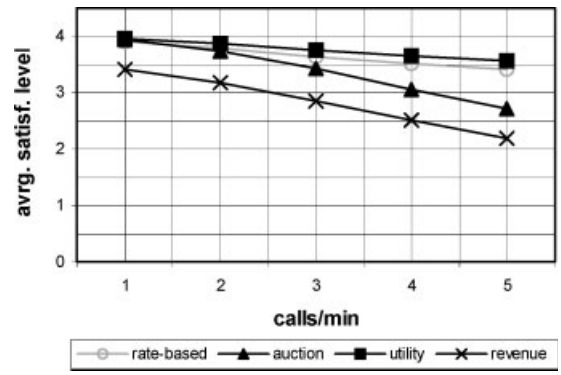

Fig. 9. Admitted calls average satisfaction level.

policy clearly outperforms the rbbp (Fig. 10). The rbbp slightly outperforms the revenue-based policy.

As described earlier, the comprehensive network satisfaction performance metric is very important for network design. The simulation results for S_{comp}, shown in Figure 11, show that the proposed policies perform much better than the rbbp when the comprehensive grade of service is assessed. The proposed

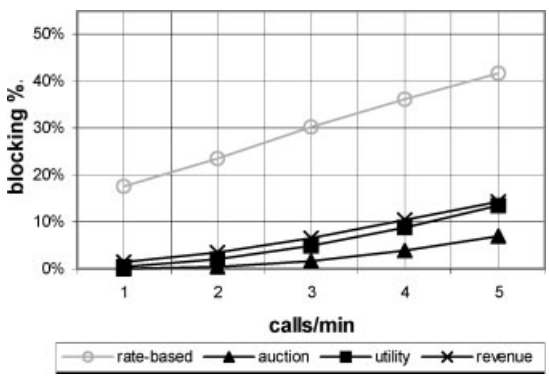

Fig. 7. Call blocking percentage.

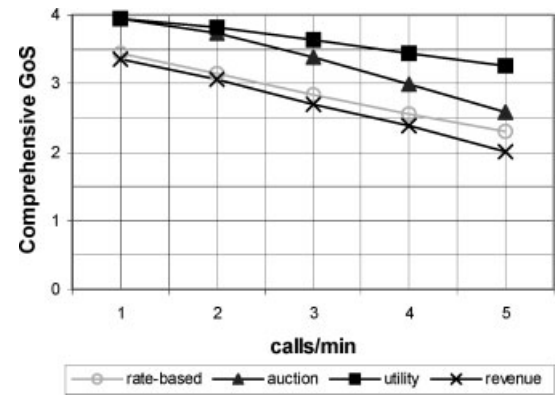

Fig. 10. Comprehensive grade of service.

Copyright © 2003 John Wiley & Sons, Ltd.

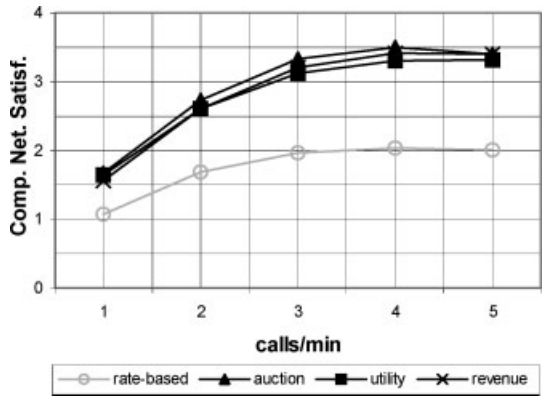

Fig. 11. Comprehensive network satisfaction.

policies are capable of generating more revenue without affecting the customer average satisfaction levels. Their performance improves as the traffic load increases to 4 calls/min. Beyond 4 calls/min, the decrease in the customer's average satisfaction level outweighs the increase in the network revenue.

5. Conclusions

This paper proposes a novel QoS-aware ICRA framework that provides soft QoS services in terms of customer QoS perception and provides adaptive resource use based on the traffic conditions and customer QoS perception. We also propose the QoS profile as a new way for customers to efficiently convey their QoS requirements. A new QoS-based charging scheme uses the QoS profile to charge customers dynamically based on the network conditions and the amount of resource they consume. This charging scheme persuades customers to reveal their true QoS requirements and increases the customer's confidence in the dynamic charging scheme by using their budgets as the maximum limit for the call price.

A major advantage of the proposed framework is the flexibility to support different ICRA policies, which allows service providers to implement the ICRA policy that best fits their marketing strategy. Three ICRA policies have been introduced in this paper as examples. The proposed policies outperform other wireless resource allocation policies in various aspects. The simulation results show that improving resource allocation efficiency significantly increases network revenue and reduces both CDR and CBR. The simulation also shows that service provider satisfaction and customer satisfaction contradict each other. Not surprisingly, increasing the network revenue decreases the customer's average satisfaction level and *vice versa*.

The reservation-less handoff scheme significantly reduces the CDR, but degrades customer QoS marginally. It affects the resources utilization level and slightly reduces the network revenue and the average customer satisfaction levels. The network revenue is affected because the price adjustment process may lower the BMP to prevent the premature termination of an ongoing call.

The two major features of the framework are (1) explicit handling of the trade-off between customer QoS perception and network resource utilization and (2) the flexibility to express numerous different network management policies and objectives. These are key attributes needed in any charging and resource allocation system for next-generation wireless networks.

References

1. Bahl P, Chlamtac I, Farago A. Resource assignment for integrated services in wireless ATM networks. *International Journal of Communication Systems* 1998; **11**: 29–41.
2. Bender P, Black P, Grob M, Padovani R, Sindhushayana N, Viterbi A. CDMA/HDR: A bandwidth efficient high-speed wireless data service for Nomadic users. *IEEE Communications Magazine* 2000; **38**(7): 70–77.
3. Das SK, Jayaram R, Kakani NK, Sen SK. A call admission and control policy for quality-of-service (QoS) provisioning in next generation wireless networks. *Wireless Networks* 2000; **6**(1): 17–30.
4. Dasilva L. Pricing for QoS-enabled networks: a survey. *IEEE Communications Surveys*, Second Quarter 2000; 2–8.
5. El-Kadi M, Olariu S, Abdel-Wahab H. A rate-based borrowing policy for QoS provisioning in multimedia wireless networks. *IEEE Transactions on Parallel and Distributed Systems* 2002; **13**(1): 156–167.
6. Frodigh M, Parkvall S, Roobol C, Johansson P, Larsson P. Future genereation wireless networks. *IEEE Personal Communications Magazine* 2001; **8**(5): 10–17.
7. Fulp EW, Reeves D. The fairness and utility of pricing network resources using competitive markets. *Computer Networks Journal* 2000; **33**(1).
8. Gupta A, Stahl DO, Whinston AB. *Priority Pricing of Integrated Services Networks, Internet Economics*, McKnight, Bailey (eds), MIT Press, 1997; 323–378.
9. IEEE 802.16 broadband wireless access working group. Quality of service classes for BWA. July 22, 1999.
10. ISO/IEC JTC1/SC29/WG11/N2425. MPEG-4 audio verification test results: audio on internet. October 1998.
11. ITU-R BT.500-10. Methodology for the subjective assessment of the quality of television pictures. 2000.
12. Kikuchi H. (M + 1)st-Price Auction. Proceedings of The Fifth International Conference on Financial Cryptography' 01, IFCA, February 2001; pp.291–298.
13. Kim S, Kwon T, Choi Y. Call admission control for prioritized adaptive multimedia services in wireless/mobile networks. *IEEE Vehicular Technology Conference* (VTC'00 spring), Tokyo, May 2000.

Copyright © 2003 John Wiley & Sons, Ltd.

14. Kwon T, Das S, Park I Choi Y. Bandwidth adaptation algorithms with multi-objectives for adaptive multimedia services in wireless/mobile networks. ACM WoWMoM'99, Seattle, August 1999; pp. 51–58.
15. Mackie-Mason JK. A smart market for resource reservation in a multiple QoS network. 25th Annual Telecom Policy Research Conference, Alexandria, VA, September 28, 1997.
16. Malewicz G, Shvartsman AA. An auction-based flexible pricing policy for renegotiated QoS connections and its evaluation. *Proceedings of Seventh International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems* (*MASCOTS' 99*),1999.
17. Oliveira C, Kim JB, Suda T. An adaptive bandwidth reservation policy for high-speed multimedia wireless networks. *IEEE journal on Selected Areas in Communication* 1998; **16**(6): 858–874.
18. Odlyzko A. Paris metro pricing: the minimalist differentiated solution. *Proceedings of the IEEE/IFIP International Workshop on Quality of Service IWQoS'99*, June 1999; 159–161.
19. Sinclair RD. What does the world income distribution look like? The Maxwell School of Syracuse University, Department of Economics, November 2000, version 2.2
20. Singh S. Quality of service guarantees in mobile computing. *Journal of Computer Communications* 1996; **19**(4): 359–371.
21. Songhurst D, Kelly F. Charging policies for multiservice networks. ITC15. Amsterdam, 1997, pp.781–790.

Authors' Biographies

Walid Ibrahim received his B.Sc. degree from Cairo University in 1992 and his Ph.D. from Carleton University in 2002. Since then he has been working as a post doctoral researcher in the Department of Systems and Computer Engineering, Carleton University. His main ares of research include resource allocation and charging in wireless data networks, real time system design and applied optimization techniques.

John W. Chinneck is a professor in the Department of Systems and Computer Engineering at Carleton University. His main research interests are in applied optimization, including the development of computer tools that assist in the formulation and analysis of large and complex models, as well as the development of improved solution algorithms and their applications.

Shalini Periyalwar received the Ph.D. in Electrical Engineering in 1992 from Dalhousie University, Canada. She held an NSERC-Industry funded position as assistant professor in the Department of Electrical Engineering at Dalhousie University from 1991 to 1994, specializing in research in the area of coding and modulation, following which she joined Nortel Networks. She has led teams which have contributed to radio resource management features and system capacity evaluation of Nortel Networks' North American TDMA products, followed by contributions to IS-95 and UMTS access system design and standards. She is currently working on the design of high capacity wireless networks in Nortel Networks' Wireless Technology Labs. She has published several journal and conference papers, and holds three patents. Her research interests include all aspects of system design for cellular and multi-hop networks.

Copyright © 2003 John Wiley & Sons, Ltd.