Fair resource sharing scheduling for cellular data services with QoS provisioning

J.-L.Chen and H.-C.Chao

Abstract: To provide cellular data services with QoS provisions, a shared resource scheme, based on optimisation theory and LaGrange λ-calculus, has been developed. This scheme can generate a fair schedule for a diverse mix of traffic with diverse QoS requirements in a limited radio spectrum. The authors define the acceptance indication (AI) as the QoS measurement of the shared resource scheme. The experimental results show that as much as 2.5% improvement in the mean acceptance rate is obtained relative to other existing schemes.

1 Introduction

The cellular communication architecture with its best effort service model is inadequate for new classes of applications that require QoS assurance [1–3]. IETF defined the implementing concept of differentiated services architecture that offers services of various quality levels to facilitate competitive differentiation in a wire-based Internet [4, 5]. Based on the traffic classification and policing of the service architecture, solutions to providing QoS and accommodation for a large number of mobile users range from providing a substantial radio spectrum increase to using various resource allocation approaches. However, the spectrum resource is very limited, and thus many QoS provisioning mechanisms based on resource allocation schemes were proposed to effectively manage the spectrum and provide service assurances.

The proposed resource allocation schemes may fall into the following groups [6–8]:

Group 1: a maximum number of time slots allocated to a request

Group 2: a minimum number of time slots allocated to a request

Group 3: a maximum number of time slots allocated to a request unless there are not enough available resources

Group 4: resources allocated according to the estimated blocking performance.

The schemes in group 1 will cause a high blocking probability under heavy loads. In this circumstance, few mobile users are simultaneously accommodated. The schemes in group 2 will accomplish a low data rate even if the traffic is light. The above groups are not adequate for real-time applications in cellular networks. The schemes in group 4 can provide good blocking performance. However, as the traffic becomes heavy, the performance of these schemes is similar to that in group 2. The schemes in group 3 are similar to those in group 1, but they provide a flexible technique against high blocking probability. Based on the group 3 conceptual model, we propose a fair shared resource scheme for cellular data services with QoS provisioning.

2 QoS aspects in cellular data services

Mobile communications usually depend on a wireless network that has a cellular architecture. This architecture is a hierarchical structure consisting of a backbone network, mobile switching centres (MSC), base stations (BS) and mobile units, shown in Fig. 1. The backbone network is a wired network connecting the existing wired links to the MSC or MSC to MSC. The MSC connected to the BS is a special switch tailored for mobile applications. A controller embedded in the MSC accomplishes the call control. The BS manages the communication activity of a covered geographic area, called a cell, where the mobile unit stays. A BS is usually in the centre of a cell and neighbouring cells overlap with each other to ensure communications continuity while the mobile users move from one cell to another.

Owing to the advanced WCDMA and WAP applications, mobile multimedia services will increasingly be utilised in our daily life. Table 1 summarises the service classifications with which we are concerned. Basically, three types of service are differentiated according to the traffic characteristics. Type I and III services require bounded delay and guaranteed rate. Type II service, like the conventional data service, does require loss-free transmission, but does not require a guaranteed rate nor bounded delay. Following the service classifications, we tried to determine a QoS indicator that is suitable for cellular data services.

Table 1: Traffic characteristics

Service type	Type I	Type II	Type III
Example	Voice phone	File transfer	Interactive video
Packet delay	bounded	sensitive	bounded
Transmission rate	guaranteed 0–8Kbit/s	non-guaranteed 8–128Kbit/s	guaranteed 0–64Kbit/s
Bit error rate	< 0.001	0	< 0.00001

IEE Proceedings online no. 20010247

DOI: 10.1049/ip-com:20010247

Paper first received 1st June 2000 and in revised form 4th January 2001

J.-L. Chen is with the Department of Computer Science and Information Engineering, National Dong Hwa University, Hualien, Taiwan, Republic of China

H.-C. Chao is with the Department of Electrical Engineering, National Dong Hwa University, Hualien, Taiwan, Republic of China

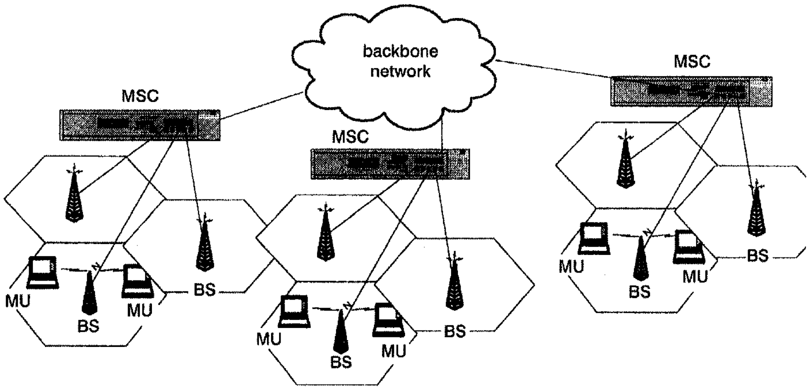


Fig. 1 *Cellular networks*
MU = mobile unit

When a mobile user wants to communicate with another, a channel (several time slots or fundamental code channel) must be requested from the base station in the initial stage, which may succeed or fail. If there are available channels, the mobile user will be assigned a channel for communication. If there are no free channels, the request will be rejected. On receiving a channel, the mobile user starts to communicate with others and may complete his call at the original cell where a new channel was requested or through other cells. If the mobile user completes his call through other cells, more than two different channels must be used during the call duration (to avoid interference). The procedure for requesting a new channel while the mobile user moves across a cell's boundary is generally called a handoff. If the handoff succeeds, the mobile user continues his communication without interruption. If the handoff fails, the mobile user's call is force-terminated [9, 10].

Based on the call life flow, a request and satisfactory service for a mobile user is accomplished with:

(*a*) the request not blocked in the initial stage

(*b*) service not force-terminated in the handoff stage

(*c*) a high quality of service during the service hold time.

Committing to mobile users' expectations, we defined an 'acceptance indication' (AI) as the QoS measurement in cellular data services. Therefore,

$$AI = K_1(1 - P_b) + K_2(1 - P_f) + K_3P_s \quad (1)$$

$$P_b = \sum_{B_0} p(i, j, k) \quad (2)$$

$$P_f = p(C_h, j, k)|_{j=C_f - C_h, k=C_d, C=C_f+C_d} \quad (3)$$

$$P_s = \mu[P_cT_c](C_{i,Allocate}/C_{i,Max}) \quad (4)$$

$$K_1 + K_2 + K_3 = 1 \quad (5)$$

where:

P_b = estimated probability that a request will be blocked

P_f = estimated probability that service will be force-terminated

P_s = percentage that indicates user satisfaction for the call connection

P_c = estimated probability that a call will be completed

C_h = number of reserved channels (time slots) for handoff calls in a cell

C_f = fixed channels (time slots) in a cell

C_d = dynamic channels (time slots) in a cell

$C_{i,Allocate}$ = expected number of time slots allocated to a request for the *i*th service

$C_{i,Max}$ = maximum requirement capacity at a request for the *i*th service

C = total available resources at the scheduling time

B_0 = $\{(i, j, k)|j = C_f - C_h,\ k = C_d,\ 0 \le i \le C_h,\ C = C_f + C_d\}$

T_c = expected effective service time for a completed call

μ = mean service rate

K_1, K_2, K_3 = weighting values that concern mobile users.

In the above parameters, P_s indicates the mobile user's satisfaction with the call connection, and $1/\mu$ is the mean service time for a completed call. To increase flexibility in a service network, a hybrid channel allocation scheme is adopted in our research. In this scheme, the channels are divided into fixed and dynamic sets. The fixed channels, C_f, are assigned to different cells and all of the users share the dynamic channel, C_d.

Eqn. 4 is derived from [6], and these specified parameters were estimated using an OpNet simulator. Eqn. 5 gives the weighting values for different service classes corresponding to the performance indices. For example, service assurance is more important than a service request in a voice phone service, and therefore K_1 is less than K_2 and K_3. Mobile users may negotiate with service providers about the three weighting values.

3 LaGrange λ-calculus for QoS provisioning

In a cellular data network, there are several applications that may be simultaneously requested. In this circumstance, the traffic characteristics may be different for most mobile users and thus the required QoS assurances are different. Meeting the differentiated QoS is difficult, especially for limited resources. From the service providers' point of view, the highest acceptance by all mobile users is expected. Mathematically speaking, the problem may be stated very concisely. That is, an objective function, AI_T, is equal to the total acceptance indication for serving the indicated requests. The issue is to maximise AI_T subject to the constraint that the sum of the requested resources must be less than or equal to the available resources. That is:

objective function

$$\max(AI_T), \quad AI_T = AI_1 + AI_2 + AI_3 + \ldots + AI_n \quad \text{for } n \text{ service request} \quad (6)$$

subject to

$$1.\ \sum_{j=1}^{n} C_{j,Allocate} \le C \Rightarrow \phi \simeq 0 \le C - \sum_{j=1}^{n} C_{j,Allocate} \qquad (7)$$

$$2.\ C_{i,Min} \le C_{i,Allocate} \le C_{i,Max}, \quad i = 1, 2, \ldots, n \qquad (8)$$

where:

AI_i = acceptance indication for ith service

$C_{i,Min}$ = minimum requirement capacity at a request for ith service.

This is a constrained optimisation problem that may be attacked formally using advanced calculus methods that involve the LaGrange λ-calculus [10]. To establish the necessary conditions for an extreme AI_T value, add the constraint function to the AI_T after $\varnothing$ has been multiplied by a multiplier, λ. This is known as LaGrange λ-calculus and is shown in eqn. 9,

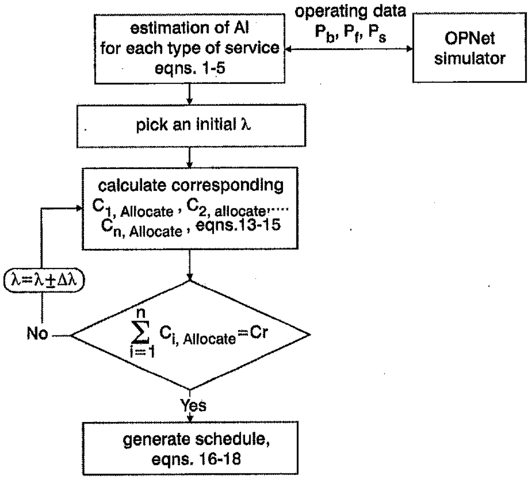
$$\zeta = AI_T + \lambda\phi \qquad (9)$$

The necessary conditions for an extreme AI_T value results when we take the first derivative of the λ-calculus with respect to each of the independent values and set the derivatives equal to zero. Based on the application domain, we only take the derivative of the λ-calculus with respect to the $C_{i,Allocate}$ values at a scheduling time give the set of equations shown as eqns. 10 and 11:

$$\frac{\partial \zeta}{\partial C_{i,Allocate}} = \frac{\partial AI_i}{\partial C_{i,Allocate}} - \lambda = 0, \quad i = 1, 2, \ldots, n \qquad (10)$$

$$\lambda = \frac{\partial AI_1}{\partial C_{1,Allocate}} = \frac{\partial AI_2}{\partial C_{2,Allocate}} = \ldots \qquad (11)$$

The necessary condition for the existence of a maximum AI_T for cellular data services is that the incremental AI of all of the mobile users is equal to λ. Of course, to produce this necessary condition we must add the constraint equation that the sum of the $C_{i,Allocate}$ values must be less than or equal to C. In addition, there are two inequalities that must be satisfied for each request. That is, the $C_{i,Allocate} \ge C_{i,Min}$ and $C_{i,Allocate} \le C_{i,Max}$. The operating flow chart is shown in Fig. 2. The operating scenarios are described as follows.



estimation of AI for each type of service eqns. 1-5
operating data
P_b, P_f, P_s
OPNet simulator
pick an initial λ
calculate corresponding $C_{1,Allocate}$, $C_{2,allocate}$,..., $C_{n,Allocate}$, eqns.13-15
λ=λ±Δλ
No
$\sum_{i=1}^{n} C_{i,Allocate}=Cr$
Yes
generate schedule, eqns. 16-18

Fig.2 *Flow chart of λ-calculus for QoS provisioning*

Step 1: Estimation of AI for each type of service. Many treatments were performed using an OpNet simulator according to the $C_{i,Allocate}$ variation, and thus a set of AI values for the specified service class was estimated. Through the curve fitting technique for operating data filtering, the AI function for each service class may be treated for the linear or quadratic rate case. For example,

$$AI_i = a_i + b_i C_{i,Allocate} + c_i C^2_{i,Allocate}, \quad a, b, c = \text{constant} \qquad (12)$$

Step 2: λ-calculus. Based on Step I, we assume two types of services whose incremental AI values are represented by the following equations:

$$\frac{dAI_1}{dC_{1,Allocate}} = 2c_1 C_{1,Allocate} + b_1 = \lambda \qquad (13)$$

$$\frac{dAI_2}{dC_{2,Allocate}} = 2c_2 C_{2,Allocate} + b_2 = \lambda \qquad (14)$$

$$C_r = \min\{C, C_{1,Max} + C_{2,Max}\} \qquad (15)$$

Therefore, $C_{1,Allocate}$ and $C_{2,Allocate}$ may be estimated while a λ-value is selected.

Step 3: Schedule generation. The iterative process of finding the λ-value stops when $\Sigma_{i=1}^{n} C_{i,Allocate} = C_r$. Two cases were discussed in the schedule generation:

Case 1: $C_r = (C_{1,Max} + C_{2,Max})$, thus:

$$C_{1,Allocate} = C_{1,Max}$$
$$C_{2,Allocate} = C_{2,Max}$$

Case 2: $C_r = C$; we recognise the inequality constraints, then the necessary conditions may be expressed as shown in the set of equations making up eqns. 16–18. The incremental rate relation with constraints is illustrated in Fig. 3:

$$\frac{dAI_i}{dC_{i,Allocate}} = \lambda \text{ for } C_{i,Min} < C_{i,Allocate} < C_{i,Max} \qquad (16)$$

$$\frac{dAI_i}{dC_{i,Allocate}} \le \lambda \text{ for } C_{i,Allocate} = C_{i,Max} \qquad (17)$$

$$\frac{dAI_i}{dC_{i,Allocate}} \ge \lambda \text{ for } C_{i,Allocate} = C_{i,Min} \text{ or } 0 \text{ (blocked)} \qquad (18)$$

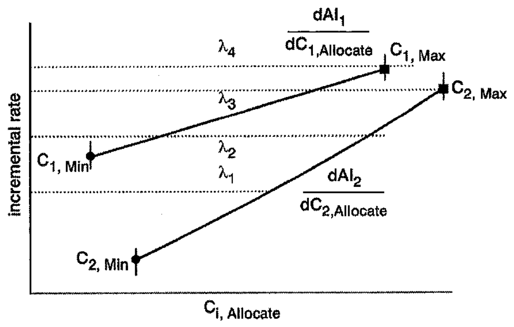


Fig.3 *Incremental rates with constraints*

From the observations in Fig. 3, three situations are identified:

Situation I: $\lambda = \lambda_2$ or λ_3; both the $C_{1,Allocate}$ and $C_{2,Allocate}$ values will fall into the range $\{C_{i,Min}, C_{i,Max}\}$.