Trade-offs Between Tariffs and QoS in Mobile Telephony Networks: 
an Integrated Design Approach

M.Ajmone Marsan, A.Bianco, M.Calderini, C.Cambini, M.Meo
Politecnico di Torino,
Corso Duca degli Abruzzi 24, 10129 Torino, Italy
{ajmone,bianco,michela}@polito.it
calderini@athena.polito.it, cambini@lep.polito.it

Abstract

In this paper we present an analytical approach for the joint optimization of quality of service (in terms of call blocking probability) and tariffs for a mobile telephony network. The analytical approach is based on the combination of traditional telecommunication system design techniques and econometric approaches for profit maximization.

1 Introduction

The traditional approach to design and capacity planning of a telecommunication system aims at the identification of the minimum set of resources that allow the system to meet a specified group of constraints on the quality of service (QoS) offered to end users, under given system operating rules, concerning both the internal system elements, and the end users behaviors (the internal system operating rules correspond to the system management algorithms; the end users behavior is described through a workload – or traffic – model, which is generally expressed in mathematical form).

The minimization adopts as metrics the costs of the resources that are necessary for the system implementation.

The result of the traditional approach to design and capacity planning of a telecommunication system is thus a minimum cost design for the specified QoS. In a monopolistic market, the selection of the QoS level to be used in the design process is arbitrary, and the tariffs for end users are determined as a result of the system and operations cost.

In a competitive market environment, where a number of telecommunication system operators offer services with different QoS levels and tariffs to end users, the selection of the QoS constraints to be used in the system design process is itself an output of the optimization process, because different QoS levels translate into different system costs, which in turn translate into different tariffs; both tariffs and QoS are key elements for the acquisition of market shares.

This means that, in a competitive telecommunications market, the system design process comprises a number of steps:

1. identification of the system management algorithms and of the relevant QoS parameters
2. definition of a parametric traffic model (parameters allow the scaling of traffic)
3. definition of a parametric system model (parameters are the numbers of resources)
4. selection of a range of different target QoS levels
5. computation of the functions that specify the amount of resources that allow each one of the selected target QoS levels to be reached, versus the workload parameters
6. association of cost functions to resources
7. computation of the functions that specify the system cost that allows each one of the target QoS levels to be reached, versus the traffic parameters
8. parametrization of the cost function: estimation of the demand and revenue functions in the QoS and tariffs variables
9. identification of the optimal strategy for the telecommunications system operator, i.e. profit maximization in QoS level and tariffs.

In this paper we follow these steps, focusing, as an example, on a simple element of a telecommunication system,
i.e., one cell of a telecommunication network providing mobile telephony services according to the GSM standard.

The approach that we describe in this paper for telecommunication system design and planning differs from previous attempts to a joint optimization of the traffic and economic aspects of a telecommunications system: the relationship between QoS and traffic on one side, and system cost on the other, is made explicit through a process that directly involves the engineering design of the network, rather than the usual generic assumption about the technical characteristics of the network.

2 Step 1 – System Management Algorithms and QoS Parameters

The selection of the specific telecommunication system to be considered allows the identification of the system management algorithms (step 1 of the telecommunication system design process).

Indeed, the GSM cell that we consider in this paper offers basic telephony services by allocating resources to each conversation (call). The resources necessary to set up a call in a cell amount to a free time slot on one of the frequency pairs available in the cell. The time/frequency resources necessary to serve a call request form a bidirectional channel. While the number of slots per frequency is a constant (8, out of which normally 1 is devoted to signalling and 7 can be allocated to calls in progress1), the number of frequencies in a cell (and thus the number of channels) is a design parameter, which determines the system cost and QoS.

The system supports users' mobility by means of handover procedures that allow the network to follow the user movements during a call.

A cell receives requests of new call setup from users which reside in the area covered by the cell, and receives handover requests from users which are moving from neighboring cells into the considered one. In both cases the request fails when the cell cannot allocate a channel for the service; i.e., when all channels are busy serving other calls.

Assuming that once a channel is allocated, the service received by the user is of good quality (this is normally the case in circuit-switched telecommunication systems, like GSM), the QoS perceived by the user can be measured in terms of the probability that a user request fails. Denote this probability with Q. Q is the QoS parameter which will be used throughout the paper for the description of the system design approach.

3 Step 2 – Workload Model

For the implementation of steps 2 (definition of a workload model) and 3 (definition of a system model) we develop Markovian models.

Markovian models have been traditionally used for the design and planning of mobile cellular telephony networks, considering one cell at a time (see for example [1, 2, 3, 4, 5]); this means that the analysis focuses on the description of the behavior of a single cell, and approximates the description of the interaction between adjacent cells.

The dynamics that govern the cell workload are the arrival processes of call requests (either new calls or incoming handovers) and of channel releases (either because of a call termination or because of an outgoing handover).

In the development of the workload model, the following assumptions are introduced:

- the aggregate process of new call requests within the cell is Poisson with parameter λ; λ is estimated from the user population and the system geometry

- the flow of incoming handover requests from other cells is Poisson with rate λh; λh is derived by balancing the incoming and outgoing handover flows, as explained below

- the call duration is an exponentially distributed random variable with parameter μ; μ is obtained from the observation of the user behavior

- the time between two successive handover requests of a call (the call dwell time) is an exponentially distributed random variable with parameter μh; μh is obtained from the observation of the user mobility.

Cells are studied in isolation from the rest of the system. The interaction with adjacent cells is taken into account by means of the incoming handover flow, which is assumed to be equal to the handover flow out of the considered cell. This approach is justified by the results presented in [6], where it was shown that more complex (multi-cell) models do not lead to significant improvements in the accuracy of performance predictions.

The incoming handover rate for the considered cell must be evaluated numerically, since it cannot be a-priori derived from the model parameters. An iterative procedure is used to balance the incoming and outgoing handover rates, assuming that the incoming handover rate at step j is equal to the outgoing handover rate computed at step j - 1. The iterative procedure is stopped when equilibrium is reached.

Two different workload scenarios are considered, where users have different mobility behaviors. In the first scenario one handover per call is required on average: μh = μ; in the second one, 10 handovers per call are required on average:
\( \mu_h = 10 \mu \). The user mobility is determined by the end user behavior in the environment in which the cell is settled. For example, the choice of moderate mobility may be suitable for an urban area, while high mobility may be apt for modeling a freeway. In both cases the average call duration is \( 1/\mu = 180 \) s.

### 4 Step 3 – System Model

Each cell is characterized by the maximum number of calls that can be simultaneously activated in it; we shall denote this parameter by \( N \) for the cell under consideration.

Due to the flow balance assumption described above, a cell can be modeled as a \( M/M/K/0 \) queue, where:

- \( K = N \) servers represent the channels which are available in the cell;
- the arrival process is the superposition of the arrival processes of new call requests and of incoming handovers; the resulting total arrival rate is \( \lambda + \lambda_h \);
- the service time represents the channel holding time, that is the interval between the instant in which a user starts using a channel and the instant in which the channel is released either because of the call termination or because the user is moving out of the cell; it is an exponentially distributed random variable with parameter \( \mu + \mu_h \).

The state \( s \) of the cell is the number of busy channels, which coincides with the number of active connections. \( s \) can assume all integer values between 0 and \( N \). The steady-state probabilities \( \pi(s) \) that \( s \) connections are active can be computed from the formula:

\[
\pi(s) = \frac{\pi(0) \rho^s}{s!}
\]

with

\[
\pi(0) = \left( \sum_{s=0}^{N} \frac{\rho^s}{s!} \right)^{-1}
\]

where \( \rho \) is the traffic intensity, \( \rho = \frac{\lambda + \lambda_h}{\mu + \mu_h} \).

The blocking probability \( Q \) that a call cannot be accepted in the system due to lack of available channels is equal to the probability \( \pi(N) \):

\[
Q = \pi(N) = \frac{\pi(0) \rho^N}{N!}.
\]

This is the well-known Erlang B formula.

The blocking probability \( Q \) is the QoS parameter used in the joint optimization of QoS and tariffs.

Given a value of \( Q \) and a specification of the user behavior, in terms of call request rate, call duration and mobility (i.e. in terms of \( \lambda, \mu \) and \( \mu_h \)), the number of channels necessary to guarantee \( Q \) can be derived by inverting equation (3).

### 5 Steps 4 and 5 – QoS Levels and Curves

We now describe the results that can be obtained for step 5 of the design procedure (computation of the functions that specify the amount of resources that allow each one of the selected target QoS levels to be reached, versus the workload parameters), after a set of target QoS levels has been selected (step 4), in the case of the telecommunication system under consideration.

Consider the first workload scenario of low user mobility. Left plots in Fig. 1 show the number of channels that are required to guarantee the target QoS level \( Q \) equal to \( 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4} \). Of course, as the target \( Q \) decreases, the number of channels increases, almost doubling when tightening the QoS requirement from \( Q = 10^{-1} \) to \( Q = 10^{-4} \). The flat part of the curve for low values of the input traffic is due to the introduction of a lower bound for the number of channels per cell. This bound is set to 7, which corresponds to the number of traffic channels provided by one radio frequency. When the GSM cell is equipped with the number of channels shown in left plots of Fig. 1, the actual \( Q \) perceived by end users is as reported in right plots of the same figure. The channel utilization in terms of the average number of busy channels is plotted in Fig. 2.

Considering the high user mobility scenario (results are not shown for the sake of brevity) we can observe that the increase in the end user mobility makes the QoS dimensioning slightly less demanding. This is due to the decrease in the channel holding time which makes the traffic intensity decrease; this effect is particularly evident for low values of target QoS.

### 6 Steps 6 and 7 – Cost Functions and Curves

Steps 6 and 7 of the telecommunication system design process require the association of cost functions to the different resources necessary for the system implementation, and the derivation of the curves that specify the system cost that allows each of the target QoS levels to be reached, versus the workload parameters.

The equipment necessary to activate a GSM cell is termed “base station”. Roughly speaking, a base station is composed of a set of transceivers, each one of which controls one frequency pair, and an intelligent controller, that handles part of the signalling functions, collecting service requests, and establishing connections to the closest switching center of the mobile telephony network. Transceivers
reside in BTS’s (Base Transceiver Stations), which are placed over the GSM network service area to provide an effective coverage. The intelligent controller is called BSC (Base Station Controller); each BSC is connected to a number of BTS’s by means of (normally wired) 2 Mb/s links. BSCs are then connected to MSCs (Mobile Switching Center), which collect and switch voice traffic. MSCs are the digital exchanges of the PSTN.

Let us first focus on installation costs. Cost parameters derive from a rough cut estimation. To establish a GSM network, several MSCs and BSCs have to be installed. The installation cost of an MSC can be estimated about 1,500 kUSD dollars (in the following we will use USD for dollars and kUSD for thousands of dollars); the installation of a BSC is about 300 kUSD.

Since BSCs and MSCs serve a large number of cells, we partition their installation costs between the served cells, thus we can define the cost of a cell taking into account the cost of all the elements which the cell needs to be operated. Referring to our region, two MSCs serve the traffic of North West Italy, with about 1,000 cells; in the same region, 3 BSCs are employed. Therefore, the cell share of the installation cost of the MSCs is 2 · 1,500/1,000 = 3.0 kUSD; similarly, for the BSCs cost share per cell 3 · 300/1,000 = 0.9 kUSD.

The installation cost of a base station depends mainly on the geography of the area which the cell must cover. In a urban area, for example, preexisting buildings and structures are usually exploited as sites for the installation. In the countryside, instead, ad-hoc structures are to be erected, causing the installation cost to increase. In other, even more inaccessible environments, such as mountains and narrow valleys, the cost of the base station installation can grow even more. Focusing on base stations in urban area, a reasonable estimate of the installation cost of a BTS is 50 kUSD dollars.

Installation costs used in the paper are listed in Table 1.

<table>
<thead>
<tr>
<th>Component</th>
<th>Installation cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSC</td>
<td>3,000</td>
</tr>
<tr>
<td>BSC</td>
<td>900</td>
</tr>
<tr>
<td>BTS</td>
<td>50,000</td>
</tr>
</tbody>
</table>

Table 1. Installation costs in dollars

Once an area has been covered with a BTS site, service
providers have to decide upon the number of frequencies (hence channels) to be activated in the cell. A channel is specified by a time slot and a frequency; each frequency carries eight time slots, one devoted to signalling, seven to user traffic. The cost of a frequency is thus shared by seven user traffic channels.

The cost of activating a new frequency at an already installed base station is given by the cost of mounting a transceiver on the BTS and possibly the cost of adding a link between the BTS and the BSC, if the existing one is not able to sustain the traffic increment from the BTS to the BSC. Links between the BTS and the BSC have a capacity 2 Mb/s and can carry up to 21 traffic channels, given that a 2 Mb/s channel is shared by three BTSs. Therefore, adding a new frequency not necessarily leads to the need for a new 2 Mb/s link from BTS to BSC.

The fixed initial cost of mounting a new transceiver is estimated to be about 3 KUSD; we call this a “setup” cost. The cost of a new link between BTS and BSC is instead given by a fixed initial setup cost, estimated to be 250 USD, and a monthly cost of renting the link, equal to 350 USD per month. These data are summarized in Table 2.

<table>
<thead>
<tr>
<th>Component</th>
<th>Setup</th>
<th>Monthly</th>
<th>Carried traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transceiver</td>
<td>3,000</td>
<td>none</td>
<td>7 channels</td>
</tr>
<tr>
<td>Link BTS–BSC</td>
<td>250</td>
<td>350/month</td>
<td>21 channels</td>
</tr>
</tbody>
</table>

Table 2. Costs in dollars for increasing traffic channels in a cell

The above costs are instrumental in computing the curves that specify the system cost that allows each of the target QoS levels to be reached, starting from the curves determined in the previous section, that provide the number of resources that allows each of the target QoS levels to be reached.

We focus on the case of low mobility (one handover per call on average) already considered in the previous section, and set a target QoS level ranging from $10^{-1}$ to $10^{-4}$. The number of channels necessary to provide the desired QoS is shown in the left plots of Fig. 1. With the costs reported in Table 2, it is possible to derive the setup and monthly costs of equipping the cell with a given number of channels; results are shown in Fig. 3. Installation costs reported in Table 1 are not considered in the plots since they do not depend on the number of channels in the cell.

2Actually, each 2 Mb/s channel can carry the data of 10 frequencies; in order to simplify matters, we assume for simplicity that each cell of a 3-cell clover (the most common cell configuration) can load on the 2 Mb/s channel the data of 3 frequencies, and that each frequency carries 7 data channels and 1 signalling channel.

7 Step 8 – The economic model

The purpose of this section is to set up a model which enables us to determine the optimal QoS and tariffs. The goal of a telecommunication network operator is to maximize its profit function which is the difference of two components: the revenue function and the cost function. The former can be written in the generic form: $R(p, Q)$, where $p$ is the price (or tariff) per second, and $Q$ is the QoS offered by the system. The latter can be written as: $C(Q, \rho)$, where $\rho$ is the traffic intensity in the cell. Accordingly, the profit function will be defined as: $\Pi(p, Q, \rho) = R(p, Q) - C(Q, \rho)$.

Such functions must be made explicit. First, the estimation of the cost function is performed, second, some assumptions are made on revenues in the profit function.

7.1 The cost function

The generic cost function $C$ of a mobile telephony operator can be written in the form $C = C(Q, \rho)$, where $C$ represents long-run total costs, $Q$ the system QoS, and $\rho$ the traffic intensity in a single cell. In our model, we define long run total costs as the investment cost needed to install the communication equipment. We largely simplify the matter by neglecting operating costs which represent a small fraction of total costs. This assumption is fairly realistic in the telecommunications industry. Investment costs are redeemed over the life time of the network which is estimated to be equal to five years. Such a short time lapse is due to the technological innovation which pushes operators to change quite often the systems they use.

In the representation of the cost function, we adopt the so-called translog flexible functional form, which has been extensively used in many telecommunications studies [10, 11, 13, 14, 15, 16]. The translog function can be interpreted as a second order Taylor’s approximation of an arbitrary cost function [8, 9, 12]. In particular, the most popular translog cost function, which we use in the following sections, approximates the logarithm of $C(Q, \rho)$ as a second order (quadratic) approximation in the logarithms of $(Q, \rho)$, that is:

$$
\ln C = a_0 + a_1 \ln(p) + a_2 \ln(Q) + a_3 (\ln(p))^2 + a_4 [\ln(Q)]^2 + a_5 \ln(p) \ln(Q)
$$

(4)

where $a_i$ ($i = 1, 2, \cdots, 5$) are the coefficients to be estimated. Since total costs are expressed in logarithms, the first order coefficients are to be interpreted as cost elasticities evaluated at the sample median. In particular:

- $a_0$ is a measure of the level of total fixed cost
- $a_4$ is the cost elasticity with respect to the traffic variable, where the cost elasticity is given by: $\frac{\partial C}{\partial p} \cdot \frac{p}{C}$

550
Figure 3. Setup cost (left plots) and monthly cost (right plots) necessary to guarantee QoS equal to $10^{-1}$ and $10^{-4}$ for different values of input traffic; $1/\mu = 180$ s, $\mu_h = \mu$

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_0$</td>
<td>2.0474</td>
<td>0.0270</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>-0.2072</td>
<td>0.0087</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>-0.0551</td>
<td>0.0063</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>0.0666</td>
<td>0.0018</td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td>-0.0038</td>
<td>0.0004</td>
</tr>
<tr>
<td>$\alpha_5$</td>
<td>-0.0224</td>
<td>0.0009</td>
</tr>
</tbody>
</table>

Table 3. Cost function coefficient estimates

- $\alpha_3$ is the cost elasticity with respect to the QoS $Q$ of the mobile service, given by $\frac{\partial C}{\partial Q}$
- $\alpha_1$ measures the change in the cost elasticity of the traffic due to a change in the traffic level
- $\alpha_4$ measures the change in the cost elasticity of the QoS due to a change in the QoS level
- $\alpha_5$ measures the change in the cost elasticity of the QoS/traffic due to a change in the traffic/QoS (cross effect).

A simple least square linear regression on the available data observations allows us to estimate the coefficients and their associated standard errors (see Table 3). Note that we verified that for every parameter estimate the confidence interval with confidence level 0.9 does not include 0; this confirms that the chosen translog flexible functional form (4) is apt to describe the cost function of our system.

On the basis of the estimated parameters, we can draw the following conclusions.

1. The cost elasticity with respect to the traffic level is negative in the term $\alpha_1$ and positive in $\alpha_3$. This is due to a border effect of the least square linear regression. For realistic values of traffic $\rho$ the term with coefficient $\alpha_3$ in (4) dominates; this implies that an increase in the number of calls will increase total costs.

2. The cost elasticity with respect to the QoS level is negative; this implies that a decrease of $Q$ will increase total costs. Thus, an increase in quality requires an increase in total costs.

7.1.1 The demand and revenue functions

The revenue function will be derived through standard assumptions about the demand function. This latter is defined as: $D(p, Q) = \exp(-bp + a(1 - Q))$, where $p$ is the price per time unit and $Q$ is the QoS. The demand function is expressed in terms of seconds spent on the phone by consumers. This form is compatible with standard economic assumptions about demand functions. Demanded quantity is decreasing in price and in $Q$. Therefore, the revenue function can be defined as: $R(p, Q) = pD(p, Q) = p\exp(-bp + a(1 - Q))$.

As already stated, we miss the data that would allow us to estimate the parameters of the function: $a, b$. This is not a difficult exercise, but requires knowing how responsive is the total time spent by users on mobile calls to shifts in price ($b$) and in $Q$ ($a$). This could be obtained through a standard econometric model based on historical data of mobile telephony operators. Given the unavailability of empirical data, we will simulate our model choosing arbitrary (but realistic) values for $a$ and $b$, leaving aside the estimation issue.


Figure 4. Profits as a function of Price and Quality for a=1, b=1


Figure 5. Profits as a function of Price and Quality for a=2, b=0.5

8 Step 9 - Profit maximization

Substituting the expressions of the revenue and cost functions in the profit function we derive:

\[ \Pi(p, Q, \rho) = R(p, Q) - C(Q, \rho) = p \exp(-bp + a(1 - Q)) - C(Q, \rho) = p \exp(-bp + a(1 - Q)) - \exp(2.0474 - 0.2072 \ln \rho) \exp(-0.0551 \ln(Q) + 0.0666[\ln \rho]^2) - 0.0224 \ln \rho \ln(Q)) \]

The profit function is expressed in monetary units (e.g. in Dollars). This allows us to simply define our maximization problem in the form of:

\[ \max_{P, Q} \Pi \quad \text{s.t.} \quad \Delta t \exp(-bp + a(1 - Q)) = \Delta t \rho (1 - Q) \]

The constraint implies that demand in time span \( \Delta t \) is equal to the actual traffic that, in the same time span is generated in the cell, i.e. the volume of traffic (\( \rho \)) weighted by the call success probability \( 1 - Q \). The solution of this problem cannot be obtained in closed form, due to computational constraints. However, we can simulate the equation in order to have a rough cut approximation of the optimal strategy for the operator in terms of tariffs and QoS. As stated above, due to lack of empirical data on the demand side, we have to make an assumption about the values of the parameters \( a \) and \( b \) in the demand function. For the sake of exposition, we simulate the profit function for the following pairs: \( (a = 1, b = 1) \) and \( (a = 2, b = 0.5) \). These pairs have been chosen within a range which is normally considered relevant in empirical economic literature, identifying normal levels of elasticity for fixed telephony [17]. In absence of accurate estimations for mobile communications, we have considered slightly greater values as reasonable, given consumer’s economic characteristics. The numerical results obtained with the two pairs of values are reported in Figs. 4 and 5, respectively.

The plots of the profit function show that, for different values of the parameters \( a \) and \( b \), both tariffs and QoS maximize the operator’s profit at different levels. The managerial relevance of this result is that, when \( a \) and \( b \) are known through econometric demand studies, it would be possible to determine the optimal combination of price and quality to be supplied.

Parametrically, when the price coefficient \( (b) \) decreases, the optimal price increases, whilst for greater values of \( b \) the optimal price decreases. In fact, the interpretation of coefficient \( b \) can be given as the sensitivity of the demanded quantity to price. As far as QoS is concerned, it has to be remarked that profit is sensitive to very small adjustments in QoS level. For increasing sensitivity to QoS \( (a) \), the optimal quality level increases \( (Q \) decreases).

It has to be noted that, whilst for the price coefficient \( (b) \) we have a fairly good parametrical estimation, the demand response to small variations in the QoS level is largely unexplored. Therefore, a consistent implementation of the model should require a sound econometric study of the demand as a function of QoS.
9 Conclusions and further research

We developed a quantitative approach for the integration of economic models and traditional telecommunications network design approaches. The rationale beyond the proposed approach is the consideration that standard economic models fall short of their goals when trying to provide a sound support to the operator's strategic decisions in terms of tariffs and QoS. This is due, in our view, to the fact that many economic models fail to capture in full the technological and engineering dimension that is at the basis of any economic decision. Our attempt tries to fill this gap, building a cost function that explicitly takes into account the technological constraints involved in the design of a mobile telephony network.

This paper should be taken as a first attempt to integrate two different approaches to efficiently design a mobile telephony network. We have tried to provide an example of what a serious integration effort could consist of. We are confident that the integration of the two approaches could yield a remarkable added value to the definition of a useful strategic model.

Acknowledgments

The authors would like to thank Raffaele Menolascino of Motorola Technology Center (Italy) for suggestions on the cost function and for providing data and Tommaso Valletti of Politecnico di Torino for useful discussions.

References


