Ann Oper Res (2012) 199:179–191
DOI 10.1007/s10479-011-0993-1

# Optimal number of bypasses: minimizing cost of calls to wireless phones under Calling Party Pays

**Eduardo González · Leonardo D. Epstein · Verónica Godoy**

Published online: 4 October 2011


**Abstract** In telecommunications, Calling Party Pays is a billing formula that prescribes that the person who makes the call pays its full cost. Under CPP land-line to wireless phone calls have a high cost for many organizations. They can reduce this cost at the expense of installing wireless bypasses to replace land-line to wireless traffic with wireless-to-wireless traffic, when the latter is cheaper than the former. Thus, for a given time-horizon, the cost of the project is a trade-off between traffic to-wireless and the number of bypasses. We present a method to determine the number of bypasses that minimizes the expected cost of the project. This method takes into account hourly varying traffic intensity. Our method takes advantage of parallels with inventory models for rental items. Examples illustrate the economic value of our approach.

**Keywords** Communications · Telephony · Wireless bypasses · Calling Party Pays · Rental situations · Inventory models

## 1 Introduction

In telecommunications, Calling Party Pays (CPP) is a billing formula that prescribes that the person who makes the call pays its full cost. Under CPP some corporations incur high expenditures associated with their land-line to mobile telephone calls. Comparing fees, a time unit of land-line to mobile connection is typically more expensive than a time unit of land-line to land-line connection. Hence, a corporation may seek to reduce these expenditures with the use of wireless bypasses. A wireless bypass (WBP) is a device that enables direct

E. González (✉) · L.D. Epstein · V. Godoy
Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez, Diagonal Las Torres 2640 Peñalolén, Santiago, Chile
e-mail: eduardo.gonzalez@uai.cl

L.D. Epstein
e-mail: leonardo.epstein@uai.cl

V. Godoy
e-mail: vgodoy@uai.cl

connection of a corporation's internal telephony system to any commercial cellular network, via an already existing Private Automatic Branch Exchange (PABX) system. Thus, a WBP reduces the cost of land-line to mobile calls by replacing land-line to wireless traffic with cheaper wireless to wireless traffic.

The initial outlay to purchase WBPs to save later on calls to mobile phones leads to consider the problem of determining the number of WPBs that maximizes the value of the savings of the project for a given time horizon. Our review of the literature suggests that no formal approaches are available to solve this problem, hence the importance of the methods this article proposes.

The system we model consists of $n$ WBPs and $m$ overflow channels. An overflow channel (OFC) is a land-line assigned to route calls to mobile phones when no WBPs are available. Note that a WBP in use may become available while OFCs are in use. This article concerns the situation when $m$ land-lines are in place and one wishes to determine $n$ to maximize the expected present value of the savings.

In the baseline situation there are $m$ OFCs and $n = 0$ WBPs. The baseline expected cost $C_{\text{Base}}$ is the expected present value of calls that go through these OFCs. This baseline expected cost does not depend on the number of WBPs. In this sense it is constant. With a system with $n > 0$ WBPs, the expected cost $C(n)$ is the expected present value of calls that go either through OFCs or through WBPs. The expected savings is the difference $S(n) = C_{\text{Base}} - C(n)$. Hence, maximizing expected savings is equivalent to minimizing the expected cost.

The situation we model connects closely with inventory models for rental situations: a caller borrows a line from an inventory to place a call, then she returns the line to the inventory when she completes the call. This returned line becomes available for future callers. The cost associated with the use of a line is proportional to the length of the call. Hence, the situation we model has a special feature: the cost of rental is proportional to the length of time between the borrowing and the returning times. Tainiter (1964) seems to have pioneered models for rental situations with a single type of items. Later, Jain (1966) proposed a model for rental situations where if one's inventory is exhausted, then one can obtain items from a third party. Finally, Whisler (1967) made closely related contributions. These three papers provide basic models that use a general distribution for the inter-demand times and exponential distribution for the length of rentals.

The organization of the rest of this article is as follows. Section 2, develops a two-state transition model. Section 3 develops the expected cost as a function of $n$ and $m$. This development incorporates historical hourly calling rate data into the expected cost. Section 3 includes a description of a simple strategy to obtain an initial value for the optimal value of $n$. This initial value is the optimal value of a single state version of the model. Section 4 presents an application closely related to the problem that motivated this work. Finally, Sect. 5 discusses a number of interpretations of the models and features of cost functions. It also describes lines of future research.

## 2 A transition model with two state-variables

We model the system of calls to wireless phones with two state-variables: the number of WBPs in use at time $t$, $X_B(t)$, and the number of OFCs at time $t$, $X_O(t)$. We assume that the time between calls is exponentially distributed and that the duration of the calls is also exponentially distributed. Our approach accommodates time-varying traffic patterns with the use of approximating step-functions, that is, functions that are constant over a finite number

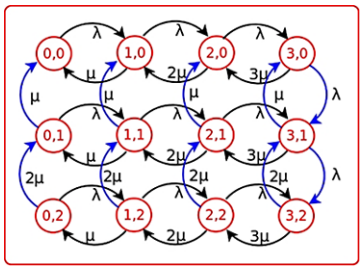


**Fig. 1** State transition diagram for 3 WBPs and 2 OFCs

of time intervals. These step-functions allow one to use a model with steady-state transition probabilities within each interval. We let $h$, $h = 1, \ldots, H$, index these time intervals. For concreteness, Sect. 3 uses the time intervals $]h-1, h]$, $h = 1, \ldots, H = 24$, but this choice may be modified to suit specific situations.

For each time interval, we assume the time between calls exponentially distributed with mean $1/\lambda^{(h)}$, that is, $\lambda^{(h)}$ is the hourly calling rate. We further assume that the duration of the calls is exponentially distributed with mean $1/\mu^{(h)}$, that is, $\mu^{(h)}$ is the service rate. Then the expected number of busy channels per hour is $\lambda^{(h)}/\mu^{(h)}$, which represents traffic, measured in Erlangs. Equivalently, a call that keeps a channel busy during one hour features a 1 Erlang traffic.

Below we present a simple model for steady-state transition probabilities. In the telephony situation we address, usually the traffic is of high intensity, which leads easily to hundreds of calls in a one hour interval. Thus, the intensity is approximately constant in short intervals. This approximation in combination with large counts of calls per interval allows one to use hour-specific transition probabilities. With hour-specific probabilities we obtain hour-specific expected costs which we then add to obtain the expected cost of the project.

Figure 1 displays the state transition diagram for a system with $n = 3$ WBPs and $m = 2$ OFCs, with an infinite population of callers. With a state transition diagram such as the one in Fig. 1, one can write a set of linear equations for the steady-state probabilities associated with this system. A standard reference for state transition diagrams and their use to derive steady-state equations is Hebuterne (1987).

Let $i$ be the number of busy WBPs, and $j$ the number of busy OFCs. The model consists of a sequence of steady-state models, $h = 1, \ldots, H$, one for each time interval. Accordingly, let $p_{i,j}^{(h)} = P^{(h)}(X_B = i, X_O = j)$ be the probability that there are $i$ WBPs and $j$ OFCs busy at time $t$ in the $h$ time interval.

It is convenient to rearrange the $p_{i,j}^{(h)}$ probabilities into an $[(n+m) \times 1]$-vector $\boldsymbol{q}$. This rearrangement can use $k = (i+1) + j \cdot (n+1)$, $i = 0, \ldots, n$; $l = 0, \ldots, m$, to index the components of $\boldsymbol{q}$.

The number of states is $k_{\max} = (n+1) + m \cdot (n+1) = (m+1) \cdot (n+1)$. The first state is $(i, j) = (0, 0)$, which corresponds to $k_{\min} = 1$. The balance of the steady-state probabilities implies that there are $(m+1) \cdot (n+1) - 1$ linearly independent equations and one linearly dependent equation. One can replace one of these equations with the linearly independent equation,

$$\sum_{i=0}^{n} \sum_{j=0}^{m} p_{i,j}^{(h)} = 1 \quad (h = 1, \ldots, H), \tag{1}$$

to ensure that the solution is a probability distribution. For instance, the system that corresponds to the diagram of Fig. 1 is,

$$\begin{pmatrix}
\lambda & -\mu & 0 & 0 & -\mu & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-\lambda & \mu+\lambda & -2\mu & 0 & 0 & -\mu & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -\lambda & 2\mu+\lambda & -3\mu & 0 & 0 & -\mu & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -\lambda & 3\mu+\lambda & 0.0 & 0 & 0 & -\mu & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.0 & \mu+\lambda & -\mu & 0 & 0 & -2\mu & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -\lambda & 2\mu+\lambda & -2\mu & 0 & 0 & -2\mu & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -\lambda & 3\mu+\lambda & -3\mu & 0 & 0 & -2\mu & 0 \\
0 & 0 & 0 & -\lambda & 0 & 0 & -\lambda & 4\mu+\lambda & 0.0 & 0 & 0 & -2\mu \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0 & 2\mu+\lambda & -\mu & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda & 3\mu+\lambda & -2\mu & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda & 4\mu+\lambda & -3\mu \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix}$$

$$\times \boldsymbol{q} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \tag{2}$$

where we have omitted the $h$ index in $\lambda^{(h)}$ and $\mu^{(h)}$.

We next return to the description of the general model. The linear system is,

$$\boldsymbol{\Phi q} = \boldsymbol{e},$$

where $\boldsymbol{\Phi}$ is a $k_{\max} \times k_{\max}$ matrix and $\boldsymbol{e}$ is a $k_{\max} \times 1$ vector with zeroes everywhere except in the last row where it has a 1.

## 3 The expected cost function: definition and minimization

### 3.1 The expected cost

Let $T$ be the horizon (months) of the project and assume that at the end of the project the WBPs have zero residual value. Let $r$ be the monthly discount rate, $b$ the monthly traffic growth rate. Let $D$ be the number of workdays per month. To incorporate into the model the hour-to-hour traffic variability, let $E_h$ be the traffic intensity during the time interval $]h-1, h]$ (hour), $h = 1, 2, \ldots, H = 24$. Only for concreteness, we assume that the hourly traffic pattern $E_h$, $h = 1, 2, \ldots, H = 24$, is the same from day to day, but this assumption is not essential to our approach. The assumption that traffic intensity is constant during the time window $]h-1, h]$ implies that in this window, the calling rate $\lambda^{(h)}$ and the service rate $\mu^{(h)}$

are such that $E_h = \lambda^{(h)}/\mu^{(h)}$. The intensities $E_h$ may be estimated from the total duration of calls in a window $]h-1, h]$, as we did in the situation that motivated this research. More concretely, if $S_D^{(h)}$ is the aggregate duration of all calls in the one hour interval $]h-1, h]$, then $E_h$ may be estimated with $S_D^{(h)}/60$. With the number of calls during this interval, $N^{(h)}$, the expected duration $1/\mu^{(h)}$, may be estimated with $S_D^{(h)}/N^{(h)}$, whence an estimate of $\mu^{(h)}$. Finally, $\lambda^{(h)} = E_h\mu^{(h)}$ may be estimated with $N^{(h)}/60$.

With the assumptions that the traffic pattern is the same from day to day, that call durations are independent, and with data from a history of $K$ days, one can build an estimate of $E_h$ as follows: Let $k = 1, \ldots, K$ index the days, let $S_D^{(h,k)}$ be the aggregate duration of all calls in the window $]h-1, h]$ of day $k$. Then $S_D^{(h,+)} = \sum_{k=1}^{K} S_D^{(h,k)}$ is the aggregate duration of all calls in the windows $]h-1, h]$ of days $k = 1, \ldots, K$. Hence, $\hat{E}_h = S_D^{(h,+)}/(60 \cdot K)$ is an estimate of $E_h$. Notice that if $E_{h,k}$ is the traffic in window $]h-1, h]$ and if one uses $\hat{E}_{h,k} = S_D^{(h,k)}/60$ to estimate $E_{h,k}$, then the average of the $K$ daily estimates, namely $(\sum_{k=1}^{K} \hat{E}_{h,k})/K$, is precisely $\hat{E}_h$. Indeed, $\hat{E}_h = (\sum_{k=1}^{K} \hat{E}_{h,k})/K = (\sum_{k=1}^{K} S_D^{(h,k)}/60)/K = \sum_{k=1}^{K} S_D^{(h,k)}/(60 \cdot K)$.

If $N^{(h,k)}$ is the number of calls in the interval $]h-1, h]$ of day $k$, then the expected duration $1/\mu^{(h)}$ may be estimated with $S_D^{(h,+)}/N^{(h,+)}$, where $N^{(h,+)} = \sum_{k=1}^{K} N^{(h,k)}$, whence $\hat{\mu}^{(h)} = 1/(S_D^{(h,+)}/N^{(h,+)}) = N^{(h,+)}/S_D^{(h,+)}$ estimates $\mu^{(h)}$. Finally, $\lambda^{(h)} = E_h\mu^{(h)}$ may be estimated with $\hat{\lambda}^{(h)} = \hat{E}_h\hat{\mu}^{(h)}$.

We note that $\hat{E}_h$ is the maximum likelihood estimate (MLE) of $E_h$, $\hat{\mu}^{(h)}$ is the MLE of $\mu^{(h)}$, and $\hat{\lambda}^{(h)}$ is the MLE of $\lambda^{(h)}$, if the call durations are independent. These estimates have the support of statistical principles, and use all data available.

The transition probabilities $p_{i,j}^{(h)}$ depend on $E_h$. Hence we shall use $p(i, j, E_h, n, m)$ instead of $p_{i,j}^{(h)}$ if this precision is needed. In several formulas the number $m$ of OFCs is fixed. In these cases we use $p_m(i, j, E_h, n)$.

The expected present cost of the project is a function of $n$, the number of WBPs, and $m$, the number of OFCs. Calling fees are reported in USD/min. Let $V_B^{(h)}$ [USD/ min] be the cost of WBP's use and let $V_O^{(h)}$ [USD/min] be the cost of OFCs use. Let $C_B$ be the unit cost of a WBP unit, $C_O$ the unit cost of an OFC, and $C_I$ the fixed cost of installing the bypass system. The expected cost/min of connection time when $i$ WBPs and $j$ OFCs are in use during the interval $]h-1, h]$ is $c(i, j, h) = (iV_B^{(h)} + jV_O^{(h)})p^{(h)}(i, j, n, m)$. Hence, the expected associated cost of connection during the full one hour interval $]h-1, h]$ is $60 \times c(i, j, h)$. Let $r$ be the monthly discount rate. In the original situation we modeled, the organization that was evaluating the WBP project had an agreement with the telephony provider that prescribed a monthly reduction of fees. Let $b$ represents this monthly reduction of fees. The formula for the expected present cost simplifies if one defines $\theta = (1-b)/(1-r)$. Finally, the expected present cost function for the project is,

$$C(n \mid m) = 60 \cdot D \cdot \sum_{l=1}^{T} \left(\frac{1-b}{1+r}\right)^{l-1} \left\{ \sum_{h=1}^{24} \sum_{i=0}^{n} \sum_{j=0}^{m} \left(i \cdot V_B^{(h)} + j \cdot V_O^{(h)}\right) \cdot p\,(i, j, E_h, n, m) \right\} + n \cdot C_B + m \cdot C_O + C_I \tag{3}$$

$$= 60 \cdot D \cdot \sum_{l=1}^{T} \theta^{l-1} \left\{ \sum_{h=1}^{24} \sum_{i=0}^{n} \sum_{j=0}^{m} \left(i \cdot V_B^{(h)} + j \cdot V_O^{(h)}\right) \cdot p\,(i, j, E_h, n, m) \right\} + n \cdot C_B + m \cdot C_O + C_I. \tag{4}$$

In (4) the sum $\sum_{l=1}^{T} \theta^{l-1}$ is $L = (1-\theta^T)/(1-\theta)$. The most common practical situation is that of a PABX with only internal extensions and $m$ trunks in place. In this situation, the project consists of evaluating the expected savings associated with the purchase of $n$ WBPs. For this reason we proceed with a fixed number of $m$ OFCs. Then (4) becomes,

$$C(n|m) = 60 \cdot D \cdot L \cdot \left\{ \sum_{h=1}^{24} \sum_{i=0}^{n} \sum_{j=0}^{m} \left( i \cdot V_B^{(h)} + j \cdot V_O^{(h)} \right) \cdot p_m(i, j, E_h, n) \right\} + n \cdot C_B + m \cdot C_O + C_I. \tag{5}$$

The optimal value of $n$, call it $\hat{n}$, minimizes $C(n|m)$. Our simple minimization algorithm uses $\Delta C(n|m) = C(n+1|m) - C(n|m)$, which is,

$$\Delta C(n|m) = C_B + 60 \cdot D \cdot L \cdot \sum_{h=1}^{24} \sum_{j=0}^{m} \left[ \sum_{i=0}^{n+1} \left( i_B^{(h)} + j V_O^{(h)} \right) \cdot p_m(i, j, E_h, n+1) - \sum_{i=0}^{n} \left( i V_B^{(h)} + j V_O^{(h)} \right) \cdot p_m(i, j, E_h, n) \right]. \tag{6}$$

The expected traffic for one day of operation is,

$$T(n|m) = \sum_{h=1}^{24} \sum_{i=0}^{n} \sum_{j=0}^{m} E_h \cdot (i+j) \cdot p_m(i, j, E_h, n). \tag{7}$$

The search for $\hat{n}$ can be sped-up with a strategy that provides a starting point. The next subsection describes this strategy.

### 3.2 A simple strategy to obtain a starting point

To determine an initial value $n_0$ for the search of the minimizer $\hat{n}$ of $C(n)$, we develop a simpler problem with only one state-variable, the number of WBPs. We derive the expected cost function $C^{(1)}(n)$ for this simpler problem, whose minimization requires many fewer computations than the minimization of $C(n)$, and use its minimum as an initial value to minimize $C(n)$.

This development starts with a stochastic model for the number of busy channels during the interval $]h-1, h]$, $h = 1, 2, \ldots, 24$. Let $n \geq 0$ be the number of WBP channels installed. This simplified model assumes that there is an unlimited number of OFCs available to carry traffic when no WBPs are available.

Let $X^{(h)}(t)$ be the number of busy channels at time $t$ during the one hour window $]h-1, h[$. Therefore the number of WBP channels busy at time $t$ is,

$$Y^{(h)}(t) = \begin{cases} X^{(h)}(t), & \text{if } X^{(h)}(t) < n, \\ n, & \text{if } X^{(h)}(t) \geq n. \end{cases}$$

We use the working assumption that the expected number of busy channels is approximately constant during $]h-1, h]$, thus we assume, $E(X^{(h)}(t)) = \mathcal{E}_h$, but the $\mathcal{E}_h$, $h = 1, \ldots, 24$, may differ from one another. If this assumption is inadequate, then the analysis may use windows shorter than one hour. In certain situations, past data may suggest that it is adequate to

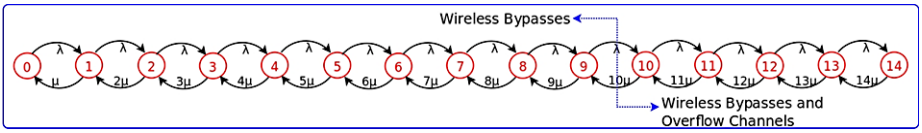


**Fig. 2** Simplified model for 14 channels with negligible traffic to fixed phones

assume that $E(X^{(h)}(t))$ is approximately constant for windows longer than one hour. Certainly, our approach does not require $E(X^{(h)}(t))$ to be approximately constant within one hour windows. We use this assumption for concreteness only. Note that if $\mathcal{E}(t) = E(X^{(h)}(t))$ is continuous for $t \in ]0, 24]$ [h] then $\mathcal{E}(t)$ is uniformly continuous and hence can be arbitrarily well approximated with a step function. This rational supports our working assumption that $E(X^{(h)}(t))$ is constant, $\mathcal{E}_h$, for $t \in ]h-1, h]$. One may assess whether the window width, call it $w$, is sufficiently small. For the specific objective of determining the optimal number of WBPs, one can proceed in a practical fashion as follows: select a starting value of $w_1$ for the window width and then determine the optimal number $n(w_1)$ of WBPs. Next, select a smaller starting window width $w_2 < w_1$, and compute the corresponding optimal number of WBPs, $n(w_2)$. If $n(w_1)$ and $n(w_2)$ do not differ significantly then $n(w_1)$ may be used as an initial value for the minimization of $C(n|m)$ in (5). Otherwise, set $w_1 \longleftarrow w_2$ and set a new smaller value for the window $w_2$. One continues with this procedure until $n(w_1)$ and $n(w_2)$ are satisfactorily close. It is important to note that the solution to this sub-section's simplified problem need not be computed exactly because it will be used only as an initial value for the minimization of $C(n|m)$ in (5).

We complete the specification of our simplified model with the assumption that $X^{(h)}(t)$ follows a Poisson distribution with parameter $\mathcal{E}_h$. This specification agrees with the $M/M/\infty$ queue model in steady-state with unlimited service within each window $]h-1, h]$, where the number of channels in use $X^{(h)}(t)$, follows a Poisson distribution. In this specification, if $\lambda^{(h)}$ is the arrival rate and $\mu^{(h)}$ is the service rate for window $[h, h+1[$, then $\mathcal{E}_h = \lambda^{(h)}/\mu^{(h)}$ (see, Gross and Harris 1985). The queue model $M/M/\infty$ views the number of individuals in the queue as the population size in a birth and death process. The problem that concerns us parallels the $M/M/\infty$ model, where the population size corresponds to the number of busy channels $X^{(h)}(t)$.

In the example of Fig. 2, there are 14 channels available where the first nine represent WBPs that differ from the five OFCs in their associated traffic and costs.

Therefore, the probability function of the number of WBPs in use at time $t \in ]h-1, h]$ is,

$$p^*(i, \mathcal{E}_h) = P(Y^{(h)}(t) = i, \mathcal{E}_h) = \begin{cases} \frac{e^{-\mathcal{E}_h}\mathcal{E}_h^i}{i!}, & i = 0, 1, 2, \ldots, n-1, \\ \sum_{j=n}^{\infty} \frac{e^{-\mathcal{E}_h}\mathcal{E}_h^j}{j!}, & i = n. \end{cases} \tag{8}$$

With $n$ WBPs the expected present cost of establishing a WBP consists of two terms,

$$C^{(1)}(n) = \text{Cost of establishing the WBP}(n) + \text{Cost of traffic}(n). \tag{9}$$

Fees and traffic vary hourly, hence to obtain the total cost one must add cost terms specific for $h = 1, \ldots, 24$.

As the bypass channels have priority for calls to wireless phones, the expected traffic during $t$ in $]h-1,h]$ in the WBPs is,

$$T^{(1)}(n)=\sum_{i=0}^{n} ip(i,\mathcal{E}_h)+\sum_{i=n+1}^{\infty} np(i,\mathcal{E}_h)=\sum_{i=0}^{n} ip^*(i,\mathcal{E}_h). \tag{10}$$

During the one-hour interval $]h-1,h]$, the billed traffic is,

$$C_{\mathrm{WBP}}^{(1)(h)}(n)=60V_B^{(h)}\sum_{i=0}^{n} i\cdot p(i,\mathcal{E}_h)+60nV_B^{(h)}\sum_{i=n+1}^{\infty} p(i,\mathcal{E}_h)=60V_B^{(h)}\sum_{i=0}^{n} i\cdot ip^*(i,\mathcal{E}_h). \tag{11}$$

On the other hand, when one adds OFCs to serve all overflow calls, then the overflow traffic through OFCs is,

$$C_{\mathrm{OFC}}^{(1)(h)}(n)=60V_O^{(h)}\sum_{i=n+1}^{\infty}(i-n)\cdot p(i,\mathcal{E}_h). \tag{12}$$

Then, with (11) and (12) one obtains,

$$\text{Cost of traffic}(n)=C_{\mathrm{WBP}}^{(1)}(n)+C_{\mathrm{OFC}}^{(1)}(n)=D\cdot L\sum_{h=1}^{24}\Big[C_{\mathrm{WBP}}^{(1)(h)}(n)+C_{\mathrm{OFC}}^{(1)(h)}(n)\Big]. \tag{13}$$

Finally, the present cost of the project is,

$$C^{(1)}(n)=n\cdot C_{\mathrm{WBP}}+\text{Cost of traffic}(n). \tag{14}$$

### 3.3 Expected cost minimization

We find the minimum $\hat{n}$ searching for the least $n$ that satisfies,

$$\Delta C^{(1)}(n)=C^{(1)}(n+1)-C^{(1)}(n)\geq 0. \tag{15}$$

This is appropriate with the assumption that $C^{(1)}(n)$ decreases from $n=0$ to the minimum and then increases.

$$\Delta C^{(1)}(n)=C_n+60DL\sum_{h=1}^{24}\big\{V_B^{(h)}p(n+1,\mathcal{E}_h)-V_O^{(h)}p(n+1,\mathcal{E}_h)\big\}. \tag{16}$$

## 4 An example

Qualitatively, this example resembles the situation in the organization where this problem arose. Figure 3 exhibits the average traffic. For ease of explanation we assume that the expected traffic pattern is the same day to day and that a step function with 1 hour intervals provides and adequate approximation for it. Cost per minute of wireless to wireless calls is USD 0.944 at any hour of the day, while cost per minute of land-line to wireless calls is USD 0.36 from 8 until 20 hours of each work day, and USD 0.185 the rest of the day. Service rate is assumed 1 [min$^{-1}$], so that the arrival call rate is readily obtained from the traffic intensity profile. An agreement with the telephony provider prescribed a monthly reduction

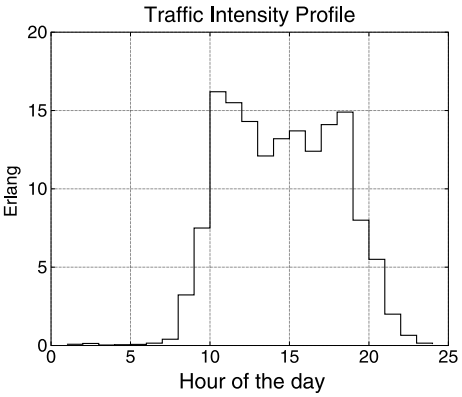


**Fig. 3** Traffic intensity profile

rate of fees of $b = 0.3\%$. We use $r = 0.4\%$ for the monthly discount rate. The total number of working days per month is $D = 22$ and the project horizon is $T = 24$ months. The unit cost of a WBP is USD 500 and the unit cost of each OFC is a sunk cost. At the time a project like this one is evaluated, the organization has in place a sufficiently large number of OFCs, say $m_a$, to serve most of the traffic. With a personal computer, finding the minimizer of the expected cost (5) takes an impractically long time. Thus, one may consider solving the problem with a smaller value of $m$, say $m_l$. As the discussion bellow shows, one must be careful to interpret the resulting optimal solution that uses $m_l$. We advance, that the key to obtain a good approximation for the minimizer of (5) is to use a value $m_l$, that ensures only a small loss of served traffic compared to the traffic served with the actual $m_a$.

We analyze cases with two values for $m_l$: one with $m_l = 8$ OFCs and a second one with $m_l = 5$ OFCs. Near the optimal number $\hat{n}$, one expects a low usage of OFCs because WBPs have priority over OFCs. Figure 4 exhibits the expected present cost functions for both cases. With 8 OFCs, the optimal number of WBPs is $\hat{n} = 26$ and the present cost value is USD 680.0 thousand. With 5 OFCs, the optimal number of WBPs is also $\hat{n} = 26$ and the present cost value is USD 679.8 thousand.

One notices that in both plots of Fig. 4 the functions behave similarly near their respective minima ($n = 26$), but behave quite differently in the range of small numbers of WBPs.

More specifically, with $n = 5$ WBPs and $m = 5$ OFCs (Fig. 4, right panel) the expected cost is about 800 [$10^3$ USD], whereas with 8 OFCs (Fig. 4, left panel) the expected cost is about 1,000 [$10^3$ USD]. Thus, it would appear, paradoxically, that with fewer OFCs one attains a lower cost. To explain this situation one must attend to the loss of traffic served when reducing the number of OFCs from 8 to 5. For instance, with 5 OFCs (Fig. 5, right panel) and $n = 5$ WBPs, the expected daily traffic served is about 90 Erlang, whereas with 8 OFCs (Fig. 5, left panel) and $n = 5$ WBPs, the expected daily traffic served is about 118 Erlang.

Figure 5 displays the traffic as a function of the number of WBPs for 5 OFCs and 8 OFCs. One sees that with the minimum number of WBPs, the number of OFCs needed is very small. Usual quality of service standards, state that during the peak hour, the loss rate should not be greater than 2%. This requirement is similar to imposing a maximum traffic loss of 0.1% during the full day. In this case, while the total traffic for 5 OFCs needs more than 22 WBPs, with 8 OFCs, 19 WBPs suffice to meet the required quality of service standard.

With the simplified one state-variable model, the algorithm finds quickly the minimizer to obtain an initial value for the two state-variable model. In fact, the initial guess suggests

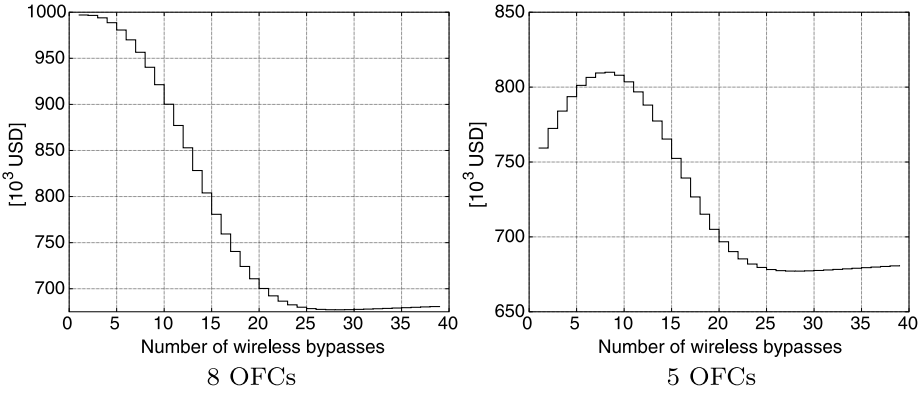


**Fig. 4** Expected cost functions ($10^3$ [USD]) vs. number of wireless bypasses

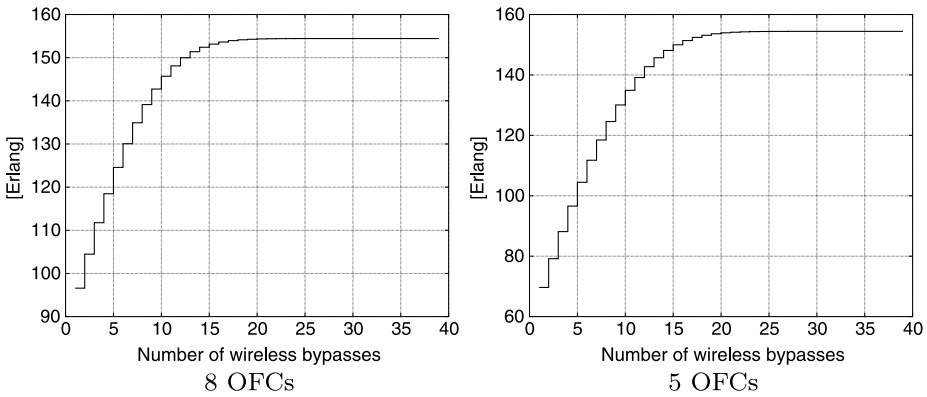


**Fig. 5** Expected traffic intensity for one work day vs. number of wireless bypasses

starting the search of $\hat{n}$ in the neighborhood of 22 WBPs. This simplified model underestimates the expected cost in USD 224 thousand, which strongly suggests one to use next the two state-variable model in combination with the use of the initial guess. We emphasize that if one uses QoS standards as criterion instead of expected cost, then one would have been lead to much higher costs. For example, and taking data from Fig. 4, with 8 OFCs in place, the total cost is $C(19|8) =$USD 712.0 thousand. With 5 OFCs, the total cost is $C(22|5) =$ USD 684.4 thousand. Far worse, the QoS standard would have been met with 27 OFCs and 0 WBPs, at a very high cost, $C(27|0) =$ USD 1.51 million. The proposals from the use of QoS standards remain undefined: which one of the 28 choices of $(n, m)$, $n, m = 0, 1, \ldots, 27$, $n + m = 27$, is the best one?

## 5 Concluding remarks

A communications manager must be aware of the consequences of using a non-optimal number of WBPs. In fact, after installing WBPs an organization may experience a seemingly paradoxical cost increase for calls to wireless phones as the following situation shows.

Suppose there are $m = 5$ OFCs available and then one WBP is added. One might expect a cost reduction associated with calls to wireless phones because the WBP may now carry calls at a cheaper rate that before were carried through an OFC. However, if the $m = 5$ OFCs did not satisfy demand prior to the installation of the WBP, then the WBP may now carry calls that before the OFCs denied service to. That is, this additional WBP does not replace an OFC but carries additional calls. This increased cost pays for additional quality of service. WBPs start replacing OFCs to carry calls only when the number of WBPs is sufficiently large so that both OFCs and WBPs satisfy most of the demand. Figures 4 and 5 illustrate this point. Each figure compares two scenarios, one with $m = 8$ OFCs (left panels), and the other with $m = 5$ OFCs (right panels). Figure 4 displays expected cost function and Fig. 5 displays traffic intensities. With 5 OFCs, the right panel in Fig. 4 shows that the cost increases as the number of WBP increases from 0 to 7. In this range, traffic intensity also increases sharply (Fig. 5, right panel). With $m = 8$ OFCs the expected cost (Fig. 4, right panel) decreases as $n$ increases from zero to $\hat{n} = 26$ WBPs.

This paper opens a number of interesting lines for future research. Here we briefly touch on just a few of these. First, a more comprehensive model should incorporate the value of lost calls into the model and quality of service (probability of service denial) constraints. These models may further increase the organization's profit. Second, some situations may concern more than two types of lines, or types of rental items. For instance, a library may contemplate to offer online access to a limited number of readers, and make available a given number of paper copies. Additional copies are available via interlibrary loans. Third, in some organizations the population of callers is finite. In such cases, one may model individual calling patterns to take into account type of callers, schedules, or other caller-specific characteristics. Fourth, the OFCs may be used to carry calls to both wireless and land-line phones. Modeling a system with shared use of overflow channels may requiere many state-variables. A related problem arises when a set of shared lines serves as overflow to a group of priority land-lines. Fifth, certainly it is of interest to develop tractable models for the times between calls and for the duration of calls, other than the exponential distributions that this paper uses. Sixth, the problems where the decision variables are the number of WBPs and the number of OFCs. This situation is relevant when one plans a telephony system prior to the start of its operation. Seventh, the model may incorporate additional phenomena that take place in telephony systems: periods of busy signals and "no answers" (NA). Especially in organizations that insistently attempt to contact customers, busy signals or NAs lead to later retrials. Retrials hold channels unavailable to other callers. If the aggregated hold-up time is significant, it may affect the optimal number of bypasses. In addition, in countries that use CPP, charges may not apply until the receiver answers the call. Thus, the time the channels are busy and the time with traffic subject to billing may be very different. These considerations lead to interesting practical modifications to our method that one may consider to accommodate special characteristics of a telephony system. Finally, There are some connections between this article and problems where service capacity of a server must be switched between customers. Rosa-Hatko and Gunn (1997) review methods and areas of applications. These connections can be helpful to identify additional applications for the approach we have presented here. In the situation we model the telephone lines available, either OFCs or WBPs, may be viewed as a service facility with two servers: the first is the WBP system, with limited capacity, and the second consists of the OFCs. In Rosa-Hatko and Gunn (1997) there is a number of type of customers and the server switches modes according to the customer's type. In the situation we model there are no waiting lines, there is a single type of customers, but at the time of requesting service, that is at the time of placing a call to a wireless phone, the customer may be switched from the WBP system to

the OFCs if the capacity of the former is found to be exhausted. It is worth noting that our approach can be readily adapted to accommodate types of customers with calling and the service rates, $\lambda^{(h)}$ and $\mu^{(h)}$, that depend on customers' covariates.

An important part of the literature on the management of telephony operations concentrate on call and contact centers where most of the traffic is inbound. Koole and Mandelbaum (2002) provide a thorough and insightful review of queueing models for call centers. There are parallels between inbound/staffing problems and outbound/number-of-channels problems as staff size and number of channels represent available servers. These parallels can be expoited to advance models for outbound/number-of-channels situations, but as we have discussed, in this latter situation expected cost seems to be a more appropriate criterion to select number-of-channels than a QoS criterion.

This article considers a problem of salient importance both in telephony and telecommunications engineering as in management practice. The problem we addressed emerged in an organization that faced the decision to buy wireless bypasses to save on expensive calls to mobile phones in the future. One of the features that sets apart the situation we model is the availability of two types of lines, WBP and OFC, the different costs associated with the use of each of them, and the consequent priority of a WBP over an OFC. A more familiar and analogous situation is that of a library that has $n$ copies of a volume available for lending. When none of these are available, the library can make available to its users up to $m$ additional copies via inter-library loans, but at a higher cost. Thus, in different guises this same problem may arise in applications where an agent or a business holds an inventory of items to rent or lend, and the agent can procure additional items at higher prices when his inventory is insufficient to satisfy demand.

Our approach considers a model with two state-variables. One state-variable is the number of bypasses in use and the second is the number of overflow lines in use. Examining the state transition diagram, we develop a model for the steady-state transition probabilities in terms of parameters of the distributions of the times between calls that arrive at the system and the duration of the calls. With direct estimates of these parameters one estimates the transition probabilities and then writes the expected cost for a given time horizon. The optimal number of bypasses minimizes the expected cost. Additionally we propose a strategy to select an initial guess to start the search for the optimum. This strategy uses the expected cost of a simpler model with a single state-variable. The minimizer of this simpler problem serves as initial guess for the two state-variables. The simpler model may itself be appropriate for certain situations.

Examples illustrate the economic value of our approach, especially when one compares the reduction of costs our method provides with the costs associated with a Quality of Service criterion. We have used our method in a commercial Bank. The method provided very significant costs reductions, by aiding the telecommunications manager to decide on a single key element: the number of bypasses to purchase.

**Acknowledgements** The authors would like to thank the editors and two anonymous reviewers for their constructive and insightful comments.

## References

Gross, D., & Harris, C. M. (1985). *Fundamentals of queueing theory* (2nd ed.). New York: Wiley.
Hebuterne, G. (1987). *Traffic in switching systems*. Norwood: Artech House.
Jain, H. C. (1966). An inventory problem applied to a rental situation. *Australian Journal of Statistics*, *8*(3), 154–162.

Koole, G., & Mandelbaum, A. (2002). Queueing models of call centers: an introduction. *Annals of Operations Research*, *113*, 41–59.

Rosa-Hatko, G., & Gunn, E. A. (1997). Queues with switchover: a review and critique. *Annals of Operations Research*, *69*, 299–322.

Tainiter, M. (1964). Some stochastic inventory models for rental situations. *Management Science*, *11*(2), 316–326.

Whisler, W. D. (1967). A stochastic inventory model for rented equipment. *Management Science*, *13*(9), 640–647.