Pricing a threshold-queue with hysteresis

Louis-Marie Le Ny[1] and Bruno Tuffin[2]

[1] IRISA-Université de Rennes 1, Campus universitaire de Beaulieu
35042 Rennes Cedex, France
leny@irisa.fr
[2] IRISA-INRIA, Campus universitaire de Beaulieu
35042 Rennes Cedex, France
btuffin@irisa.fr
WWW home page:
http://www.irisa.fr/armor/lesmembres/Tuffin/Tuffin_en.htm

Abstract. In this paper, we consider pricing schemes at an M/M/1 queue with infinite potential demand but where the number of customers in the queue depends on both the price and the offered quality of service (QoS). Our model aims at comparing the optimal revenue for three different strategies: first the case of a fixed price; second the case where there is a threshold on the queue occupancy such that a larger charge is imposed for an occupancy above the threshold (in order to maintain a given QoS); third the case of a threshold-queue with hysteresis in order to avoid costly oscillations around the aforementioned threshold. In all those three situations, we determine the equilibrium number of customers, and the parameters (price(s), threshold(s)) optimizing the revenue. We can therefore find the optimal strategy from the provider's point of view. In this paper, we use a static demand because session lengths are assumed large with respect to the queue dynamics and therefore consider the system in steady-state. We then show that, contrary to what could be expected, a policy without threshold is recommended.

1 Introduction

In many situations, users sharing a common facility can be modeled as queueing systems. This occurs in manufacturing, computer science, transport (airplanes on a runaway, cars on a highway...) for instance, but also in communication networks where data are processed through routers.

In those cases, as the number of users increases, congestion occurs, resulting in a decrease in terms of offered quality of service (QoS). As a consequence less customers will be willing to use the facility, which should decrease demand. This should lead to an (uncontroled) equilibrium.

Pricing is a common way for the facility owner (the service provider in the case of the Internet) to control demand and provide return on investment. Determining the price selection is of interest since a too high price would lead to no customer and no profit, while a too low price would lead to a too low revenue, with the number of customers driven only by the offered QoS.

In this paper, we investigate the impact of several pricing policies on the maximal revenue of the seller (the facility owner) where the facility is modelled by an M/M/1 queue. We study three different policies. In the first one (the *fixed policy*), the price is fixed. In the second one (the *threshold policy*), there is a threshold on the queue occupancy such that a higher price is charged above the threshold in order to limit low QoS levels. In the third case (the *hysteresis policy*), we introduce hysteresis for switching between low and high prices, that might be costly from a signalling and/or engineering point of view. In each case, we determine the equilibirum number of customers in the queue for fixed prices and thresholds (if any). We then determine the prices and thresholds leading to a maximum revenue. This helps the facility owner in determining the policy to use. Our model is based on the assumption that when a flow sends a packet, it does not care about the current state of the queue because session lengths are supposed to be long with respect to the transient evolution of the queue, meaning that it better considers the queue in steady-state to make a decision about sending traffic or not. We show that under this assumption, the particular case of a policy without threshold is the most efficient.

Pricing has already been studied for different scheduling policies, especially in the case of heterogeneous users [1–3], but to our knowledge there exists no investigation relating pricing, quality of service and threshold policies.

The paper is organized as follows. We present the basic model in Section 2. Section 3 deals with the *fixed policy* and determines the optimal revenue. Section 4 is for the *threshold policy* while Section 5 is for the *hysteresis policy*. Section 6 makes a formal comparison between the three policies, and Section 7 is devoted to conclusions and directions for future research.

2 Basic model

In what follows we will have in mind the facility as a network router and users being networking data flows.

Consider an M/M/1 queue with service rate μ and where users generate packets according to a Poisson process with rate λ. We consider an infinite population of infinitesimal potential users. Each of them applies (selfishly) for service as soon as his utility is positive. The utility function characterizes the users behaviour and is a well-known notion in economics and game theory in general [4, 5]. The *utility* a user gets depends on the mean packet delay $\bar{D}$, but also on the *average* price per packet/customer $\bar{p}$ (so that users do not know the instantaneous network status, but rather work based on average value that they estimate, which is a valid assumption for long sessions):

$$U = f(\bar{D}) - \bar{p}$$

where f is a strictly decreasing function representing the valuation of a user for a QoS of interest, the delay (or response time). Note that it would seem interesting to add the jitter, that is the delay variance, to the utility function

(as a dis-utility), but since for an M/M/1 queue delay variance equals delay, it can be incorporated in function f without loss of generality.

Note also that, from our notion of utility and the analysis that will follow, users do not care about the state of the queue when they enter, they are only interested in average value, meaning that they are active for a long time.

Therefore users/sources send packets as soon as their utility is positive. A first point will be to find out wether an equilibrium number of sources N^* exists or not, and if it is unique in that case.

On the other hand, from the network provider's point of view, the goal is to maximize the revenue $\mathcal{R} = \lambda \bar{p} N^*$ minus the operational cost κO_{sc}, which will be further defined later, but basically representing the cost of switching between prices, at the packet/customer level.

Throughout the paper, we will especially focus on the special case $f(\bar{D}) = 1/\bar{D}^\beta$ with $\beta > 0$.

3 The M/M/1 queue without threshold

Consider the classical M/M/1 queue and a fixed per packet price p (so that $\bar{p} = p$).

If there are N active sources, then the packet arrival rate is $N\lambda$, leading to a traffic load $\rho = \frac{\mu}{N\lambda}$ and an average delay $\frac{1}{\mu - N\lambda}$.

The utility function

$$U(N) = f\left(\frac{1}{\mu - N\lambda}\right) - p$$

is strictly decreasing with the number N of sources.

Given the infinite potential number of sources, N will increase if $U(N) > 0$ and decrease if it is strictly negative and $N > 0$. As a consequence, there is a unique equilibrium number N^*, verifying $U(N^*) = 0$ if this equation has a (then unique) solution, and $N^* = 0$ if $U(0) < 0$. This is summarized in the following proposition.

Proposition 1. *There is a unique equilibrium number of sources given by*

$$N^* = \begin{cases} \frac{1}{\lambda}\left(\mu - \frac{1}{f^{-1}(p)}\right) \text{ if } \mu - \frac{1}{f^{-1}(p)} \geq 0, \\ 0 \qquad\qquad\qquad\qquad \text{otherwise.} \end{cases}$$

The solution $N^* = 0$ corresponds to the case where the price p is so high that no users will use the queue.

Since the price is fixed, there is no operational cost for switching between prices, and the goal of the provider is this to optimize the network revenue $\mathcal{R} = \lambda p N^*$.

Proposition 2. *When $f(\bar{D}) = 1/\bar{D}^\beta$, the optimal price is*

$$p^* = \left(\frac{\mu\beta}{\beta + 1}\right)^\beta$$

resulting in the optimal revenue $\frac{\mu}{\beta+1}\left(\frac{\mu\beta}{\beta+1}\right)^{\beta}$.

Proof. Replacing N^* by its value, we have $\mathcal{R} = \mu p - p^{1+1/\beta}$. Therefore $\frac{\partial \mathcal{R}}{\partial p} = \mu - (1+1/\beta)p^{1/\beta} = 0$ immediately gives the result. □

4 The M/M/1 threshold queue

Consider the same M/M/1 queue, with a threshold K_s such that the price is P_L if there are $n < K_s$ packets in the queue, and $p_U > p_L$ otherwise.

We still have $\bar{D} = \frac{1}{\mu - N\lambda}$ when there are N active sources, since those users are only interested in average values (and as a consequence arrival and departure rates do not depend on the state of the queue). Remark again that the utility of users/sources depends only on the average price and delay instead of instantaneous ones (the reason why we can still use the M/M/1 formula); this assumption is especially valid in the case of long file tranfers for instance.

Remind that the steady-state probability of having n packets in the queue is $\pi_n = (1-\rho)\rho^n$, with again $\rho = \frac{N\lambda}{\mu}$. The average price is thus

$$\begin{aligned}\bar{p} &= p_L \sum_{n=0}^{K_s-1} (1-\rho)\rho^n + p_U \sum_{k=K_s}^{\infty} (1-\rho)\rho^n \\ &= p_L(1-\rho^{K_s}) + p_U\rho^{K_s} \\ &= p_L + (p_U - p_L)\rho^{K_s}.\end{aligned}$$

Remark that $\bar{p}$ increases with N.

From now on, we let $f(\bar{D}) = 1/\bar{D}^{\beta} = (\mu - N\lambda)^{\beta}$.

The utility of each source is then given by

$$U(N) = (\mu - N\lambda)^{\beta} - p_L - (p_U - p_L)\left(\frac{N\lambda}{\mu}\right)^{K_s}$$

which is strictly decreasing with N. We therefore have the following result.

Proposition 3. *There is a unique equilibrium number N^* of sources given by the unique solution of equation $(\mu - N^*\lambda)^{\beta} - p_L - (p_U - p_L)\left(\frac{N^*\lambda}{\mu}\right)^{K_s} = 0$ if $\mu^{\beta} - p_L \geq 0$, and 0 otherwise.*

Proof. Again, if we study the dynamics of the model with an infinite potential number of customers, the number N of sources that will send packets increases if $U(N) > 0$ and decreases if it is strictly negative and $N > 0$. From the decreasingness of the utility function, there is a unique equilibrium number of sources N^*, verifying $U(N^*) = 0$ if this equation has a (then unique) solution, i.e. if $U(0) = \mu^{\beta} - p_L \geq 0$ and $N^* = 0$ if $U(0) < 0$. This gives the proposition. □

In a second part, the goal of the service provider is to optimize its benefits, that is its revenue $\mathcal{R} = \lambda\bar{p}N^*$ minus its operational costs κO_{sc}, with O_{sc} the

steady-state frequency of price oscillations between p_U and p_L per unit of time, $O_{sc} = N^*\lambda\pi_{K_s-1} + \mu\pi_{K_s}$ and κ the cost for each such individual operation.

This optimization has to be carried out over K_s, p_L and p_U, using the value of N^* obtained in Proposition 3.

5 The M/M/1 threshold queue with hysteresis

We now assume that in order to avoid costly and cumbersome oscillations around the threshold of the previous section, there is a forward threshold K_U and a backward thershold K_L. When the number of customers in the system goes beyond the forward threshold K_U, the price is immediately upgraded to p_U. Similarly, when the number of customers falls at the reverse threshold K_L (with $K_L < K_U$), it is decreased to p_L. In between, the price is not changed, avoiding oscillations. Following the analysis in [6], the stationary occupancy distribution can be determined. The system is modelled by a Markov chain defined over the state space $\mathcal{S} = \{(n,k)|0 \leq n \leq K_u \text{ if } k = 0,\ K_L + 1 \leq n \text{ if } k = 1\}$. In this characterization, n is the number of customers in the queue while $k = 0$ corresponds to the case where the price is p_L and $k = 1$ to the case where it is p_U (see Figure 1). As a consequence, the corresponding state when there are $K_L + 1 \leq n \leq K_U$ customers depends on which threshold was reached last.

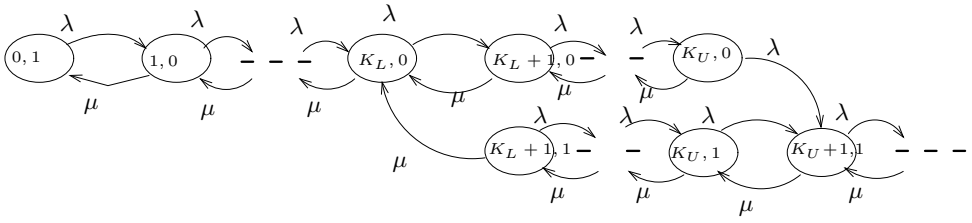

Fig. 1. State spac for the threshold-queue with hysteresis. For states tagged by 0, the price is p_L while it is p_U for states tagged by 1.

From [6], the steady-state probabilities are, with $\rho = \frac{N\lambda}{\mu}$ and $\alpha = \frac{\rho^{K_U}(1-\rho)}{1-\rho^{K_U-K_L+1}}$:

$$\begin{aligned}
\forall 0 \leq n \leq K_L,\ \pi_{n,0} &= \rho^n \pi_{0,0} \\
\forall K_L < n \leq K_U,\ \pi_{n,0} &= \left(\rho^n - \frac{\rho^{K_U+1}(1-\rho^{n-K_L})}{1-\rho^{K_U-K_L+1}}\right)\pi_{0,0} \\
\forall K_L + 1 \leq n \leq K_U + 1,\ \pi_{n,1} &= \alpha\rho\frac{1-\rho^{n-K_L}}{1-\rho}\pi_{0,0} \\
\forall n \geq K_U + 2,\ \pi_{n,1} &= \alpha\rho\frac{\rho^{n-K_U-1}(1-\rho^{K_U+1-K_L})}{1-\rho}\pi_{0,0}.
\end{aligned}$$

By straightforward computations (or by noting that it also correspond to an M/M/1 queue if we aggregate the states in terms of occupancy), it can be verified that $\pi_{0,0} = 1 - \rho$.

Therefore, the average price is given by:

$$\begin{aligned}
\bar{p} &= p_L \sum_{n=0}^{K_U} \pi_{n,0} + p_U \sum_{n=K_L+1}^{\infty} \pi_{n,1} \\
&= (1-\rho) \left[p_L \sum_{n=0}^{K_L} \rho^n + p_L \sum_{n=K_L+1}^{K_U} \left(\rho^n - \frac{\rho^{K_U+1}(1-\rho^{n-K_L})}{1-\rho^{K_U+1-K_L}} \right) \right. \\
&\quad \left. + p_U \sum_{K_U+1}^{\infty} \rho^n + p_U \sum_{n=K_L+1}^{K_U} \frac{\rho^{K_U+1}(1-\rho^{n-K_L})}{1-\rho^{K_U+1-K_L}} \right]. \\
&= (1-\rho) \left[p_L \sum_{n=0}^{K_U} \rho^n + p_U \sum_{K_U+1}^{\infty} \rho^n \right. \\
&\quad \left. + (p_U - p_L) \sum_{n=K_L+1}^{K_U} \frac{\rho^{K_U+1}(1-\rho^{n-K_L})}{1-\rho^{K_U+1-K_L}} \right].
\end{aligned}$$

Finally,

$$\bar{p} = p_L + (p_U - p_L)(K_U - K_L + 1) \frac{\rho^{K_U+1}(1-\rho)}{1-\rho^{K_U-K_L+1}}. \tag{1}$$

Again, the utility for N sources is given by

$$U(N) = (\mu - N\lambda)^\beta - \bar{p}.$$

Also, this function is strictly decreasing with N: $(\mu - N\lambda)^\beta$ decreases with N (as soon as $N < \mu/\lambda$), as well as $\bar{p}$ since in (1),

$$\frac{\rho^{K_U+1}(1-\rho)}{1-\rho^{K_U-K_L+1}} = \frac{1}{\sum_{k=0}^{K_U-K_L} \rho^{k-K_U-1}}.$$

Indeed, as $k - K_U - 1 < 0 \ \forall k \leq K_U - K_L$, $\bar{p}$ is increasing with N, which shows the result.

As a consequence, we have again the steady-state result:

Proposition 4. *There is unique equilibrium number of sources given by the unique solution N^* of equation $(\mu - N^*\lambda)^\beta - p_L - (p_U - p_L)(K_U - K_L + 1)\frac{\rho^{K_U+1}(1-\rho)}{1-\rho^{K_U-K_L+1}} = 0$ if $\mu^\beta - p_L \geq 0$, and 0 otherwise.*

Proof. The proof follows by exactly the same arguments that in previous sections. □

Again, the goal of the provider is to optimize its average benefits per unit of time:

$$\lambda \bar{p} N^* + \kappa(N^* \lambda \pi_{K_U,0} + \mu \pi_{K_L+1,1})$$

over all possible values of K_L, K_U, p_L and p_U.

6 Comparison of the three policies

The aim of this section is to show that the optimal revenue is obtained in the special case of no threshold, even if the other strategies are generalizations (which cannot lead to a smaller optimal revenue).

Let's forget about the switching cost in thresholds policies. The revenue for the three policies is then given by $\mathcal{R} = \lambda \bar{p} N^*$, with optimal N^* and $\bar{p}$ that may be different. Note that N^* is a function of $\bar{p}$, by, in all three cases, $U(N) = f(\bar{D}) - \bar{p}$ becoming zero for N^*. We therefore have with exactly the same functional dependence because average delay $\bar{D}$ is given by the same formula for the three policies.

As a consequence, in the most general case of thresholds with hysteresis, every pair of price p_U and p_L and every threshold value K_s, or K_U and K_L, will lead to a given $\bar{p}$ and a corresponding revenue. On the other hand, fixing $p = \bar{p}$ in the policy without threshold will give the same N^* because demand is the same, and therefore the same revenue. Thus pricing without threshold drives to the optimal revenue for this model. A similar analysis can of course be realized in the case of thresholds without hysteresis.

Numerically, this result is easily confirmed: the best threshold policy (with and without hysteresis) in terms of revenue is by pushing the thresholds as far as possible, resulting therefore in the policy without thresholds.

7 Conclusions

This paper investigates three pricing strategies at an M/M/1 queue, and compares the maximal revenues obtained in each case. In each case, we have been able to prove the existence, unicity (and determine) the equilibrium number of sources, and explained how optimal values can be computed. Note that *fixed policy* is a special case of *threshold policy*, itself a particular case if *hysteresis policy*. Nonetheless, under the assumption that a user does not care about the current state of a queue when he sends packets because session lenghts are large with respect to the dynamics of the queue, meaning that he rather see the queue in steady-state, then the particular case of a *fixed policy* is shown to be the one providing the best revenue.

As a direction for future work, we would like to investigate the best policy when sessions may be short and decisions of sending packets depend dynamically on the queue length. This complicates the analysis because it requires to make a transient analysis of the queue.

References

1. Mendelson, H., Whang, S.: Optimal incentive-compatible priority pricing for the M/M/1 queue. Operations Research **38**(5) (1990) 870–883
2. Hayel, Y., Ros, D., Tuffin, B.: Less-than-Best-Effort Services: Pricing and Scheduling. In: IEEE INFOCOM 2004, Hong-Kong, China (2004)

3. Haviv, M.: The Aumann-Shapley price mechanism for allocating congestion costs. Operations research Letters **29** (2001) 211–215
4. Courcoubetis, C., Weber, R.: Pricing Communication Networks—Economics, Technology and Modelling. Wiley (2003)
5. Osborne, M., Rubenstein, A.: A Course on Game Theory. MIT Press (1994)
6. Le Ny, L., Tuffin, B.: A simple analysis of heterogeneous multi-server threshold queues with hysteresis. In: Proceedings of Applied Telecommunication Symposium (ATS), San Diego, USA (2002)