5^e *Conférence Francophone de MOdélisation et SIMulation*
"Modélisation et simulation pour l'analyse et l'optimisation des systèmes industriels et logistiques"
MOSIM'04 – du 1^{er} *au* 3 *septembre* 2004 *- Nantes (France)*

AN ELEMENTARY MODEL FOR CUSTOMER FIDELITY

Max-Olivier Hongler[1], **Naoufel Cheikhrouhou** [2] **and Rémy Glardon**[2]

Swiss Federal Institute of Technology at Lausanne –1015 Lausanne – Switzerland
[1] Laboratoire de Production et Robotique
max.hongler@epfl.ch

[2] Laboratoire de Gestion et Procédés de Production
remy.glardon@epfl.ch, naoufel.cheikhrouhou@epfl.ch

RÉSUMÉ : *Customer fidelisation is an important matter not only for marketing but also for production managers that need to understand and analyse the behavior of their customers in order to develop a production strategy. In this context, an idealized version of customer fidelity problems is studied, using a simple queueing model with state dependent reentrant flow. Two particular systems are studied and their performances discussed: the M/M/1 and the G/G/1 queues.*

MOTS-CLÉS : *customer fidelity, queueing networks, state-dependant queue*

1. INTODUCTION

The fidelity of customers to a particular service is an important issue in running a business in a fluctuating demand environment. Indeed, having a group of known customers with known habits and requirements often enables to reduce the business costs. In a recent issue of the newspaper [Peclet 03], it is mentioned that gaining a new customer costs about five times more than gaining the fidelity of an already existing customer. The fidelity of customers is usually gained by simultaneously offering several service benefits; i.e reduction of the service costs, decrease of the waiting time by introducing priority policies, introduction of customized solutions, etc. Clearly, the decision to stay with a given service provider strongly depends, for each particular customer, on its satisfaction which is based on its past experience with the service provider. In addition to the flow of regular (i.e. loyal) customers, it is clearly important for a service provider to acquire new customers. Accordingly, the service traffic offered by the provider will basically be composed of two types of customers i) the new arriving customers and ii) the existing (i.e. loyal) customers who already have experienced the server provider and are satisfied with it.

The present paper is a simple tentative to mathematically explore the traffic flows resulting from this mixed customer flows. In our idealization, it is assumed that the customer satisfaction is exclusively based on their waiting time before service. More precisely, the longer a customer waits in the queue formed upstream of the service provider, the more likely it is that he will leave the system. The dynamics of the system will therefore be of the type of a queuing system with state dependent feedback. While state dependent queuing systems are abundantly studied in the literature (for recent reviews see [Falin et al. 97] and [Dshalalow 97]), to our knowledge, little attention has so far been devoted to state dependent feedback, a situation directly inspired by the present customer fidelisation problem.

This paper is organized as follows. In section 2, we introduce the general features of our model. In section 3, we focus on the analytical study which can be performed when Poisson processes are used. The section 4 is dedicated to the extension of the model based on G/G/1 queues. The conditions characterizing a stationary regime are explicitly discussed and the concept of fidelity factor is introduced.

2. ELEMENTARY MODEL

Consider a facility S which delivers a single type of service. Depending on the customers, the service time is not a constant but includes some random elements. We shall model the service time t_s by a random variable (r.v.) with cumulative distribution function (CDF) $B(x)$. We adopt the notation:

$$\text{Prob}\{x \leq t_s \leq x + dx\} = dB(x) = b(x)dx$$

where $b(x)$ is the probability density function (PDF). The first moment (i.e. the average service time) will be written as:

$$\frac{1}{\mu} = \int_0^\infty x dB(x) \; = \int_0^\infty x b(x)\, dx, \qquad (1)$$

Hence the parameter μ stands for the service rate.

The customers arriving to the service facility are described by a random point process which we shall assume is a renewal process with independent and identically distributed inter-arrival times t_a governed by the CDF $A(x)$. We define:

$$\text{Prob}\,\{x \leq t_a \leq x + dx\} = dA(x) = a(x)dx$$

and the first moment (i.e. the average inter-arrival time) will be written as:

$$\frac{1}{\lambda} = \int_0^\infty x dA(x) \; = \int_0^\infty x a(x)\, dx, \qquad (2)$$

After leaving the service facility, a customer has to adopt one of the two following alternatives:

a) leave the system forever

b) line up once again upstream of the service facility in order to receive another service.

The service discipline will be, from now on, restricted to the FIFO discipline. The choice between the alternatives a) or b) depends upon the last waiting time τ_q in the queue experienced by the customer. More precisely, If the waiting time $\tau_q > R$ with R being a critical time delay, the customer will choose option a) (i.e. leave the system) and the option b) will be selected when $0 \leq \tau_q \leq R$. The schematic flow diagram of customers is represented in Figure 1. Note that λ_i and λ_r are respectively the flow rate of new customers and the flow rate of remaining customers.

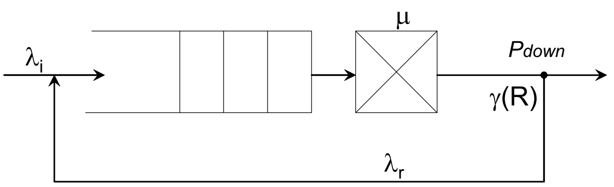

Figure 1: Sketch of the customers flow

The branching ratio between the options a) or b) does reflect the impatience of customers having experienced a long waiting time and then decide to leave the system and try to find another server with lower waiting time. Let us emphasize that the present model differs with the usual models of impatient customer queueing models. Indeed in the present paper, the customers do systematically enter the queue at least once. Accordingly, a new customer never leaves the queue before experimenting a first service. We now write the PDF of the waiting time τ_q as:

$$\text{Prob}\,\{x \leq \tau_q \leq x + dx\} = \omega_q(x)dx \qquad (3)$$

Define now by $\gamma(R) \in [0, 1]$, the proportion of customers who remain in the system.

$$\text{Prob}\,\{0 \leq \tau_q \leq R\} = \gamma(R) =$$

$$\int_0^R \omega_q(x)dx, 0 \leq \gamma(R) \leq 1 \qquad (4)$$

Clearly, the larger R the larger the proportion of customers staying in the system. In the limit case where $R \to 0$, the behavior of the feedback queue will tend to the ordinary (open) queueing system.

3. ANALYSIS FOR THE $M/M/1$ QUEUE DYNAMICS

To explore analytically the model introduced in Section 1, the following simplifying assumptions are made:

$$A(x) = 1 - e^{-\lambda x} \qquad \text{and} \qquad B(x) = 1 - e^{-\mu x} \qquad (5)$$

In view of Eq.(5), the queuing model sketched in Figure 1, the following properties directly holds [Medhi 03].

1) arrivals and services occurs according to Poisson processes.

2) the process resulting from multiplexing and separating Poisson processes is itself a Poisson process.

3) the departure process of a $M/M/1$ queue with an infinite capacity of the waiting room, in the stationary regime, is a Poisson process with the same rate as the input Poisson process.

Using these properties and focusing on the stationary regime of the system sketched in Figure 1, we may write:

$$\lambda_i = \lambda_s, \qquad \text{implied by the stationarity condition.} (6)$$

As the flow up-stream of point P_{down} in Figure 1 is a Poisson process with intensity $(\lambda_i + \lambda_r)$, we use the properties 2) and 3) above to write:

$$\lambda_r = \gamma(R)\,(\lambda_r + \lambda_i) \tag{7}$$

implying that $\gamma(R)\,\lambda_i = (1-\gamma(R))\,\lambda_r$ and hence using Eq.(6):

$$\lambda_i = \lambda_s = [1 - \gamma(R)]\,(\lambda_r + \lambda_i)\,. \tag{8}$$

It is well known that for $M/M/1$ queues with arrival rate $\lambda = \lambda_r + \lambda_i$ and service rate μ, the probability density

$$\text{Prob}\,\{x \leq W \leq x + dx\} = \omega_q(x)\,dx$$

which characterizes the waiting time W as:

$$\omega_q(x) = [\mu - (\lambda_r + \lambda_i)]\,e^{-[\mu-(\lambda_r+\lambda_i)]\,x}. \tag{9}$$

In particular, the mean waiting time $\langle W\rangle$ reads as:

$$\langle W\rangle = \int_0^\infty x\,\omega_q(x)\,dx = \frac{1}{\mu - (\lambda_i + \lambda_r)} \tag{10}$$

In view of Eq.(9), the branching ratio $\gamma(R)$ will be:

$$\begin{aligned} &\text{Prob}\,\{0 \leq W \leq R\} = \\ \gamma(R) &= \int_0^R [\mu - (\lambda_r + \lambda_i)]\,e^{-[\mu-(\lambda_r+\lambda_i)]\,x}dx = \\ &= 1 - e^{-[\mu-(\lambda_r+\lambda_i)]\,R} \end{aligned} \tag{11}$$

Using Eqs.(7) and (11), we directly deduce:

$$\lambda_r = \left[1 - e^{-[\mu-(\lambda_r+\lambda_i)]\,R}\right]\,(\lambda_i + \lambda_r) \tag{12}$$

Hence, given the control parameters μ, λ_i and R, the solution of the transcendent Eq.(12) determines the flow rate λ_r.

Let us now introduce the input traffic rate $\rho_i = \frac{\lambda_i}{\mu} \in [0,1]$ and the total traffic rate $\rho = \frac{\lambda_i+\lambda_r}{\mu} \in [0,1]$. In view of Eq.(12), we immediately have:

$$\rho_i = \rho e^{-\mu R(1-\rho)} =: F(\rho) \tag{13}$$

Defining $F(\rho) = \rho e^{-\mu R(1-\rho)}$, it is immediate to verify that $F(\rho)$ is monotonously increasing with $F(0) = 0$ and $F(1) = 1$. Hence a single traffic value $\rho^* \in [0,1]$ solves the transcendent Eq.(13).

3.1 Fidelity factor

Assuming that the stationary regime is reached, it is easy to estimate the average sojourn time in the queue. Indeed, for the standard, (i.e open) $M/M/1$ queueing system, we have that the mean waiting time $\langle W\rangle$ is given by [Medhi 03]:

$$\langle W\rangle = \frac{1}{\mu - (\lambda_i + \lambda_r)}, \qquad (\mu > (\lambda_i + \lambda_r)) \tag{14}$$

Note at this stage that the probability to reenter k times into the queue and then leave the system will be given by:

$$(1 - \gamma(R))\,[\gamma(R)]^k \tag{15}$$

Accordingly, the mean sojourn inside the system $\mathcal{F}$ (i.e. *the fidelity* factor), can be calculated by using Eq.(15) and reads as:

$$\mathcal{F} = \sum_{k=1}^{\infty} \frac{k\;(1 - \gamma(R))\,[\gamma(R)]^k}{(\mu - (\lambda_i + \lambda_r))} = \frac{\gamma(R)\langle W\rangle}{[1 - \gamma(R)]}, \tag{16}$$

where in Eq.(16) we use $\sum_{k=0}^{\infty} k\,q^k = \frac{q}{(1-q)^2}$.

Observe that Eqs.(11) and (14), enable to rewrite Eq.(16) as:

$$\begin{aligned} \mathcal{F} = \langle W\rangle \left[e^{\mu R(1-\rho)} - 1\right] = \\ = \langle W\rangle \left[e^{\frac{R}{\langle W\rangle}} - 1\right] = \langle W\rangle \frac{\lambda_r}{\lambda_i} \end{aligned} \tag{17}$$

Let us now define:

$$\pi_{ri} = \frac{\lambda_r}{\lambda_i} \tag{18}$$

With the definition given in Eq.(18) and in view of Eq.(17), the number $\pi_{ri} = \frac{\mathcal{F}}{\langle W\rangle}$ directly represents the average number of returns into the queue.

To illustrate one possibility to use our model, let's consider now the following situation:

3.2 Determination of the Fidelity factor

For a given input traffic ρ_i of new customers, we would like to ensure a fixed server global utilization $\rho = \rho_i + \rho_r$. Determine the required fidelity factor $\mathcal{F}$ and its associated critical time delay R.

In the stationary regime, we directly have:

$$\rho = \rho_i + \rho_r = (1+\pi_{ri})\rho_i \implies \pi_{ri} = \frac{\rho - \rho_i}{\rho_i}$$

In view of Eq.(17), we can determine the fidelity factor $\mathcal{F}$ as:

$$\mathcal{F} = \frac{\rho - \rho_i}{\mu \rho_i (1-\rho)}$$

and using the fact that $(\langle W \rangle)^{-1} = \mu(1-\rho)$, Eq.(13) directly yields:

$$\frac{R}{\langle W \rangle} = \ln\left(\frac{\rho}{\rho_i}\right)$$

Consider for example the case $\rho = 0.9$ and $\rho_i = 0.2$. These values imply that $\mu\,\mathcal{F} = 35$ and hence $\mu \langle W \rangle = 10$. Therefore the average number of returns π_{ri} will be:

$$\pi_{ri} = \frac{\mathcal{F}}{\langle W \rangle} = 3.5 \quad \text{and} \quad \frac{R_{\rho_i=0.2}}{\langle W \rangle} = \ln(4.5) \simeq 1.5$$

Observe that when the input of new customers is reduced to $\rho_i = 0.1$, a stationary regime implies :

$$\pi_{ri} = \frac{\mathcal{F}}{\langle W \rangle} = 8 \quad \text{and} \quad \frac{R_{\rho_i=0.1}}{\langle W \rangle} = \ln(9) \simeq 2.19$$

As the mean waiting time (measured in units of the mean service time) $\mu\langle W \rangle = 10$ remains unchanged. The customers in that case have to be much more patient, to ensure the same server utilization $\rho = 0.9$; (i.e. $\frac{R_{\rho_i=0.1}}{R_{\rho_i=0.2}} = 2.19/1.5 \simeq 1,46$)

4. GENERAL SERVICE TIME-$G/G/1$ QUEUES

The general case which arises when the arrival and service times are drawn from an arbitrary CDF $A(x)$ and $B(x)$ is obviously more complex to discuss analytically. Indeed, for an arbitrary traffic load, the system with feedback will not belong to the usual class of queueing models. This is due to the following features:

i) the statistical properties of the output process of a simple G/G/1 queues are not known in general. One however knows that the output process is generally not a renewal process [Daley 76]. In addition, due to the presence of the feedback, it becomes clear that we generally cannot expect the modified arrival process loading the server to be a renewal process. Hence one of the basic assumptions needed to define a "classical" queuing model is not fulfilled.

ii) due to the presence of the feedback, it is likely that there will be complex correlations between the arrival and the output processes of the model. How to deal with such correlations effects is clearly beyond the scope covered by ordinary $G/G/1$ queueing models.

In view of the previous remarks, to further discuss the behavior of the present feedback queueing model, one will necessarily rely on approximation methods. First, we note that for the applications we have in mind, the service facility will always be very busy (high traffic regime). Therefore, despite the fact that the original feedback system does not rigorously behave as a $G/G/1$ queue, we expect that in heavy traffic regimes, the output process and hence the feedback flow of customers can be, as a first approximation, assimilated to diffusive processes. Let us formally write C_r^2 to be the coefficient of variations (CV) of the process which describes the feedback customers. We write $\tilde{C}_a^2$ for the CV of the effective arrival process viewed by the server. Neglecting, in the high traffic regime, the correlations between the output and the effective input process viewed by the server, we can approximately write:

$$\tilde{C}_a^2 = C_a^2 + C_r^2 \tag{19}$$

with C_a^2 being the CV associated with the external arrival process, (i.e. characterized by $A(x)$).

Let us now recall that for a $G/G/1$ queue in the heavy traffic regime $\rho = \frac{\lambda}{\mu} \lesssim 1$, the diffusion approximation method holds and based on this approximation, one can approximately derive [Medhi 03]:

$$\text{Prob}\,\{x \leq W \leq x + dx\} = \omega_q(x)\,dx = \eta e^{-\eta x}\,dx \tag{20}$$

with:

$$\eta = \frac{2(1-\rho)}{\lambda(\tilde{C}_a^2 + C_b^2)} \tag{21}$$

with C_b^2 being the CV of the server process characterized by $B(x)$.

Observe that with the mapping $\eta \mapsto \mu(1-\rho)$, Eq.(20) reduces to the density Eq.(9) derived for the $M/M/1$

queue. Hence for the high traffic regimes of the server, one can directly use our previous results provided one identifies:

$$\eta = \mu(1-\rho) \qquad \text{for } \rho \lesssim 1$$

It now remains to calculate the CV C_r^2. A refined estimate of C_r^2 occurring in Eq.(19) promised to be a difficult task. Indeed, the feedback process is itself the result of a Bernoulli sampling (characterized by the parameter $\gamma(R)$) of the output process which is not a renewal process. At this preliminary stage of the work, one simply can study the sensitivity of the final results on the parameter C_r^2. An indicative value for C_r^2 will be to take $C_r^2 = 1$ which corresponds to the $M/M/1$ dynamics for which exact results have been obtained in the previous section. In view of Eq.(21), an increase of the value of C_r^2 will decrease the factor η and therefore increase the mean waiting time $\langle W \rangle = \eta^{-1}$. This does ultimately reduce the fidelity factor. This clearly shows that it is important to try to reduce the variability of the flow of feedback customers.

5. CONCLUSION

In this work, an analytical model for customer fidelity is presented, based on queuing networks. When the system behaves as an M/M/1 queue, the performance of the system with feedback is an M/M/1 queue dependant on the limit of the customer patience. This work is a first approach to develop a production strategy based on customer satisfaction and evaluation of the patience threshold. On the other hand, the G/G/1 queue model is more complex and needs further improvement as the output is not a renewal process. All the results have been checked by comparison with simulation runs of the same models.

REFERENCES

[Daley 76] D. J. Daley "Queuing Output Process". Adv. Apppl, Prob. **8**, (1976),395-415.

[Dshalalow 97] J. H. Dshalalow "Queuing systems with state dependent parameters " in "*Frontier in Queueing*", Ed. J. H. Dshalalow, CRC Press, (1997), 61-117.

[Falin et al 97] G. I. Falin and J. G. C. Templeton "*Retrial Queues*", Chapman and Hall, (1997),

[Medhi 03] J. Medhi "*Stochastic Models in Queueing Theory*". Academic Press (second edition), (2003).

[Peclet 03] J.-Cl. Peclet. "Client roi ou client pigeon: comment faire la diffirence ?", Journal "*Le Temps*", first page of the 22^{nd} December 2003 edition.