Queuing Processes

Abstract.

Queuing Processes is a subject that have been overlooked by research in economic applications. The objective of this paper is to give some economic insights of the queuing applications that have been tried over the years, as well as to propose some new ways of tackling the problem.

Keywords: Review, Poisson Process, Markov Process, Design and Control of queues.

1. Introduction

This paper is meant to be review of the applications of the queuing theory in economic literature, as well as (hopefully) give a new start in this domain. The first part of this paper is meant to give a short history of development of queuing theory and to familiarize the reader with the specific jargon of this language.

The second part is devoted to economic applications of queuing theory. The classic customer service model is presented and some extensions of this model.

Part three is trying to present some queuing applications from a new perspective. There are two main development areas in recent queuing literature; unfortunately none of them have been fully implemented in economic applications. Consider a resource allocation problem, for example, using linear programming to determine how much of each product *should* be made, within the limitations of the resources needed to make the products. This is a prescriptive model since it prescribes the optimal course of action to

follow. This type of approach to the problem has the generic name of Design and Control of queues.

The other approach to the problem is to observe a real live situation, and to try to figure out the parameters involved. This sounds like an old song because all the queuing theory is doing just that: put a real live situation into a known mathematical pattern and apply formulas. *Simulation* isn't just as simple though, because most of the real live problems don't have a definite mathematical model just for that case, and even if there is one it might not have closed form solutions. Also, the next step for simulation is to try to improve the model by trying diverse "laboratory situations" and see which one improves the model the best.

Part I. Basics of queuing theory.

1.1 History

The theory of queues was initiated by the Danish mathematician A. K. Erlang, who in 1909 published "The theory of Probabilities and Telephone Conversation". He observed that a telephone system was generally characterized by either (1) Poisson input (the number of calls), exponential holding (service) time, and multiple channels (servers), or (2) Poisson input, constant holding time and a single channel. Erlang was also responsible in his later works for the notion of stationary equilibrium and for the first consideration of the optimization of a queuing system.

Applications of the theory to the telephony were soon appearing. In 1927, E. C. Molina published "Application of the Theory of Probability to Telephone Trunking

Problems", and one year later Thornton Fry printed "Probability and its Engineering Uses" which expand much of Erlang's earlier work. In the early 1930's Felix Pollaczeck did some further pioneering work on Poisson input, arbitrary output, and single and multiple channel problems. Other names working in the same field during that period included Kolmogorov and Khintchine in Russia, Crommelin in France and Palm in Sweden. The work in queuing theory picked up momentum rather slowly in its early days, but in 1950 started to accelerate and there have been a great deal of work in the area since then.

1.2 Early applications

There are many valuable applications of the theory, most of which have been well documented in the literature of probability, operations research, management science and industrial engineering. Examples include: traffic flow (vehicles, aircraft, people, communications), scheduling (patients at the doctor, programs on a computer), and facility design (banks, post offices, fast-food restaurants).

Queuing theory originated as a very practical subject but much of literature up to middle 1980's was of little direct practical value. Since then the emphasis in literature on finding the exact solution of queuing problems with clever mathematical tricks is now becoming secondary to model building and the direct use of these techniques in decision making. Most real problems do not correspond exactly to a mathematical model and do not always have closed-form solutions, but most of the time we are able to conduct computational analysis and find approximate solutions. We have to thank for this to our every day companion, the computer.

1.3 Characteristics

The mechanism of the queuing process is very simple. Customers (not necessarily human customers) are arriving for service, waiting for service if it is not immediate, and leaving the system as soon as they are served.

There are six basic characteristics of queuing processes which provide an adequate description of a queuing system: (1) arrival pattern of customers, (2) service pattern of servers, (3) number of service channels, (4) system capacity and (5) queue discipline.

In usual queuing systems the arrival pattern of customers is stochastic and it is thus necessary to know the probability distribution describing the time between successive customer arrivals (interarrival times). Also the arrival pattern can change with time so we differentiate between stationary and nonstationary arrival patterns. The same discussion applies to the service pattern of servers, a probability distribution is needed to describe the sequence of customer service times. Queue discipline refers to the manner in which customers are selected for service when a queue has formed. The most common discipline is first come, first served (FCFS), but there are many others like last come, first served (LCFS) which is applicable in many inventory systems as it is easier to reach the nearest item; randomly selecting for service (RSS) independent of the arrival time of the customer; and a variety of priority schemes, the customers with higher priority being served ahead of the lower priority customers regardless of the order in which they arrived to the system.

1.4 Notation

As shorthand for describing queuing processes, a notation has evolved, due to Kendall (1953), which is now standard throughout the queuing literature. A queuing process is described as *A/B/X/Y/Z*, where *A* indicates the interarrival-time distribution, *B* the service pattern described by the probability distribution for service time, *X* the number of servers, *Y* the restriction on system capacity and *Z* the queue discipline. Table 1 bellow summarizes some of the most common symbols.

Characteristic	Symbol	Explanation
Interarrival time Distribution *(A)* and Service-time Distribution *(B)*	M	Exponential
	D	Deterministic
	E_k	Erlang type *k (k=1,2,…)*
	H_k	Mixture of *k* exponentials
	G	General distribution
Number of servers *(X)*	*1,2,…,∞*	
Restriction on system capacity	*1,2,…,∞*	
Queue discipline	FCFS	First come first served
	LCFS	Last come, first served
	RSS	Random selection for service
	PR	Priority
	GD	General discipline

In many situations only the first three symbols are used. Current practice is to omit the service capacity symbol if no restriction is imposed *(Y=∞)*, and to omit the queue discipline if it is first come first served *(Z=FCFS)*. The symbol G represents a general probability distribution; all we know is that the interarrival times are independent and identically distributed.

Part II. Review of literature. Economic Examples.

Surprisingly there are not so many economic applications studied yet. Some of the latest applications include:

2.1 Customer service

The classical model of clients waiting for service while incurring waiting cost to get reward from service have been implemented extensively but for the practical applications it has not surpass the model M/M/1. There have been attempts to construct more complex models. For example Kleinrock (1968) propose a model M/M/1 with a little different service discipline. Instead of FCFS discipline customers are using bribe to gain a better position in line. A customer paying a bribe will be served before the ones who paid smaller bribes in the queue, but after the people in the queue who paid larger bribes. This model is appropriate to some undeveloped countries where for gaining an audience you are encouraged to pay a fee to the clerk doing registration. The model is also appropriate to an auction process. Leff (1970) suggested that such a model may have beneficial effects serving as a catalizator for an otherwise sluggish economy. Myrdal (1968, chap. 20), argues that the corrupt official may deliberately cause delays in order to attract more bribes. If this is the case then the efficiency argument is gone. This opinion is answered by Lui, who in 1985 showed that, if the server does not decide the amounts of bribe payments, but the customers themselves are doing that, there exists Nash equilibrium. Under the extra condition, that the server is interested in speeding up the process in order to gain more bribes, the outcome is also socially optimal.

Luski (1976) discusses the notion of equilibrium in a queuing system with two servers (M/M/2 model). Specifically there are two firms offering the same kind of service, each fixing its price as to maximize profit. The decision of customers in joining one of the queues or not joining at all is made according to their cost of waiting, the expected waiting time and the price of the service in each firm. He showed that if the firms are not cooperating, and if the service time for each firm $\mu < \lambda/2$, each firm will sell at a different price in order that an equilibrium exists. In the alternate case $\mu > \lambda/2$ the firms will sell at the same price (the same mechanism that governs the real life). Not only that but he showed that if the firms are cooperating it is possible to raise joint profits by playing with prices (again logical: more or less like a monopolistic model).

2.2 Derivatives of the Customer Service Model: Bottleneck model.

This model is an extension of the previous M/M/1 model. Customers are arriving at a facility and are faced with the decision of entering or not in the system (balking phenomenon). Each customer has a cost per unit of service and waiting time, c, and receives a benefit R if he is served by the facility. In addition to the previous models the system charges a fee θ for the service and this toll determines a critical queue size. This model is pertinent with some real life situations, like today's Internet providers or toll paying highways. P. Naor (1969) has shown that the critical length of the queue, which maximizes the social welfare function is greater then the one which maximizes the expected revenue per unit time. In other words, the revenue-maximizing toll exceeds the socially optimal toll. Edelson and Hildebrand (1975) have shown that this is not the case if no balking is permitted and the same toll is Nash and Pareto Optimal.

R. Arnott, A. de Palma and R. Lindsey (1995-1996), basing their work on a model initially developed by Vickrey (1963,1969) and similar with the previous one, generalized this problem. They allowed multiple servers (service time is deterministic), and also the customer, instead of just deciding whether or not to enter the system, now has the extra choice of when to use the facility. The decision is made accordingly to information that each client has, information that is stochastic with some distribution. The authors define better information from the social efficiency of the model point of view. The model is very applicable. Roads with today's severe congestion problems and with the information that each driver can obtain (listening radio, information panels etc…) are the best example of such application. Arnott et all have shown that, if the customers decide only whether or not to use the facility, better information is more efficient. They have shown also that if the customers decide both whether and when to use the facility then the above result doesn't hold anymore and better information can be harmful when the congestion is unpriced. However if efficient tolling (different toll prices during the congestion hours) is applied the above result holds.

Another extension of the M/M/1 model is an article by Joseph Daniel (1995). He applies this queue to an airport, modeling arrivals and service time until departures as stochastic. Also he is applying his thus constructed model to a real life case using data from the Minneapolis-St. Paul Airport during the firs week of May 1990. He allows the arrival rate to be modeled by a time dependent Poisson Process, in fact dividing the time into equal intervals during which the arrival rate is constant. The costs incurred are the usual ones: queuing cost and service cost, where by service cost he understands the price

paid for the difference from the scheduled time of departure. The social planner can impose a congestion fee, which will depend of the scheduled time of operation.

What is remarkable about this article is that it estimates the parameters of the model and then uses those parameters to simulate the data. Using simulation allows him to find the optimal congestion toll and the optimal number of servers (i.e. by building additional runways).

Part III. Examples in mathematical form

3.1 Insurance risk

We make the following assumptions regarding the business of an insurance company:

(i) The number of claims arising in a time interval (0,t] has a Poisson Distribution with parameter λ, $0 \leq \lambda \leq \infty$.

(ii) The amounts of successive claims are independent random variables with common distribution B(x) ($-\infty < x < \infty$), where x can take negative values in the case of ordinary whole-life annuities.

We can consider this model as a M/G/1 queuing system where the input process A(t) is just a simple Poisson process with parameter λ giving us the moments of claims. We will consider the service as being the amount of each claim, and let X(t) the total amount of claims arising in (0,t]. Assuming X(t)=0 we have:

(3.1) $X(t) = \zeta_1 + \zeta_2 + \ldots + \zeta_{A(t)}$,

where ζ_1 , ζ_2 ,… are the successive claims. Because A(t) is a Poisson process, the distribution function of X(t) is given by:

(3.2) $\chi(x,t) = P\{ X(t) \leq x \} = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} B_n(x)$

This is the compound Poisson distribution. We have:

(3.3) $EX(t) = \lambda t \int_{-\infty}^{\infty} x dB(x)$, and $VarX(t) = \lambda t \int_{-\infty}^{\infty} x^2 dB(x)$

(iii) The totality of policyholders pays the company premiums at a constant rate β. We call β the gross risk premium rate. We will assume that α and β have the same sign.

The company's reserve fund at time t is given by $Z(t) = x + \beta t - X(t)$, with initial value $Z(0) = x \geq 0$. Here $Z(t)$ may assume positive or negative values.

The objective of the company is to choose the initial reserve x large enough to avoid ruin over a finite or infinite horizon.

So, the company is concerned with the stopping time:

(3.4) $T(x) = inf\{ t > 0 : x + \beta t - X(t) < 0 \}$,

and wants to evaluate the probability of avoiding ruin over a finite or an infinite horizon, that is,

(3.5) $P\{T(x) > t\}$ *(for some* $0 < t < \infty$*)* or $P\{ T(x) = \infty \}$.

Prabhu (1998) showed that the storage process $Z(t)$ is a Markov process and he computed the limit probabilities (3.5) when t goes to ∞.

In the model described above we have seen the basic process, which is the compounded Poisson Process. An important feature of this model is that at any moment of time only a finite number of events have occurred (namely customer claims). The

resulting property of the resulting process is that its sample function takes only a finite number of jumps in each finite interval. However this description of the input is very unrealistic in certain situations. Which brings us to the next model:

3.2 Storage Model

Consider a model of a granary with a large enough (effectively infinite) capacity. Let $X(t)$ denote the input of grain into it during a time interval $(0,t]$.

In order to formulate such a process let's suppose that $X(t)$ has stationary and independent increments i.e.:

(i) For $0 \leq t_1 \leq \ldots \leq t_n$ with $n \geq 2$ we have $X(t_1), X(t_2)- X(t_1),\ldots, X(t_n)-X(t_{n-1})$ are independent

(ii) Distribution of increment $X(t_p)-X(t_{p-1})$ depends only on t_p-t_{p-1}

Suppose furthermore that the process X is right continuous, the quantity $X(t)-X(t-)$ being the jump at time t with only a finite number of jumps allowed.

Under these regularity conditions X the input process is a Levy process (a broad family containing Brownian Motion and Compounded Poisson Process among others). In particular $Z(t)=Z(0)+X(t)+\int_0^t 1_{\{Z(s)=0\}}\,ds$ is a semimartingale so all the theory of Stochastic integration can be applied (see Protter 1988).

Again we are concerned with $T(x)= inf\{t : Z(t)=0\}$the period until the first shortage appears (here $x=Z(0)$)and again Prabhu (1998) found the distribution of T under the assumption $T(0)=0$.

3.3 Livestock problem:

Suppose that we have an animal farm that grows pigs. We want to put this into a queuing model. To do this we will make the following assumptions:

Each time a new animal is born we have to pay a fixed amount of money β for doctor and other costs. In addition each animal costs us money until we service it. C denotes this cost per unit time. We will consider this model as a M/M/1/K queue where the last number K is the finite capacity of the model.

As usual arrival rate is λ service rate is μ. We want to find the optimal truncation value k, which maximizes the expected profit rate. Mathematically this problem is to find:

$$\max_{1\le k\le K}\left(\beta\lambda(1-p_k)-c\lambda(1-p_k)W\right)$$

, where the quantities p_k and W are given by:

$$p_k=\left(\frac{\lambda}{\mu}\right)^k \text{ for}: 0\le k\le K \text{ and } W=$$

3.3.1 Another option for livestock model (M/M/1 queue with controlled service rate):

Suppose now that we have an unlimited storage capacity but we can control the rate of service μ in an interval $[0, \ddot{\imath}]$. We can change the rate of service after servicing each pig. Let $S(\mu)$ be the cost per unit time for using rate μ and let $C(i)$ be the cost per unit time when there are i pigs in the system (queue plus service). We want to cut down C by choosing a faster service rate, which costs more; in other words we want to find optimal tradeoff between service cost and system cost.

Assumptions:

1. There is a μ such that $\mu > \lambda$
2. *C* is positive, nondecreasing and "convex": $C(i+2) - C(i+1) \geq C(i+1) - C(i)$
3. *S* is nonnegative and continuous on $[0, \ddot{\mu}]$; $S(0)=0$

Here the state variable is the number of customers in the system. The control variable is the choice of service rate after a customer departs.

Take $\nu = \lambda + \ddot{\mu}$ the uniform transition rate. Bertsekas(1987) shows that the problem can be reduced to this transition rate. The cost per stage is: $C(i) + S(\mu)$. Then the Bellman equation is:

$$J(0) = \min_{\mu} [C(0) + (\nu - \lambda)J(0) + \lambda J(1)]$$

$$J(i) = \min_{\mu} [C(i) + S(\mu) + \mu J(i-1) + (\nu - \lambda - \mu)J(i) + \lambda J(i+1)]$$

So we get the optimal policy: use at state i the service rate that minimizes:

$$S(\mu) - \mu [J(i) - J(i-1)]$$

3.3.2 M/M/1 model with controlled arrival rate

We will look to the same problem from yet another perspective. Suppose that the service rate is fixed λ but we can control now the birth rate into an interval $[0, \ddot{\lambda}]$. The only thing that modifies from the previous model is that we have now a cost $S(\lambda)$ per unit time associated with the specific rate of birth chosen instead of the cost specific to different service rates.

Again let $\nu = \ddot{\lambda} + \mu$, the uniform transition rate. In this case Bellman equation becomes

$$J(0) = \min_{\lambda} [C(0) + S(\lambda) + (\nu - \lambda)J(0) + \lambda J(1)]$$

$$J(i) = \min_{\lambda} [C(i) + S(\mu) + \mu J(i-1) + (\nu - \lambda - \mu)J(i) + \lambda J(i+1)]$$

Maximizing this time with respect to λ gives us the optimal policy: use at state i the birth rate that minimizes $S(\lambda) + \lambda [J(i) – J(i-1)]$.

Bibliography:

- Arnott R., de Palma A., Lindsey R., “Economics of a Bottleneck” Journal of Urban Economics 27, (January 1990), 111-130
- Arnott R., de Palma A., Lindsey R., “Does Providing Information to Drivers Reduce Traffic Congestion?”, Transportation Research 25, (no.5 1991), 309-318
- Arnott R., de Palma A., Lindsey R., “Information and Usage of Free-Access Congestible Facilities with Stochastic Capacity and Demand”, International Economic Review 37, (January 1996), 181-203
- Arnott R., de Palma A., Lindsey R., “Information and Time-of-Usage Decisions in the Bottleneck Model with Stochastic Capacity and Demand”, European Economic Review 43, (March 1999), 525-548
- Bertzekas, Dimitri P. *Dynamic programming and stochastic control*. New York: Academic Press, 1976
- Daniel Joseph, “Congestion Pricing and Capacity of Large Hub Airports: A Bottleneck Model with Stochastic Queues”, Econometrica 63, (March 1995), 327-370
- Edelson N., Hildebrand D. “Congestion Tools for Poisson Queuing Processes”, Econometrica 43, (January 1975) 81-92

- Gross, D., Harris, C. *Fundamentals of Queueing Theory*. New York: Wiley, 1998
- Kleinrock, Leonard. “Optimum Bribing for Queue Position”. Operations Res. 15 (March/April 1967): 304-318
- Leff, Nathaniel H. “Economic Development through Bureaucratic Corruption”. In *Political Corruption: Readings in Comparative Analysis*, edited by Arnold Heidenheimer. New York: Holt, Rinehart, & Winston, 1970.
- Lui, Francis T. “An Equilibrium Queuing Model of Bribery”. Journal of Political Economy 93 (August 1985): 760-781
- Luski, Israel “On Partial Equilibrium in a Queuing System with Two Servers” Review of Economic Studies 43 (October 1976) 519-525
- Myrdal, Gunnar. *Asian Drama: An inquiry into the Poverty of Nations*. New York: Pantheon, 1969.
- Naor, P. “The regulation of Queue Size by Levying Tolls”. Econometrica 37 (January 1969): 15-24
- Prabhu N.U. *Foundations of Queuing Theory*. Boston: Kluwer Academic Publishers, 1997.
- Prabhu N.U. *Stochastic storage processes: queues, insurance risk, dams, and data communication*. New York: Springer, 1998