CONGESTION IN PROCESSING SYSTEMS

Doina Corina ŞERBAN[1]

Lucrarea se referă de procesul de adoptare a deciziilor în cadrul companiilor care se confruntă cu sisteme de procesare intense în structura tehnologică proprie şi care trebuie să decidă cu privire la numărul şi capacitatea disponibilităţilor de servire şi la ordonanţarea lucrărilor în sistem. Sunt prezentate modele ce descriu câteva tipuri de sisteme simple: sisteme cu unul sau mai multe şiruri de aşteptare şi sisteme cu pierderi. Acestea pot fi folosite pentru a realiza estimări ale sistemelor aglomerate din realitatea obiectivă. Dar mai important este faptul că aceste modele pot fi aplicate şi asigură o mai bună înţelegere a unor sisteme de procesare mai puţin clare. Ideea de bază este că orice sistem de procesare poate fi tratat ca un sistem de aşteptare, cu pierderi, mai mult sau mai puţin complex.

This paper is concerned with the decision-making process of the business company that has charge of the operation of the processing systems and makes decisions relative to the number and capacity of service facilities and the scheduling of jobs in the system. There are presented models for representing some simple types of systems: single- and multiple-channel queuing systems and loss systems. These may be used to make estimates of congestion in real-world systems. But more important is that these models provide insights that can be applied to a great variety of processing systems and that are not at all obvious. The basic ideas apply to all processing systems, whether they be simple queuing systems, loss systems, or more complicated processing systems.

Key words: decision making systems, queuing systems, loss systems, congested processing systems, networks of queues.

1. Introduction

Queues or waiting lines are very common in everyday life. The system becomes congested and we wait. Most of us consider congestion with more or less good humor. Occasionally, the size of a line we encounter discourages us, we abondon the activity, and a sale is lost.

Systems such as waiting in a line for a bus or a taxi, a hospital services or a haircut, a grocery checkout, are commonly called *queuing systems*. But a better term is *procesing systems* [1].

[1] Associate professor, Department of Management, University, "Politehnica" University of Bucharest, ROMANIA

This broader term includes factories in which jobs move through several steps in the process of being manufactured, or offices in which paperwork is handled by several individuals or committees.

These are *networks of queues.* It also includes *loss systems* in which there is no queue at all.

Queuing theory is primarily concerned with processes that have variability in a arrivals of jobs into the system. The time taken to service these jobs is also generally variable. The result is congestion or waiting lines. This can be measured by the average waiting time of arrivals. There are costs associated with having jobs wait. There are also costs associated with adding more service capacity. The management challenge is to balance these costs.

Queuing theory may be used to aid in decisions about:

- The order in which customers should be processed;
- The scheduling of jobs through a manufacturing facility;
- Increasing the service speed of certain operations;
- The value of reengineering whole processes,

or to determine the optimum number of:

- Toll booths for a bridge or toll road;
- Doctors available for clinic calls;
- Repair persons servicing machines;
- Landing strips for aircraft;
- Docks for ships;
- Paramedic units available for emergency calls;
- Clerks for a spare-parts counter;
- Sevice windows for a post office

2. System structure

Fortunately for the busy manager, reasonable decisions about processing systems can frequently be based on past experience or on the facts of the current situation.

Although a number of problems encountered by an executive can be reasonably solved by the use of intuition or past experience, there will be many situations that are too complex for our intuition. In these situations, the problem can be approached by a mathematical model procedure.

Processing systems may be distinguished from one another in many different way. The most important structural distinctions and some of the common measures of system performance are [2]:

- ***Loss Systems versus Queuing Systems.*** Loss systems are those in which arrivals into the system simply exit the system when they encounter delay. In queuing systems, arrivals wait in line for service. Of cours, there can be systems that mix these two elements - potential customers may leave the system if the waiting line is too long.
- ***Arrivals.*** Jobs come into the system for service. They may come single or in batches; they may come evenly spaced in time or in a random pattern; they may come from an infinite or very large population or from a finit set.

- ***Services.*** Each job must be serviced. The service time required may be the same for each job, or service time may vary considerably in a random fashion.

- ***Single Station or Network of Stations.*** The service may be completed by a single unit. In some systems, completing a job requires having an order processed by several work units. Systems that involve more than one station are called *queuing networks.* At each station, some service time is required, and these times generally vary by station.

- ***Number of Servers.*** At each station, there may be only one server or *channel* or many.

- ***Queue Discipline.*** While jobs wait for service, they are in a waiting line or queue. There may be only one line or separate lines for each server. There may be a space limit on the waiting line, and jobs that arrive when the line is full may be turned away.

- ***Measures of Performances.*** There are various ways of judging how well a processing system is performing.Results may be evaluated over a short period of time once the system opens, or or they may be based on the long-run or equilibrium results. Generaly, the time jobs spend waiting is important, and we may lookat the average waiting time or at a measure such as the percent of jobs that wait longer than, say, 10 minutes. A related measure is the *throughput time* for a job (waiting time plus service time). The length of the waiting line is another common measure of performance. These are measures of how well the system is performing from the customer point of view.

  Other measures relate to the cost of operating the system. The system load factor or capacity utilization measures the ability of the system to handle the arrival load. Management has the option of adding more capacity.

  A given processing system can have any combination of the elements described above. Hence, there are a very large number of possible systems, and no one mathematical model can describe them all. In this paper, we focus on a few simple models that have wide applicability and that give us insight into queuing system behavior in general. Another paper will describe simulation, which can be used to model processing systems in a very general way.

3. A single-server queing model

For this model (the Pollaczek-Khintchine model), we shall consider the case in which:

- Arrivals are random, and the interarrival times come froman exponential (or Markov) probability distribution.

- Service times are also assumed to be random following a general distribution with mean μ_S and standard deviation σ_S. Service times are assumed to be independent of the arrival process.

- There is a single server or chammel.

- The queue discipline is ***FIFO,*** and there is no limit on the size of the line.

- The average interarrival and service times do not change over time. The process has been operating long enough to remove effects of the initial conditions. We are interested in the long-run or equilibrium conditions.

Using the notation above; this is the ***M/G/1*** queue. Note first that the system load factor (or capacity utilization) for this queuing model is:

$$\rho = \mu_S / \mu_A. \quad (1)$$

Thus, the capacity utilization of the system (the percent of time the facility is busy serving arrivals) is the average service time divided by the average interarrival time. This number must be less than 1 – if the service time is longer than the interarrival time, the queue will continue to grow without limit.

It is possible to solve this model mathematically for the long-run equilibrium distribution of waiting time and number waiting in line.

For this ***M/G/1*** model, our basic measure, the waiting time multiple (***WTM***), is given by:

$$WTM = [\rho/(1-\rho)][(1+cv_S^2)/2]. \quad (2)$$

From this, the mean of the distribution of waiting times can be calculated as:

$$\mu_W = \mu_S WTM. \quad (3)$$

And the Average length of the queue is:

$$\mu_L = \mu_W / \mu_A. \quad (4)$$

Consider the case in which the distribution of service times is also exponential (that is, Markov). This is the ***M/M/1*** queue. One feature of the exponential distribution is that the standard deviation equals the mean (i.e., $\sigma_S = \mu_S$). If we make this substitution back in the basic formula for the WTM, we obtain the following: *For **M/M/1** queue:*

$$WTM = \rho / (1-\rho). \quad (5)$$

A second special case is that in which the service is deterministic, the ***M/D/1*** case. That is, service time is a constant with zero standard deviation. Substituting $\sigma_S = 0$ in the basic formula, we obtain: *For **M/D/1** queue:*

$$WTM = \rho / 2(1-\rho). \quad (6)$$

The equations for the average waiting time and average queue length are the same as in the basic case.

Suppose we make no assumptions about the distributions for interarrival time and for service time - they can be general probability distributions with specified means and standard deviations. We continue to assume that the distributions are independent and the single queue has a ***FIFO*** discipline.

There is no mathematical model that can solve this case exactly. However, there is an approximation model for this situation called the *heavy-traffic approximation.* It is an accurate estimation for waiting time when the system is heavily loaded (values of ρ close to 100 percent) and a reasonable approximaton in other cases. The waiting time multiple in this case is:

$$WTM = [\rho / (1-\rho)][(cv_A^2+cv_S^2) / 2]. \quad (7)$$

In this formula, $\boldsymbol{cv_A^2}$ and $\boldsymbol{cv_S^2}$ are the coefficients of variation and measure the relative variability of interarrival and service times. The formulas for the average waiting time and average queue length are the same as in the basic case.

4. Multiple servers

In many queuing situations, there is more than one server (or channel) waiting on customers. A check-in counter at the airport with several clerks or multiple check-out stands at the supermarket are examples. In such situations, the waiting line discipline is important- there may be only one line for all channels (common at airport checkpoints) or each channel may have its own line (the usual supermarket case). For the simple model described below, we shall assume that there is a single waiting line. Customers arrive into the system with mean interarrival time μ_A. There are c servers, and each has the same average service time μ_S. The single line is serviced in a FIFO (first-in, first-out) manner, and arrivals and services are independent. The distributions for both interarrival times and service times are assumed to be Markov (exponential).

The formula for the waiting time multiple ***(WTM)*** in this case is, unfortunately, very messy and complicated. Instead of providing it, Table 1 gives the ***WTM*** for a number of useful cases: number of channels from 1 through 5 and 10, and for various levels of the system load factor. Note that the system load factor or utilization in a system with ***c*** channels is:

$$\rho_c = \mu_S / c\mu_A. \quad (8)$$

Table 1.

Waiting Time Multiple for the M/M/c Queue

System Load Factor	*Number of Service Channels, c*					
	c = 1	*c = 2*	*c = 3*	*c = 4*	*c = 5*	*c = 10*
0,10	*0,1111*	*0,0101*	*0,0014*	*0,0002*	*0,0000**	*0,0000**
0,20	*0,2500*	*0,0417*	*0,0103*	*0,0030*	*0,0010*	*0,0000**
0,30	*0,4286*	*0,0989*	*0,0333*	*0,0132*	*0,0058*	*0,0002*
0,40	*0,6667*	*0,1905*	*0,0784*	*0,0378*	*0,0199*	*0,0015*
0,50	*1,0000*	*0,3333*	*0,1579*	*0,0870*	*0,0521*	*0,0072*
0,60	*1,5000*	*0,5625*	*0,2956*	*0,1794*	*0,1181*	*0,0253*
0,70	*2,3333*	*0,9608*	*0,5470*	*0,3572*	*0,2519*	*0,0739*
0,80	*4,0000*	*1,7778*	*1,0787*	*0,7455*	*0,5541*	*0,2046*
0,90	*9,0000*	*4,2632*	*2,7235*	*1,9694*	*1,5250*	*0,6687*
0,95	*19,0000*	*9,2564*	*6,0467*	*4,4571*	*3,5112*	*1,6512*

* Less than 0,00005

As in the case for a single channel, there is no mathemetically exact formula for the case in which the arrival and service times can have a general distribution. However, there is an estimation formula *(the heavy-traffic approximation)* that provides accurate approximations when the system is heavily loaded (load factor close to 100 percent), and rough estimates in other cases. This formula provides an adjustment to the waiting time multiple for the ***M/M/c*** queue given in Table 1. If we let $\boldsymbol{WTM_{M/M}}$ represent the values in Table 1 for the appropriate system load factor and number of channels, then the waiting time multiple for ***G/G/c*** queue is given by:

$$WTM = WTM_{M/M}\,[(cv_A^2 + cv_S^2) / 2]. \tag{9}$$

The second term is an adjustment that allows for different variability in the arrival and service times. The formulas for average waiting time and queue length are the same as in the single-channel case:

Average waiting time:

$$\mu_W = \mu_S\, WTM, \tag{10}$$

Average length of waiting line:

$$\mu_L = \mu_W / \mu_A. \tag{11}$$

In some processing systems, arriving jobs cannot emter the system because there is no waiting line or queue. There are called ***loss systems***.

The ***M/G/c*** Loss system. Consider a system in which jobs arrive in a random or Markov process (that is, interarrival times are exponential) and service times can have a general distribution. There are ***c*** service channels. Jobs that arrive when all ***c*** channels are busy are lost.

The mean interarrival time is μ_A and the mean service time is μ_S. The system load factor is:

$$\rho = \mu_S / c\mu_A. \tag{12}$$

In loss systems, this may be greater than 100 percent since some arrivals actually do not enter the system for service. An important measure of performance for loss systems is the fraction of arriving jobs that are lost. The formula for this is:

$$Fraction\ lost = [(c\rho)^c / c!] / \Sigma_{k=0}^{c}[(c\rho)^k / k!]. \tag{13}$$

For the single-channel case (***c*** = 1), this becomes:

$$Fraction\ lost = \rho / (1+\rho). \tag{14}$$

For the two-channel case (***c*** = 2), it simplifies to:

$$Fraction\ lost = 2\rho^2 / (1+2\rho+2\rho^2). \tag{15}$$

For the three-channel case (***c*** = 3):

$$Fraction\ lost = [(3\rho)^3/3!] / [1+(3\rho)+(3\rho)2/2+(3\rho)^3/3!]. \tag{16}$$

In all the models discussed above, we have assumed a ***FIFO*** queue discipline. All jobs are treated equallly on a first-come, first-served basis. But it may be possible to reduce congestion in systems by the use of priorities in scheduling.

In many important applications in the real world, queues exist not just as a single station but as a part of a network of queues.

5. Conclusions

Processing systems are an important part of most business and public sector operations. They can represent flows of goods and parts through a manufacturing process or flows of paperwork through back-office systems. They can represent flows of peaple through reservation and check-in systems or telephone inquiries.

They can even represent flows of information in computer networks. Congestion is a significant and costly component of many of these systems.

This paper presented models for representing some simple types of systems: single- and multiple-channel queuing systems and loss systems. These may be used to make estimates of congestion in real-world systems.

But more important is that these models provide insights that can be applied to a great variety of processing systems and that are not at all obvious [3]. These insights are:

- ***There is a need for planned excess capacity.*** Management might like to see facilities as close to 100 percent utilized as possible. But waiting time and congestion is very sensitive to the system load factor, particularly when it gets above 70 to 80 percent. So one must plan for some idle capacity in processing systems.

- ***Variability is a major culprit in congestion.*** The amount of congestion in a processing system is directlly related to the amount of variability in interarrival and service times. Management can significantly improve congestion by reducing these sources of variability.

- ***Pooling can lead to significant improvements.*** Pooling of resources, so that jobs can be serviced by alternate service facilities, was shown by formulas and examples to significantly reduce congestion. There are many opportunities for pooling, some of which are not obvious. Cross-trainig of personnel on multiple tasks is one exemple. If a firm has a mechanic who specializes in one type of repair (engines, for exemple) and a second who only handles a second type of repair (brakes and struts, for example), there will be situations when one has a big backlog and the other is idle. However, if they can be cross-trained to handle either type of job, it becomes a two-channel system with significant reduction in congestion. This is just one exampleof the many opportunities for pooling.

- ***Scheduling matters.*** The exemples in the real life ilustrated that how jobs are scheduled through the system can significantly impact congestion.

- ***These factors can have big impacts.*** They can have orders of magnitude effects on the congestion in systems. Management can significantly improve processes.

- ***They are not intuitively obvious.*** Our intuition about processing systems is not adequate to understand the size of these effects. Hopefuly, the models in this paper will lead to a better understanding.

R E F E R E N C E S

[1] *F. Hillier, G.J. Liebennan*, Introduction to Operations Research. 6th ed. New York: McGraw Hill,1995.

[2] *H.T. Papadopoulos, C. Heavez and J. Browne,* Queuing Theory in Manufacturing Systems Analysis and Design. London: Chapman & Hall, 1993.

[3] *D.C. Şerban*, Managementul operaţiunilor – note de curs, Universitatea POILTEHNICA din Bucureşti, Şcoala de Studii Academice Postuniversitare de Management, Bucureşti, 2007.