This paper discusses a queuing model for a non-priority time-sharing environment in which all active tasks fit in a homogeneous main storage. Design parameters such as queue length and response time, as well as their distributions, can be estimated with the aid of the model. The model provides a basic frame of reference for the development of more complicated models.

A queuing model for a simple case of time sharing

by W. Chang

Of late, considerable attention has been given to time-sharing computer systems. To provide a close relationship between the computer and its users, these systems typically handle a variety of remote terminals, from typewriters to computer graphic display systems. Although many source terminals may employ the computer concurrently, the objective is to provide a form of terminal service that makes it appear to each terminal operator as if he were the only one using the computer.

The time the computer spends on scheduling, allocating, buffering, and controlling terminal input and output represents a slice of processing time that may be called "overhead." For time-sharing to succeed, the improvement in problem-solving effectiveness and user convenience must more than offset the overhead loss. One of the design objectives in time-sharing systems is therefore to minimize overhead. Because many tasks may be handled concurrently by the computer system, each task is served in turn by the computer for a time period that will be called a *quantum*. Depending upon system specifications and design considerations, the quantum may be either a constant interval, or a random interval with a probability distribution. In this paper, the quantum will be treated as a random variable. The constant quantum may then be considered a special case of the general analysis.

An important design problem is that of estimating system

response time (often called "turn-around" time). For a given task and system, response time is defined as the interval between arrival of input and departure of results. In most contexts, response time is not a constant but a random variable, and the probability distribution of this variable is a critical design criterion. Since the number of jobs presented to the system typically influences the required amount of main storage, another problem is to determine the distribution of job-queue length.

The purpose of this paper is to present techniques for estimating response time and queue length in one postulated time-sharing environment. Since various time-sharing schemes have been proposed, several mathematical models would be required to analyze them all; to avoid excessive detail, we limit ourselves here to one basic time-sharing model for the central processing unit.[1] The queuing behavior of a system is analyzed assuming that jobs have already arrived at the computer. Delays between terminal and processing unit are not included in the model, although these delays must be given consideration in the actual design of a time-sharing system.[2]

A simple model

The appropriate elementary model for a time-sharing system involves a queue with feedback.[3,4] Let us first examine the following case as shown in Figure 1. Tasks arrive at the computer in message form and wait in turn for service. During each service, one quantum of processing time is allotted. Two conditions may occur: either the task is finished after the service of a quantum (this occurs with probability q) or it is not completed (this occurs with probability p, where $p = 1 - q$). If not completed, a task returns to the queue and awaits its turn for another quantum. At the end of a service quantum, the random variable either p or q is again specified; its new value is independent of its previous value. This assumption simplifies the solution considerably.

input process

The flow of requests for service from each input source may be considered a Poisson process. Let λ_i denote the input density of the Poisson process for source i, where $i = 1, 2, \cdots, N$. Then the total input to the computer is a Poisson process with density λ, where λ denotes the sum of all λ_i.

service time

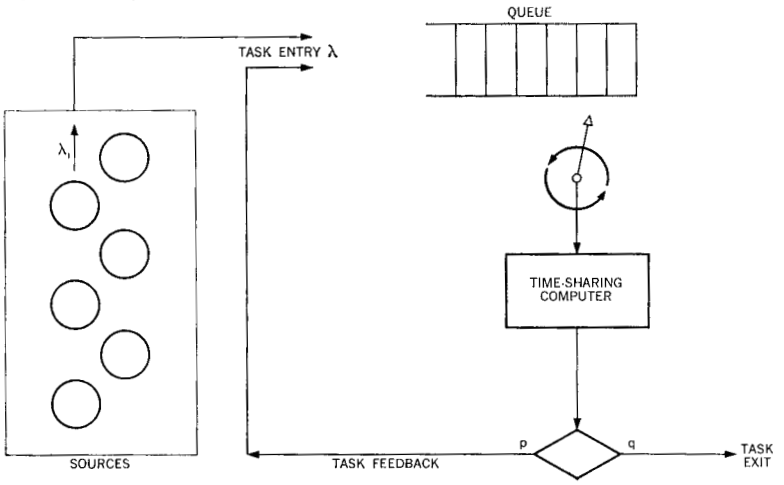
We assume the response time for a task consists of two parts, a *waiting time* and a *service time*, the latter being actual processing time. Let $H(x)$ denote the service time distribution, and define the Laplace transform[5] of $H(x)$ as

$$\psi(s) = \int_0^\infty e^{-sx}\, dH(x)$$

and the rth moment as

$$a_r = \int_0^\infty x^r\, dH(x)$$

Figure 1 Simplified model



Let $Q(x)$ be the quantum size distribution with Laplace transform

$$\phi(s) = \int_0^\infty e^{-sx}\, dQ(x)$$

and rth moment

$$c_r = \int_0^\infty x^r\, dQ(x)$$

Referring to the definition of the model, the relation between $H(x)$ and $Q(x)$ is

$$H(x) = q \sum_{n=1}^{\infty} p^{n-1} Q_n(x)$$

where $Q_n(x)$ is the nth folded convolution of $Q(x)$ with itself. The Laplace transform is

$$\psi(s) = q \sum_{n=1}^{\infty} p^{n-1} [\phi(s)]^n$$

$$= \frac{q\phi(s)}{1 - p\phi(s)} \tag{1}$$

Equation 1 is a relationship between the Laplace transforms of the service time and the quantum size distributions in the simplified time-sharing model (Figure 1). If one of the Laplace transforms is known, the other can be determined by Equation 1.

The value of q can be determined as follows. From the definition of probability expectation and the Laplace transform, the average service time is given by

$$a_1 = -\psi'(0)$$

and the average quantum size by

$$c_1 = -\phi'(0)$$

Taking the derivatives of both sides of Equation 1 and setting $s = 0$, we have (since $p + q = 1$)

$$a_1 = \frac{qc_1}{(1-p)^2} = \frac{c_1}{q} \tag{2}$$

Example 1 Suppose that service time is exponentially distributed (i.e., that $H(x) = 1 - e^{-\mu x}$) and that the average quantum size to be used is c_1. The problem is to determine the quantum size distribution.

The average service time is

$$a_1 = \int_0^\infty x\, dH(x) = \frac{1}{\mu}$$

and from Equation 2,

$$q = c_1/a_1 = \mu c_1$$

Similarly, p is given by $p = 1 - \mu c_1$. Equation 1 can be rewritten as

$$\phi(s) = \frac{\psi(s)}{q + p\psi(s)} \tag{3}$$

Since

$$\psi(s) = \int_0^\infty e^{-sx}\, dH(x) = \frac{\mu}{s+\mu}$$

we have

$$\phi(s) = \frac{\mu}{qs+\mu} = \frac{\mu/q}{s+\mu/q}$$

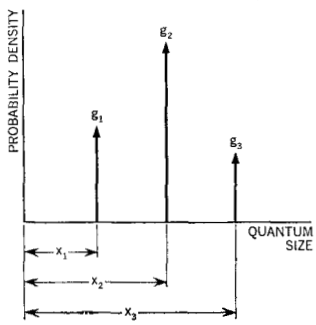
From this equation, we obtain the quantum size distribution $Q(x)$ by

$$Q(x) = 1 - e^{-\mu/qx}$$

Thus the quantum size is also exponentially distributed, and has a mean of q/μ.

Given $Q(x)$, we can approximate the service time distribution $H(x)$ by $H^*(x)$, where the expectation of $H^*(x)$ is equal to the actual service time a_1. $\psi^*(s)$, the Laplace transform of $H^*(x)$, can be determined by Equation 1.

Figure 2 Examples of quantum size and probability density



Example 2 We are given the values of three quantum sizes, x_1, x_2, and x_3, with corresponding probability densities g_1, g_2, g_3 as shown in Figure 2. The average service time a_1 is also given. Again let $H^*(x)$ denote an approximation to the service time distribution, where the expectation of $H^*(x)$ is equal to a_1. The problem is to determine $\psi^*(s)$, the Laplace transform of $H^*(x)$.

The average quantum size is

$$c_1 = g_1x_1 + g_2x_2 + g_3x_3$$

The value of q can be determined from Equation 2 and, as usual, $p = 1 - q$. The Laplace transform of the quantum size distribution is obtained from

$$\phi(s) = \int_0^\infty e^{-sx}\, dQ(x)$$

$$= g_1 e^{-sx_1} + g_2 e^{-sx_2} + g_3 e^{-sx_3}$$

and $\psi^*(s)$ from

$$\psi^*(s) = \frac{q\phi(s)}{1 - p\phi(s)}$$

$$= \frac{q[g_1 e^{-sx_1} + g_2 e^{-sx_2} + g_3 e^{-sx_3}]}{1 - p[g_1 e^{-sx_1} + g_2 e^{-sx_2} + g_3 e^{-sx_3}]}$$

In the case of constant quantum size, $c_1 = c$, then $H^*(x)$ is a geometric distribution.[4]

If the overhead loss within each quantum cannot be neglected, the above expressions must be modified. Let $Q_0(x)$ be the overhead loss within each quantum, and let $\phi_0(s)$ be its Laplace transform. The useful portion of the processing time within a quantum can be easily obtained. The Laplace transform of these useful portions is $\phi(s)/\phi_0(s)$. Replacing $\phi(s)$ in Equation 3 by $\phi(s)/\phi_0(s)$, we obtain the relation between quantum size and the actual service time.

queue size evaluation

Referring again to Figure 1, let ξ_n be the queue size in the system immediately after the completion of a nth quantum $\{n = 1, 2, 3, \cdots\}$. Define the generating function $U(z)$ as

$$U(z) = E\{z^{\xi_n}\}$$

The function $U(z)$ satisfies the relation

$$U(z) = p[U(z) - U(0)]\phi[\lambda(1 - z)]$$

$$+ q\left[\frac{U(z) - U(0)}{z}\right]\phi[\lambda(1 - z)]$$

$$+ (q + pz)U(0)\phi[\lambda(1 - z)]$$

This is true because the queue sizes at n and $n + 1$ form a Markov chain. To find $U(z) = E\{z^{\xi_{n+1}}\}$, we take into consideration that immediately after the nth quantum, three cases may prevail:

- Service was not completed; an additional quantum is desired by the current task, and the queue size is not reduced
- Service was completed; queue size is reduced by one
- Service was not required; the queue was empty before the beginning of the nth quantum

However, additional customers may have arrived during the nth quantum. These additional customers are expressed in terms of $\phi[\lambda(1 - z)]$ as discussed by Takacs.[6] Solving for $U(z)$, we have

The average queue size is equal to the number of new arrivals during the average response interval.

The second moment T_2 can be obtained if $\theta(s)$ is known. $\theta(s)$ can only be implicitly determined. For a detailed discussion of $\theta(s)$, we refer to Takacs;[3] the following is merely a summary derivation of $\theta(s)$.

Define the compound generating function $U(s, z)$ as

$$U(s, z) = \left(1 - \frac{\lambda c_1}{q}\right) \frac{\lambda z(1 - z)\{\phi(s) - \phi[\lambda(1 - z)]\}}{\{z - (q + pz)\phi[\lambda(1 - z)]\}[s - \lambda(1 - z)]}$$

Further define

$$U_1(s, z) = P_0\phi[s + \lambda(1 - z)] + U\{s + \lambda(1 - z), (q + pz)\phi[s + \lambda(1 - z)]\}$$

where

$$P_0 = 1 - \lambda a_1$$

and

$$U_{k+1}(s, z) = \phi[s + \lambda(1 - z)]U_k\{s, (q + pz)\phi[s + \lambda(1 - z)]\}$$
$$k = 1, 2, \cdots$$

From these expressions, the Laplace transform of the response time distribution is

$$\theta(s) = q \sum_{k=1}^{\infty} p^{k-1}U_k(s, 1)$$

The first two moments may be obtained as follows

$$T_1 = \frac{\lambda c_2 + 2c_1(1 - \lambda c_1)}{2(q - \lambda c_1)}$$

$$T_2 = \frac{q^2 - 2q}{6(q - \lambda c_1)^2[q^2 - q(2 + \lambda c_1) + \lambda c_1]} \cdot \{2q[6\lambda c_1^3 - 6c_1^2 - 6\lambda c_1 c_2 + 3c_2 + \lambda c_3] - [12\lambda c_1^3 - 12c_1^2 - 6\lambda c_1 c_2 + 2\lambda^2 c_1 c_3 - 3\lambda^2 c_2^2]\}$$

The second moment is useful in determining the variation of the response time.

busy period

The busy period of the computer can be obtained from a single-server queue formulation,[6] since the order of service within a busy period is immaterial in the analysis. Let $D(x)$ be the busy period distribution and let $\gamma(s)$ be its Laplace transform. $\gamma(s)$ can be obtained as the root with the smallest absolute value of the equation

$$z = \psi[s + \lambda(1 - z)]$$

The average busy period, d, may be determined as

$$d = \frac{a_2}{1 - \lambda a_1}$$

storage requirement

In this analysis we assume that the entire queue is located in main storage. To determine the requirement for queue storage, let f_i be the probability that a task requires i units of storage. Define the generating function $F(z)$ as

$$F(z) = \sum_{i=0}^{\infty} f_i z^i \tag{6}$$

Let v_i be the probability that i units of storage are used. Let $V(z)$ be the generating function

$$V(z) = \sum_{i=0}^{\infty} v_i z^i$$

Then $V(z)$ can be obtained as follows:

$$V(z) = \sum_{k=0}^{\infty} p_k [F(z)]^k = U[F(z)]$$

This formulation is the so-called compound generating function defined by Feller.[7]

The average storage required for accommodating the queues is

$$V'(1) = U'(1)F'(1)$$

The second moment of the core storage can be obtained by

$$V''(1) = U''(1)[F'(1)]^2 + U'(1)F''(1)$$

Note that $U(z)$ is used instead of $U^*(z)$ because the queue size at the completion of a quantum determines the status of the computer at operation. In the event that a constant storage size k is needed for every task present in the system, Equation 6 can be simplified to

$$F(z) = z^k$$

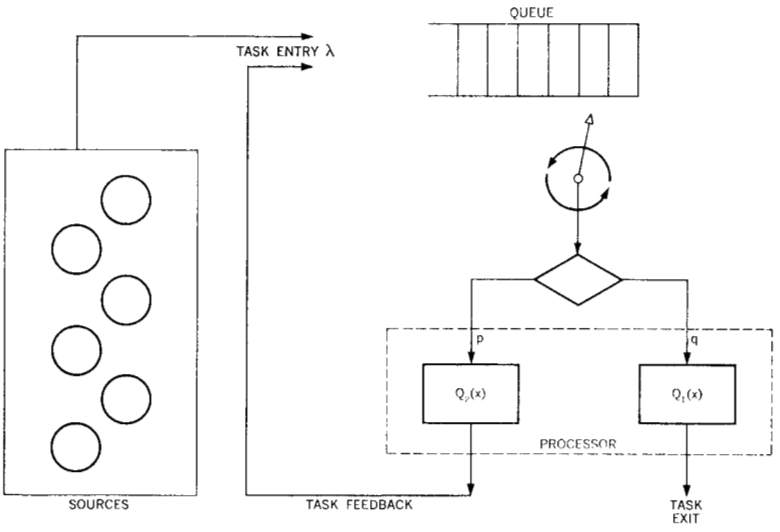
A somewhat extended model

If the binary decision involving p and q (see Figure 1) takes place before a quantum is given to a task, we can obtain the more interesting model suggested by Figure 3. In this case, if the outcome is q, the service time follows distribution $Q_1(x)$ and the task is terminated. If the outcome is p, the service time follows distribution $Q_2(x)$ and the task again joins the queue. Given this "look-ahead" capability, the model can be used to analyze one or more different quanta distributions, and to better accommodate tasks (such as short debugging runs) that can be completed within one service period. For example, if the task is completed before the end of the last quantum allotted to the task, the remaining time is usually made available to other tasks. The use of different quantum sizes can also take such actions into consideration. Define

$$\phi_1(s) = \int_0^{\infty} e^{-sx}\, dQ_1(x), \qquad b_r = \int_0^{\infty} x^r\, dQ_1(x)$$

$$\phi_2(s) = \int_0^{\infty} e^{-sx}\, dQ_2(x), \qquad e_r = \int_0^{\infty} x^r\, dQ_2(x)$$

Figure 3 Extended model



Then the Laplace transform of service time distribution for a task is

$$\psi(s) = \frac{q\phi_1(s)}{1 - p\phi_2(s)}$$

The following relation holds for the queue-size generating function immediately at the completion of a quantum:

$$U(z) = \frac{q[U(z) - U(0)]\phi_1[\lambda(1 - z)]}{z} + p[U(z) - U(0)]\phi_2[\lambda(1 - z)] + U(0)\{pz\phi_2[\lambda(1 - z)] + q\phi_1[\lambda(1 - z)]\}$$

Since $U(1) = 1$, we have

$$U(0) = q - \lambda(qb_1 + pe_1)$$

Thus

$$U(z) = \frac{[q - \lambda(qb_1 + pe_1)]\{pz(z - 1)\phi_2[\lambda(1 - z)] + q(z - 1)\phi_1[\lambda(1 - z)]\}}{z - q\phi_1[\lambda(1 - z)] - pz\phi_2[\lambda(1 - z)]}$$

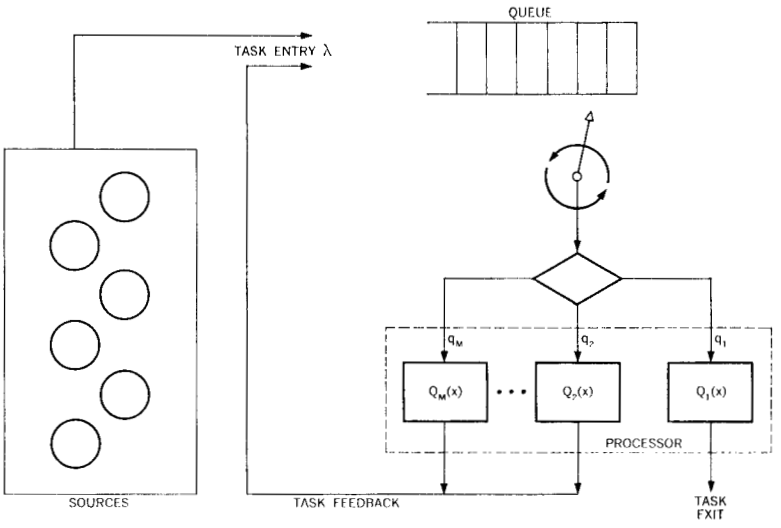
Using a similar approach, $U^*(z)$ can be determined as

$$U^*(z) = \frac{U(0)(z - 1)\phi_1[\lambda(1 - z)]}{z - q\phi_1[\lambda(1 - z)] - pz\phi_2[\lambda(1 - z)]}$$

Further extension of the model is shown in Figure 4. In this model, there are M different quantum-size distributions $Q_1(x)$, $Q_2(x) \cdots , Q_M(x)$, where

$$\sum_{i=1}^{M} q_i = 1$$

Figure 4 Further-extended model



Except in the case of q_1, all q's are feedback loops. Let $\phi_i(s)$ be the Laplace transform of $Q_i(x)$, for $i = 1, 2, \cdots, M$.
Then

$$U(z) = U(0) \frac{(z-1)\{q_1\phi_1[\lambda(1-z)] + \sum_{i=2}^{M} zq_i\phi_i[\lambda(1-z)]\}}{z - \sum_{i=2}^{M} q_i z\phi_i[\lambda(1-z)] - q_1\phi_1[\lambda(1-z)]}$$

and

$$U(0) = q_1 - \lambda \sum_{i=1}^{M} q_i[-\phi_i'(0)]$$

Similarly, we determine the queue-size generating function for the extended model immediately after the departure of a task as

$$U^*(z) = \frac{U(0)(z-1)\phi_1[\lambda(1-z)]}{z - \sum_{i=2}^{M} q_i z\phi_i[\lambda(1-z)] - q_1\phi_1[\lambda_1(1-z)]}$$

Concluding remarks

This paper discusses a mathematical model of a time-shared processing unit and an extended model using queues with feedback. The generating functions $U(z)$ for the queue size immediately after the completion of a quantum, as well as the generating functions $U^*(z)$ for the queue size immediately after the departure of a task, are derived. $U(z)$ is useful in determining main storage requirements, $U^*(z)$ in analyzing the average response time. The higher moments of response-time distribution can be obtained only through a more complicated analysis.[3] Under the limitation of these models; namely, that they consider only the processor,

the average response time is seen to be the same for all quantum sizes if the overhead loss within a quantum is negligible. The second moment of the response-time distribution decreases as quantum size increases. These observations verify, at least in view of processing-unit efficiency, that a quantum should be specified as large as possible. The amount of overhead and the variance of the response time are made smaller by taking larger quanta. Actual choices of quantum sizes must account for additional factors such as input-output requirements and terminal delays.

Although the models discussed are by no means sufficiently general to cover all time-sharing environments, they apply rather closely to some applications of medium-sized computers. Furthermore, unless we are to be entirely content with simulation exercises, time-sharing analysis merits continued efforts in the postulation and elaboration of mathematical models.

ACKNOWLEDGEMENT

The author wishes to thank H. D. Leed, P. H. Seaman, S. Shiao, and D. J. Wong for many profitable discussions related to this subject.

CITED REFERENCES AND FOOTNOTES

1. Some partially solved problems omitted from this paper are priority assignments of tasks to be processed, main storage allocation and usage, and processing controls, all of which need additional mathematical models. Effort is being made to solve these problems.
2. W. Chang and D. J. Wong, "Analysis of real-time computer systems," *Bulletin of the Operations Research Society of America* **13**, Supplement 1, Abstract B-35 (Spring 1965).
3. L. Takacs, "A single-server queue with feedback," *Bell System Technical Journal* **42**, 2, 505–519, (March 1963).
4. L. Kleinrock, "Analysis of a time-shared processor," *Naval Research Logistics Quarterly* **11**, 1, 59–73 (March 1964).
5. Methods for computing the time domain function (inverse Laplace transform) may be found in W. R. LePage, *Complex Variables and the Laplace Transform for Engineers*, McGraw Hill Book Company, New York (1961).
6. L. Takacs, *Introduction to the Theory of Queues*, Oxford University Press, New York, New York (1962).
7. W. Feller, *An Introduction to Probability Theory and its Applications*, John Wiley & Sons, Inc., New York, New York (1950).