D. INFORMATION FLOW IN LARGE COMMUNICATION NETS

In continuing our research[1] on the problems of information flow in large communication nets results have been obtained (for a single node) for two classes of queue disciplines: priority queueing, and time-shared servicing. A law of conservation has been proved which constrains the allowed variation in the average waiting times over the set of priority classes.

1. Priority Queueing

For priority queueing, the input traffic is broken up into P priority classes. Units from priority class p (p=1, 2, . . . , P) arrive in a Poisson stream with an average rate

λ_p units per second; each unit from this priority class has a total required processing time selected independently from an exponential distribution, with mean $1/\mu_p$. We define

$$\rho_p = \lambda_p/\mu_p,$$

$$\rho = \sum_{p=1}^{P} \rho_p,$$

and

$$W_o = \sum_{p=1}^{P} \rho_p/\mu_p.$$

The priority structure is such that a unit from the p^{th} priority class entering the queue at time T is assigned a number b_p, where $0 \leqslant b_1 \leqslant b_2 \leqslant \ldots \leqslant b_P$. The priority $q_p(t)$, at time t, associated with such a unit is

$$q_p(t) = (t-T)\, b_p.$$

The effect of this priority assignment is to increase a unit's priority in proportion to the time that elapsed since that unit's arrival at the system (referred to as a delay-dependent priority system).

Let us define W_p to be the expected value of the time spent in the queue for a unit from the p^{th} priority class. We then state the following theorem.

THEOREM 1: For the delay-dependent priority system described above, and for $0 \leqslant \rho < 1$,

$$W_p = \frac{\dfrac{W_o}{1-\rho} - \displaystyle\sum_{i=1}^{p-1} \rho_i W_i \left(1 - \frac{b_i}{b_p}\right)}{1 - \displaystyle\sum_{i=p+1}^{P} \rho_i \left(1 - \frac{b_p}{b_i}\right)} .$$

From a designer's point of view, the introduction of the P independent quantities b_p is an asset. Consider the problem of a system designer who is faced with assigning some priority structure to a queueing system. Let us assume that he is given the quantities λ_p, μ_p, and P, that is, he is given the desired input traffic and partitioning. From these parameters, he can easily calculate ρ. With the free parameters b_p, he can then attain any value for W_p (for this value of ρ) within broad limits. Without these additional degrees of freedom, the set W_p would be fixed (as for a commonly used priority structure[2] for which $q_p(t) = a_p$ and a_p is independent of time).

2. A Conservation Law

As one might expect, there is a certain trade-off of waiting time among the various priority classes. In particular, let us define a class of queueing disciplines as follows:

(a) Arrival statistics are Poisson with an average arrival rate λ_p for the p^{th} priority class.

(b) Service-time statistics are arbitrary with mean $1/\mu_p$ for the p^{th} priority class.

(c) All units remain in the system until completely served.

(d) The service facility is never idle if there are any units in the system.

(e) Pre-emption (the replacement of a low-priority unit in service by a higher priority unit) is allowed only if the service-time distributions are exponential, and if upon re-entry into the service facility the low-priority unit continues from the point at which its service was interrupted.

THEOREM 2: For any queue discipline and any fixed-arrival and service-time distributions that are subject to the restrictions stated above

$$\sum_{p=1}^{P} \rho_p W_p = \text{constant} = \begin{cases} \dfrac{\rho}{1-\rho} V & \rho < 1 \\ \infty & \rho \geqslant 1 \end{cases}$$

where

$$V = \frac{1}{2} \sum_{p=1}^{P} \lambda_p E(t_p^2)$$

and

$E(t_p^2)$ = second moment of the service-time distribution for priority class p.

This conservation law constrains the allowed variation in the average waiting time for any queue discipline that falls into this wide class.

3. Time-Shared Servicing

For a time-shared servicing facility, we consider time to be quantized into intervals, each of which is Q seconds in length. At the end of each time interval, a new unit arrives in the system with probability λQ (result of a Bernoulli trial); thus the average number of arrivals per second is λ. The service time of a newly arriving unit is chosen independently from a geometric distribution so that for $\sigma < 1$,

$$s_n = (1-\sigma)\,\sigma^{n-1} \qquad n = 1, 2, 3, \ldots$$

where s_n is the probability that a unit's service time is exactly n time intervals long.

The procedure for servicing is as follows: A unit upon arrival joins the end of the queue, and waits on line in a first come first served fashion until it finally arrives at the service facility. The server picks the next unit in the queue and performs a unit of service upon it. At the end of this time interval, the unit leaves the system if its service is finished; if not, it joins the end of the queue with its service partially completed. Obviously, a unit whose service time is n intervals long will be forced to join the queue a total of n times before its service is completed. Another assumption must now be made regarding the order in which events take place at the end of a time interval. We shall assume that the unit leaving the service facility is allowed to join the tail of the queue before the next unit arrives at the queue from outside the system (referred to as a late-arrival system). The case with reversed order has also been solved, but will not be reported on here, since the results are not essentially different.

Upon arrival, a unit finds some number of units, m, in the system. The expected value, E(m), of the number m is known[3] to be

$$E(m) = \frac{\rho}{1 - \rho}\sigma$$

where

$$\rho = \frac{\lambda Q}{1 - \sigma}.$$

We are now ready to state the following theorem.

THEOREM 3: The expected value, T_n, of the total time spent in the late-arrival system for a unit whose service time is nQ seconds, is

$$T_n = \frac{nQ}{1 - \rho} - \frac{\lambda Q^2}{1 - \rho}\left\{1 + \frac{(1-\sigma a)(1-a^{n-1})}{(1-\sigma)^2 (1-\rho)}\right\}$$

where

$$a = \sigma + \lambda Q.$$

Now, instead of the round-robin type of structure just described, we shall consider a strict first come first served system in which each unit waits for service in order of arrival, and, once it is in service, each unit remains until it is completely serviced. Then for T_n defined as before, we state the following theorem.

THEOREM 4: The expected value, T_n, of the total time spent in the first come first served system for a unit whose service time is nQ seconds, is

$$T_n = \frac{1}{1 - \sigma} QE(m) + nQ$$

where E(m) is as defined above.

Now if one wishes an approximate solution to the round-robin system, one might argue as follows: Each time a unit (the tagged unit, say) returns to the queue, it finds E(m) units in the system ahead of it (this is the approximation). Each of these units will spend Q seconds in the service facility before the tagged unit arrives at the service facility. Since the tagged unit must go through this process n times, the total time that it spends in the queue is nQE(m). Also, it spends exactly nQ seconds in the service facility itself. Thus, our approximate solution, T_n', turns out to be

$$T_n' = nQE(m) + nQ.$$

Comparing this solution with the result for the first come first served case, we see that there is a critical value of n, say n_{crit}, at the point $n_{crit} = \frac{1}{1-\sigma}$. In fact, we observe that the quantity $\frac{1}{1-\sigma}$ is merely the mean value, $\overline{n}$, of the number of service intervals required by a unit. Thus, the approximate solution shows us that units whose service time is greater (or less) than the average time, $\overline{n}Q$, spend more (or less) time in the round-robin system than in a strict first come first served system, that is, units with short service-time requirements are given preferential treatment over units with longer requirements. The fact that the critical length is equal to the average length is a surprisingly simple result. It has also been shown that the approximation is excellent.

It is interesting to note that the round-robin and first come first served disciplines offer an example of the validity of the conservation law. That is, if we define $W_n = T_n - nQ$, which is the average waiting time in the queue, then it is a simple algebraic exercise to show that

$$\sum_{n=1}^{\infty} \rho_n W_n \text{ (first come first served)} = \sum_{n=1}^{\infty} \rho_n W_n \text{ (round-robin)} = \frac{Q\rho^2\sigma}{(1-\rho)(1-\sigma)}$$

where

$$\rho_n = \rho s_n = \rho(1-\sigma)\,\sigma^{n-1}.$$

L. Kleinrock

References

1. L. Kleinrock, Information flow in large communication networks, Quarterly Progress Report No. 62, Research Laboratory of Electronics, M.I.T., July 15, 1961, pp. 162-163.

2. A. Cobham, Priority assignments in waiting line problems, Operations Researc Vol. 2, pp. 70-76, 1954.

3. J. R. Jackson, Some problems in queueing with dynamic priorities, Naval Research Logistics Quarterly, Vol. 7, p. 235, 1960.