normalized queuing delay due to buffering is equal to 1.25 character-service times. Since each service time equals $1/\mu = 1/240 = 4.16$ ms, the waiting time of each character is 5.06 ms. Now suppose the number of terminals increases from 48 to 96, so that the traffic intensity is less than unity, two transmission lines are needed, and the traffic intensity is still equal to 0.6. From Fig. 2, the buffer length corresponding to the desired overflow probability for two transmission lines is about 14 characters. The waiting time is about 0.8 character-service times which is equal to 3.33 ms. Although the difference between 5.06 and 3.33 ms may not be detected by a user at a terminal, a common buffer of the same size operating with two output lines can handle twice the number of input lines as with one output line. Thus, the common buffer approach permits handling a wide range of traffic without substantial variation in buffer size.

References

[1] W. W. Chu, "A study of asynchronous time division multiplexing for time-sharing computer communications," presented at the 2nd Hawaii Internatl. Conf. System Sciences (Honolulu, Hawaii), January 22–24, 1969.

[2] B. A. Powell and B. Avi-Itzhak, "Queuing systems with enforced idle time," *Operations Res.*, vol. 15, no. 16, pp. 1145–1156, November 1967.

[3] T. G. Birdsall, M. P. Ristenbatt, and S. B. Weinstein, "Analysis of asynchronous time multiplexing of speech sources," *IRE Trans. Communications Systems*. vol. CS-10, pp. 390–397, December 1962.

[4] N. M. Dor, "Guide to the length of buffer storage required for random (Poisson) input and constant output rates," *IEEE Trans. Electronic Computers* (Short Notes), vol. EC-16, pp. 683–684, October 1967.

[5] J. D. C. Little, "A proof of the queuing formula $L=\lambda w$," *Operations Res.*, vol. 9, pp. 383–387, 1961.

[6] R. W. Hamming, *Numerical Methods for Scientists and Engineers*. New York: McGraw-Hill, 1962, pp. 363–364.

[7] P. M. Morse, *Queues, Inventories and Maintenance*. New York: Wiley, 1958, pp. 15–18.

[8] W. W. Chu, "Optimal file allocation in a multiple computer system," *IEEE Trans. Computers*, vol. C-18, pp. 885–889, October 1969.

Swap-Time Considerations in Time-Shared Systems

LEONARD KLEINROCK, MEMBER, IEEE

Abstract—**Solved for is the expected swap time expended for those customers in the system of queues in general models of time-shared systems. This quantity is expressed in terms of the expected queueing time conditioned on required service time and is applied to a number of examples of interest.**

Index Terms—**Modeling and analysis, processor-sharing, queueing analysis, scheduling, swap time, time-shared.**

Introduction[1]

NUMEROUS authors have addressed themselves to the problem of solving for the average response time T in time-shared computer systems [1]–[11] and an excellent summary of such investigations is available [9]. Many of these studies condition T on the required service time (i.e., the required processing time) t which we denote by $T(t)$. Some go further and introduce an external priority system, solving for $T_p(t)$ which is the average response time for a customer from priority group p who requires t seconds of processing time [4].

Recently a new quantity, the distribution of attained service time $N_p(\tau)$ was calculated [6] for any priority feedback queueing system which satisfies Little's result (which states that the average number to be found in the system is equal to the average arrival rate of customers times the average time spent in that system). $N_p(\tau)$ is defined as the expectation of the number of customers in the system of queues from priority group p who have so far received exactly τ seconds of useful processing. A customer is said to be in the system of queues whenever he is waiting for his next quantum of service time during his request for t seconds of total service. We assume that on his nth visit to service, a customer from priority group p will receive the attention of the processing unit for $g_{pn}Q$ seconds.

Results

In this paper, we are interested in swap-time considerations. (Swap time is the time spent in removing the old customer from and bringing the new customer into service, as well as any other cost in time directly related to this operation.) Our main result is that a measure of this quantity is simply expressed in terms of $N_p(\tau)$ as follows. We direct our attention to the customers in the system of queues and we inquire as to the expected time S, which has been expended in swapping for this set of customers. The answer comes directly from $N_p(\tau)$. First we note that all customers from group p who have visited the service facility exactly n times

Manuscript received August 14, 1969; revised December 16, 1969. This work was supported by the Advanced Research Projects Agency of the Department of Defense under Contract DAHC15-69-C-0285. Reproduction in whole or in part is permitted for any purpose of the United States Government.

The author is with the School of Engineering and Applied Science, University of California, Los Angeles, Calif. 90024.

[1] This paper evolved from and is an extension of the paper in [3].

must so far have received useful processing in an amount equal to

$$\tau_n = \sum_{i=1}^{n} (g_{pi}Q - \theta_{pi}) \text{ seconds} \tag{1}$$

where θ_{pn} = (wasted) swap time used for a customer from group p on his nth visit to service.[2] Thus only τ_n will appear as a meaningful argument for $N_p(\tau)$. Clearly, then, the expected swap time expended for all customers from group p who are in the system of queues must be

$$S_p = \sum_{n=1}^{\infty} \gamma_{pn} N_p(\tau_n) \tag{2}$$

where

$$\gamma_{pn} = \sum_{i=1}^{n} \theta_{pi}. \tag{3}$$

But from Theorem 1 of [6] we have that

$$N_p(\tau_n) = \lambda_p[1 - B_p(\tau_n)][W_p(\tau_{n+1}) - W_p(\tau_n)] \tag{4}$$

where

λ_p = average arrival rate of customers from group p,

$B_p(t) = P$(required processing time for customers from pth priority group $\leq t$), and

$W_p(t)$ = expected wait in queues for customers from group p who require a total of t seconds of processing.

Finally, to solve for S (the expected swap time expended on all customers in the system of queues) we have the following.

Theorem 1: For time-shared systems ($Q>0$)

$$S \equiv \sum_{p=1}^{P} S_p = \sum_{p=1}^{P} \sum_{p=1}^{\infty} \gamma_{pn}\lambda_p[1 - B_p(\tau_n)] \cdot [W_p(\tau_{n+1}) - W_p(\tau_n)] \tag{5}$$

where

P = total number of priority groups.

Observe that S is expressed in terms of known quantities (γ_{pn}, λ_p, $B_p(\tau_n)$) and a function $W_p(\tau_n)$ which is the average conditional waiting time; this last measure is the usual one solved for in the analysis of time-shared systems.

Corollary 1: For $\theta_{pn} = \theta_p$ independent of n,

$$S = \sum_{p=1}^{P} \lambda_p\theta_p \left[\sum_{n=1}^{\infty} nb_p(\tau_{n+1}) W_p(\tau_{n+1}) - \sum_{n=1}^{\infty} [1 - B_p(\tau_n)]W_p(\tau_n) \right] \tag{6}$$

[2] We assume the obvious condition that $g_{pn}Q \geq \theta_{pn}$ for all n.

where

$$b_p(\tau_{n+1}) = B_p(\tau_{n+1}) - B_p(\tau_n)$$
$$= P[\text{customer from group } p \text{ has required service time } t \text{ such that } \tau_n < t \leq \tau_{n+1}]$$
$$= P[\text{customer from group } p \text{ requires exactly } n+1 \text{ visits to the service facility}].$$

Proof: We have $\gamma_{pn} = n\theta_p$. Substituting this in (5) and using $1 - B_p(\tau_n) = [1 - B_p(\tau_{n+1})] + [B_p(\tau_{n+1}) - B_p(\tau_n)]$ we get (6).

Corollary 2: For $\theta_{pn} = \theta_p$ and $P = 1$ (i.e., no priorities and so we drop the subscript p) we have

$$S = \lambda\theta \sum_{n=1}^{\infty} nb(\tau_{n+1}) W(\tau_{n+1}) - \lambda\theta \sum_{n=1}^{\infty} [1 - B(\tau_n)]W(\tau_n). \tag{7}$$

Proof here is immediate from Corollary 1.

We now consider Theorem 1 plus its corollaries for the *processor-shared systems* [4]. These systems are time-shared systems in which the quantum size Q is allowed to shrink to zero. In this limit for $Q \to 0$, we must obviously contain the swap time θ_{pn} in a meaningful way such that $g_{pn}Q \geq \theta_{pn}$. This we do in the (theoretically) natural way, namely, defining

$$\phi_{pn} = \frac{\theta_{pn}}{g_{pn}Q} \tag{8}$$

where we require $0 \leq \phi_{pn} \leq 1$. Thus ϕ_{pn} is that *fraction* of the nth quantum given to a customer from group p which is wasted due to swap time. In the limit as $Q \to 0$, we must then define

$$\phi_p(\tau) = \lim_{Q \to 0} \phi_{pn} \tag{9}$$

where for a given p and τ, we consider an $n = n(Q)$ increasing as Q decreases such that

$$\tau = \lim_{Q \to 0} \sum_{i=1}^{n(Q)} (g_{pi}Q - \theta_{pi}) = \lim_{Q \to 0} \sum_{i=1}^{n(Q)} g_{pi}Q(1 - \phi_{pi}). \tag{10}$$

As discussed in [4], this $Q \to 0$ limit has useful characteristics in our analysis of time-shared systems. Developing the analogous equations here, we have

$$S_p = \int_0^{\infty} \gamma_p(\tau) N_p(\tau) d\tau \tag{11}$$

where

$$\gamma_p(\tau) = \int_0^{\tau} \phi_p(t) dt. \tag{12}$$

Note that $\gamma_p(\tau)$ is the time wasted in providing τ seconds of useful service to a customer from group p. But, from Theorem 2 of [6], we have that

$$N_p(\tau) = \lambda_p[1 - B_p(\tau)] \frac{dW_p(\tau)}{d\tau} \tag{13}$$

where λ_p, $B_p(t)$, and $W_p(t)$ are as defined above and $N_p(\tau)$ is now an expected density function.

We thus have S for this case as follows.

Theorem 2: For processor-shared systems ($Q \to 0$)

$$S = \sum_{p=1}^{P} S_p = \sum_{p=1}^{P} \int_0^\infty \gamma_p(\tau)\lambda_p[1 - B_p(\tau)] \cdot \left(\frac{dW_p(\tau)}{d\tau}\right) d\tau. \tag{14}$$

This last may be interpreted as a Stieltjes integral in the case that $W_p(\tau)$ is discontinuous.

Corollary 1: For $\phi_p(\tau) = \phi_p$ independent of τ, we have

$$S = \sum_{p=1}^{P} \lambda_p \phi_p \int_0^\infty \tau[1 - B_p(\tau)] \left(\frac{dW_p(\tau)}{d\tau}\right) d\tau. \tag{15}$$

Proof here is immediate since $\gamma_p(\tau) = \tau\phi_p$ by (12).

Corollary 2: For $\phi_p(\tau) = \phi_p = \phi$ (i.e., no dependence on τ and $P = 1$ giving no priorities thus allowing us to drop all subscripts) we have

$$S = \lambda\phi \int_0^\infty \tau[1 - B(\tau)] \left(\frac{dW(\tau)}{d\tau}\right) d\tau. \tag{16}$$

(Observe that (11) through (16) for the processor-sharing case are analogous to (2) through (7) for the time-sharing case.) It is important to note in this last (simplest) case that ratio of S to $\bar{\tau}$ (the average attained service) is given simply as follows: $\bar{\tau}$ is defined by

$$\bar{\tau} = \frac{\int_0^\infty \tau N(\tau) d\tau}{\int_0^\infty N(\tau) d\tau} \tag{17}$$

where the denominator is merely $\overline{N}$ = expected number of customers in the system of queues. Since, from (11) and for $P = 1$, we have

$$S = \int_0^\infty \phi\tau N(\tau) d\tau,$$

we then obtain

$$\frac{S}{\bar{\tau}} = \phi\overline{N} \tag{18}$$

which clearly represents the ratio of average (wasted) swap time to average (useful) attained service time for the set of customers still in the system of queues. The simplicity and intuitive appeal of this last equation further supports the utility of using processor-shared models.

Examples

In order to apply the results obtained above, we must find in the published literature solutions for $W_p(t)$ (the expected wait conditioned on a service requirement of t seconds) for time-shared systems which account for swap time. Such results are not especially numerous. We do, however, find some useful examples.

Example 1: Let us apply Corollary 2 of Theorem 1 to the discrete round robin infinite-input population system [5] where

$$b(\tau_{n+1}) = (1 - \sigma)\sigma^n \quad n = 0, 1, 2, \cdots; \; 0 \le \sigma < 1 \tag{19}$$

$$P[1 \text{ arrival in interval of length } Q] = \lambda Q; \quad 0 \le \lambda Q < 1$$

$$P[0 \text{ arrival in interval of length } Q] = 1 - \lambda Q$$

$$g_{pn} = 1; \; P = 1$$

where $P[x]$ is read "probability of x." This system has an exact solution for $W(\tau_n)$, (for $\theta = 0$), and also a simple approximation to $W(\tau_n)$ given below (see [5]):

$$W(\tau_n) = W(nQ) \cong \frac{\rho n Q \sigma}{1 - \rho} \tag{20}$$

where

$$\rho = \lambda Q/(1 - \sigma). \tag{21}$$

When we consider $\theta > 0$, a number of considerations enter. The major question is now to keep the discrete model intact since, for example, if $\theta = Q/3$, then a customer requiring Q seconds of useful processing (*one* quantum for $\theta = 0$) now requires a noninteger number of quanta. No satisfactory way appears to resolve this, and so we will take two approaches. We begin with a continuous distribution of service time, namely the exponential, where $P[\text{service time} \le t] = 1 - e^{-\mu t}$. We then recognize that all customers with $n(Q-\theta) < t \le (n+1)(Q-\theta)$ will need exactly $(n+1)$ visits to the service facility (we make the simplifying, but serious assumption that unused portions of quanta are lost) and the probability that a new arrival satisfies such a condition is

$$b(\tau_{n+1}) = \int_{n(Q-\theta)}^{(n+1)(Q-\theta)} \mu e^{-\mu t} dt = (1 - e^{-\mu(Q-\theta)})[e^{-\mu(Q-\theta)}]n.$$

Comparing this to (19), we make the correspondence

$$\sigma = e^{-\mu(Q-\theta)}. \tag{22}$$

We then apply (20) and (22) to Corollary 2 of Theorem 1 to give for this example,

$$S = \lambda\theta \sum_{n=1}^{\infty} n(1 - \sigma)\sigma^n \rho n Q\sigma/(1 - \rho) - \lambda\theta \sum_{n=1}^{\infty} \sigma^{n+1} \rho n Q\sigma/(1 - \rho).$$

This gives

$$S = \frac{\lambda\theta\rho Q\sigma^2}{(1 - \rho)(1 - \sigma)^2} \tag{23}$$

where ρ and σ are given by (21) and (22), respectively. Note that for $S < \infty$, we require $\rho < 1$ which requires

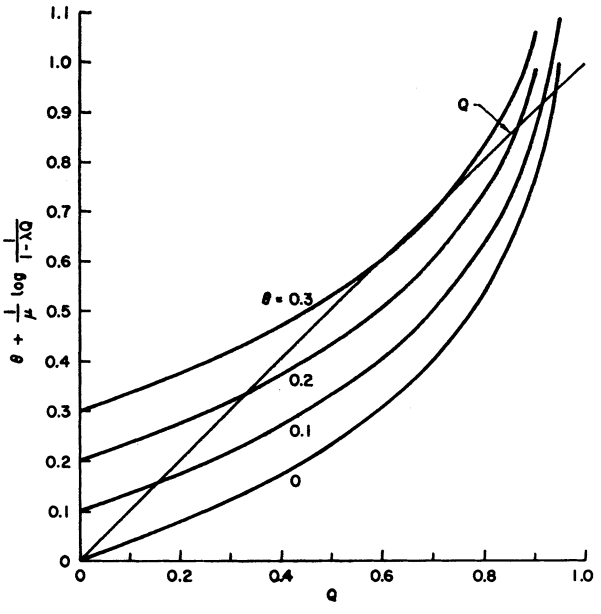

Fig. 1. Allowable values for Q shown as that region for which the curve $\theta+(1/\mu)\log 1/(1-\lambda Q)$ lies below Q. ($\lambda=1, \mu=3$) for Example 1.

$$Q > \theta + \frac{1}{\mu}\log 1/(1-\lambda Q). \tag{24}$$

Equation (24) places lower and upper bounds on Q. The lower bound is due to the restriction that the maximum *effective* service rate must exceed the average arrival rate. The upper bound is due to the wasted excess quantum and also due to the constraint that the arrival probability $\lambda Q<1$.

In Fig. 1 we show the allowed range of Q as determined from (24) for $\lambda=1$, $\mu=3$, and θ as parameter. Fig. 2 gives S as a function of Q (in its allowed range) with θ as parameter again and for the same values of λ and μ. Here we also plot the locus of optimum Q over the family of θ curves to give minimum S.

The second approach to handling the $\theta>0$ case in the first example is to allow the exponential distribution of required service time to remain, but not force it into the discrete form of (19). We then get an equivalent system with $\theta=0$ if we segment this distribution into pieces of width $Q-\theta$, each separated by a gap of size θ as shown in Fig. 3. This results in a mean service time $E(t)$ and a second moment $E(t^2)$ given by

$$E(t) = \frac{1}{\mu} + \frac{\theta}{1-e^{-\mu(Q-\theta)}}$$

$$E(t^2) = \frac{2}{\mu^2} + \frac{\theta\left(\theta+\frac{2}{\mu}\right)}{1-e^{-\mu(Q-\theta)}} + 2Q\theta\frac{e^{-\mu(Q-\theta)}}{\left[1-e^{-\mu(Q-\theta)}\right]^2}.$$

One may now use this service-time distribution in the continuous round robin model studied in [2], as well as in other models of time-shared systems, with additional care to replace Q by $Q-\theta$ in the appropriate places.

We do not carry out this exercise here.

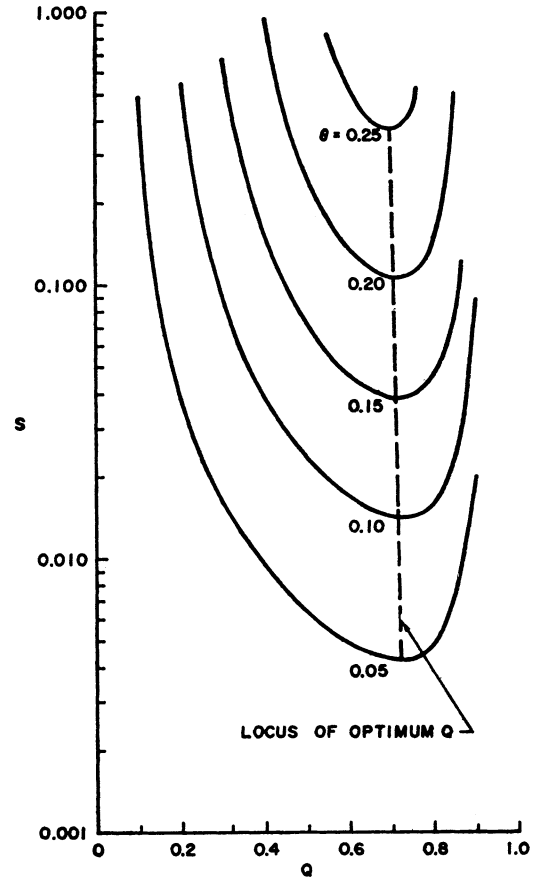

Fig. 2. Average swap time S as a function of quantum size Q with swap time (per quantum) θ as a parameter for Example 1 ($\lambda=1$, $\mu=3$).

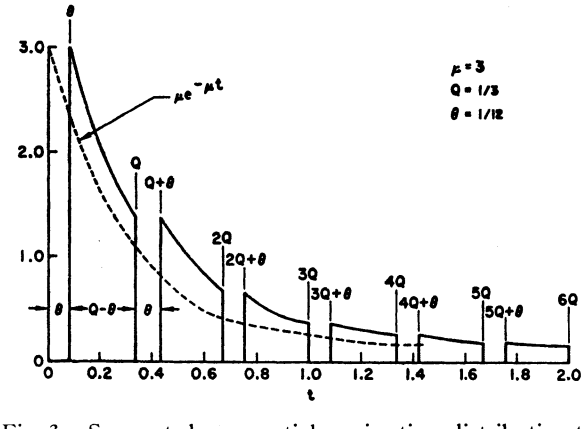

Fig. 3. Segmented exponential service-time distribution to account for swap time θ.

The following examples apply only to the processor-sharing case.

Example 2: Consider the continuous (processor-sharing) round robin infinite-population system [4]. Again we assume $g_{pn}=1$, $P=1$. Here we have Poisson arrivals at rate λ and exponential service of average duration $1/\mu$. For this system we know from [4] that for $\phi=0$,

$$W(\tau) = \rho\tau/(1-\rho) \tag{25}$$

where

$$\rho = \lambda/\mu.$$

The original exponential distribution of service time with parameter μ for this system must now be replaced with another exponential distribution with parameter $\mu(1-\phi)$ where ϕ = fraction of service time wasted (as above).

We now apply Corollary 2 of Theorem 2 to give

$$S = \lambda\phi \int_0^\infty \tau e^{-\mu(1-\phi)\tau}[\rho/(1-\rho)]d\tau.$$

This results in

$$S = \frac{\rho^2\phi}{\mu(1-\rho)(1-\phi)} \tag{26}$$

where

$$\rho = \lambda/\mu(1-\phi). \tag{27}$$

The average swap time S is plotted in Fig. 4 versus ϕ with λ/μ as parameter for this example.

It is interesting to note the application of (18) here. From [4] we have

$$\overline{N} = \rho^2/(1-\rho)$$

and from [6] we have (with the new value $\mu(1-\phi)$)

$$\bar{\tau} = 1/\mu(1-\phi).$$

Thus

$$S = \phi\overline{N}\bar{\tau} = \frac{\rho^2\phi}{\mu(1-\rho)(1-\phi)}$$

which is (26) again.

Further, we can calculate this by considering the average swap time per customer, S'. S' is the product of average service time ($=1/\mu(1-\phi)$) and the fraction of wasted (swap) time ($=\phi$). Thus, $S'=\phi/\mu(1-\phi)$. This multiplied by $\overline{N}$ must also give S, as is obvious.

We now show that the result of Example 1 can be taken to the limit of $Q=0$ with fixed $\phi=\theta/Q$ to give the result of Example 2. From Example 1, we get as $Q\to 0$

$$\begin{aligned}\rho &= \lambda Q/(1-\sigma)\\ &= \lambda Q/(1-e^{-\mu(Q-\theta)})\\ &= \lambda Q/(1-1+\mu(Q-\theta)-\cdots)\\ &= \lambda/\mu(1-\phi).\end{aligned}$$

Thus (21) limits to (27). Also, as $Q\to 0$

$$S = \frac{\lambda\theta\rho Q\sigma^2}{(1-\rho)(1-\sigma)^2}$$

$$= \frac{\lambda\left(\frac{\theta}{Q}\right)\left(\frac{\lambda}{\mu(1-\phi)}\right)Q^2e^{-2\mu(Q-\theta)}}{\left(1-\frac{\lambda}{\mu(1-\phi)}\right)(1-e^{-\mu(Q-\theta)})^2}$$

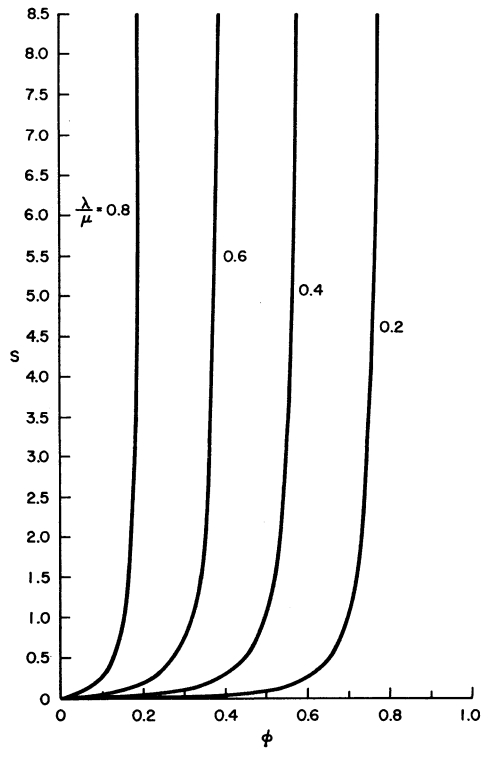

Fig. 4. Average swap time S as a function of percentage swap time ϕ with λ/μ as parameter for Example 2 ($\mu=3$).

$$= \frac{\lambda^2\phi Q^2}{[\mu(1-\phi)-\lambda][1-1+\mu(Q-\theta)-\cdots]^2}$$

$$= \frac{\lambda^2\phi}{[\mu(1-\phi)-\lambda][\mu(1-\phi)]^2}$$

$$= \frac{\rho^2\phi}{\mu(1-\phi)(1-\rho)}.$$

Thus (23) limits to (26).

Furthermore, if we wish to find the average swap time S_T for *all* customers still in system (queues plus service), we merely replace $\overline{N}=\rho^2/(1-\rho)$ with $\overline{N}_{\text{TOTAL}}=\rho/(1-\rho)$ which then gives $S_T=\rho\phi/[\mu(1-\rho)(1-\phi)]$.

Example 3: Here we consider the priority processor-shared case studied in [4]. We have $g_{pn}=g_p$, $B_p(t)=1-e^{-\mu_p t}$, and Poisson arrivals at rate λ_p for group p, $p=1, 2, \cdots, P$. Allowing swap time of form $\theta_{pn}=\theta_p$, we recognize that we must modify the service time distribution to take the form $B_p(t)=1-e^{-\mu_p(1-\phi_p)t}$. From [4] we get

$$W_p(\tau) = \frac{\tau}{g_p(1-\rho)}\sum_{i=1}^{P} g_i\rho_i \tag{28}$$

where

$$\rho_p = \lambda_p/\mu_p(1-\phi_p). \tag{29}$$

Applying Corollary 1 of Theorem 2 we get

$$S = \sum_{p=1}^{P}\frac{\lambda_p\phi_p}{g_p(1-\rho)}\int_0^\infty \tau e^{-\mu_p(1-\phi_p)\tau}\left(\sum_{i=1}^{P} g_i\rho_i\right)d\tau$$

$$S = \sum_{p=1}^{P} \frac{\rho_p \phi_p \sum_{i=1}^{P} g_i \rho_i}{g_p(1-\rho)\mu_p(1-\phi_p)} \tag{30}$$

where

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

In order to plot S, we must choose values for $\{g_p\}$, $\{\mu_p\}$, $\{\lambda_p\}$, and $\{\phi_p\}$. For the case $\lambda_p = \lambda/P$, $\mu_p = \mu$, $\phi_p = \alpha/g_p$ (where $0 \leq \alpha \leq 1$) one obtains from (30),

$$S = \left(\frac{\alpha\lambda^2}{\mu^3 P^2}\right) \frac{\sum_{p=1}^{P} \frac{1}{(g_p - \alpha)^2} \sum_{i=1}^{P} \frac{g_i^2}{g_i - \alpha}}{1 - \frac{\lambda}{\mu P} \sum_{j=1}^{P} \frac{g_j}{g_j - \alpha}}. \tag{31}$$

This last is plotted in Fig. 5 versus α with $\mu = 3$, $\lambda = 1$, $P = 5$, and for the following eight cases: $g_p = 1$, $g_p = (1.01)^p$, $g_p = (1.1)^p$, $g_p = \log_2 (p+1)$, $g_p = p$, $g_p = p^{3/2}$, $g_p = p^2$, $g_p = p^{5/2}$. Note the interesting effect where S decreases as we progress through the first four cases and then increases as we progress through the last four cases. These cases have been arranged in order of increasing discrimination between classes.

We note here only one of other additional methods for obtaining S. From [4] we have that the expected number N_p of customers in the system of queues from group p is

$$N_p = \frac{\rho_p}{g_p(1-\rho)} \sum_{i=1}^{P} g_i \rho_i. \tag{32}$$

Also, for a job of average length $(1/\mu_p(1-\phi_p))$, we spend $\phi_p/\mu_p(1-\phi_p)$ seconds swapping. Thus we must have

$$S_p = \frac{\phi_p \rho_p \sum_{i=1}^{P} g_i \rho_i}{\mu_p(1-\phi_p) g_p (1-\rho)}. \tag{33}$$

Since $S = \sum_{p=1}^{P} S_p$, we sum and obtain (30).

Example 4: For our last example, we consider the finite-population case with M consoles, exponential service with mean $1/\mu$ and exponentially distributed think-time with mean $1/\gamma$ as studied in [1], [7], and [10]. From the curves given by Adiri and Avi-Itzhak [1], we may approximate $T(\tau)$ (the average total response time) by a linear function of τ. We take this as the solution form for $T(\tau)$ in the processor-shared case $(Q \to 0)$. Thus we have (for zero swap time, $\phi = 0$)

$$T(\tau) \cong K\tau \tag{34}$$

where K is some constant. To solve for this constant, we use the well-known result for T = average (over τ) of $T(\tau)$ (see [7]).

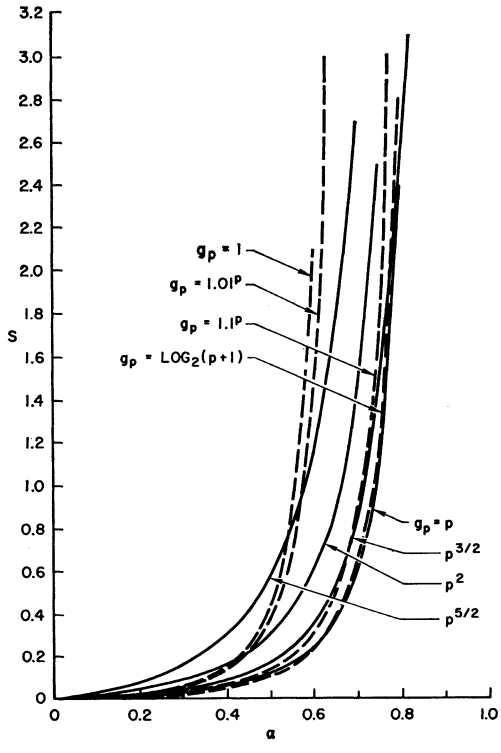

Fig. 5. Average swap time as a function of swap loss for the priority processor-shared system.

$$T = \frac{M/\mu}{1-\pi_0} - \frac{1}{\gamma} \tag{35}$$

where

$$\pi_0 = \left[\sum_{m=0}^{M} \frac{M!}{(M-m)!} \left(\frac{\gamma}{\mu}\right)^m\right]^{-1}. \tag{36}$$

But

$$T = \int_0^\infty T(\tau)\, dB(\tau) \tag{37}$$

where

$$B(\tau) = 1 - e^{-\mu\tau}.$$

From (34) and (37) we obtain

$$K = \mu T.$$

Thus, for the zero swap-time case

$$T(\tau) \cong \mu T \tau = \left[\frac{M}{1-\pi_0} - \frac{\mu}{\gamma}\right]\tau. \tag{38}$$

Now for $\phi > 0$, we merely use $\mu(1-\phi)$ rather than μ to obtain

$$T(\tau) \cong \left[\frac{M}{1-\pi_0} - \frac{\mu(1-\phi)}{\gamma}\right]\tau \tag{39}$$

where

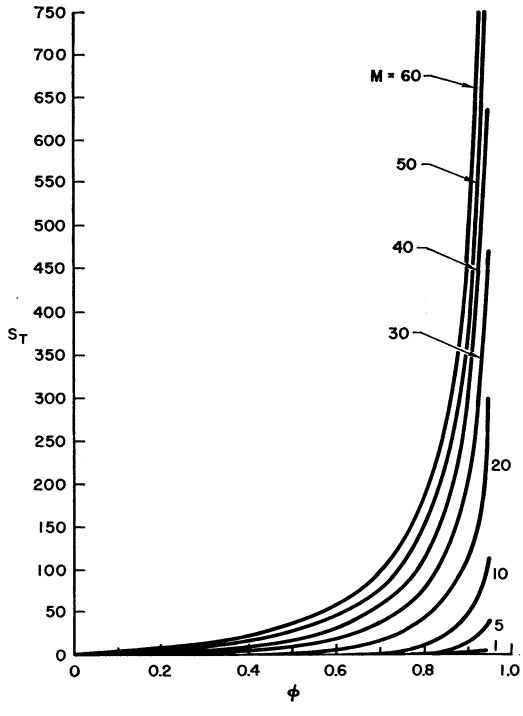

Fig. 6. Average swap time as a function of percentage swap time for the finite console processor-shared case.

$$\pi_0 = \left[\sum_{m=0}^{M} \frac{M!}{(M-m)!} \left[\frac{\gamma}{\mu(1-\phi)} \right]^m \right]^{-1}. \quad (40)$$

Applying Corollary 2 of Theorem 2, we obtain for the average system (queue plus service) swap time S_T,

$$S_T \cong \lambda\phi \int_0^\infty \tau e^{-\mu(1-\phi)\tau} \mu(1-\phi) T d\tau$$

where λ, the average input and output rate, is clearly $\lambda = \mu(1-\phi)(1-\pi_0)$. Thus

$$S_T \cong (1-\pi_0)\phi T$$

$$S_T \cong \phi \left[\frac{M}{\mu(1-\phi)} - \frac{(1-\pi_0)}{\gamma} \right]$$

where π_0 is given by (40). S_T is plotted versus ϕ for various M in Fig. 6 for $1/\mu = 0.88$ and $1/\gamma = 35.2$.

Conclusion

We have shown how to solve for the expected swap time expended on all customers in the system of queues. This we have done for the time-sharing systems ($Q>0$) in Theorem 1 and for the processor-sharing systems ($Q \to 0$) in Theorem 2. The examples given show the ease of obtaining results for the processor-sharing case.

References

[1] I. Adiri and B. Avi-Itzhak, "A time-sharing queue with a finite number of customers," *J. ACM*, vol. 16, no. 2, pp. 315–323, April 1969.
[2] E. G. Coffman and L. Kleinrock, "Feedback queueing models for time-shared systems," *J. ACM*, vol. 15, no. 4, pp. 549–576, October 1968.
[3] L. Kleinrock, "On swap time in time-shared systems," *1969 Proc. IEEE Computer Group Conf.* (Minneapolis, Minn.), pp. 37–41.
[4] ——, "Time-shared systems: A theoretical treatment," *J. ACM*, vol. 14, no. 2, pp. 242–261, April 1967.
[5] ——, "Analysis of a time-shared processor," *Naval Res. Logistics Quart.*, vol. 11, no. 10, pp. 59–73, March 1964.
[6] L. Kleinrock and E. G. Coffman, "Distribution of attained service in time-shared systems," *J. Computer Sys. Sci.*, vol. 1, no. 3, pp. 287–298, October 1967.
[7] L. Kleinrock, "Certain analytic results for time-shared processors," *1968 Proc. IFIP Cong.* (Edinburgh, Soctland), pp. D119–D125.
[8] B. Krishnamoorthi and R. C. Wood, "Time-shared computer operations with both interarrival and service times exponential," *J. ACM*, vol. 13, no. 3, pp. 317–338, July 1966.
[9] J. M. McKinney, "A survey of analytical time-sharing models," *Computing Surveys*, vol. 1, no. 2, pp. 105–116, June 1969.
[10] A. L. Scherr, *An Analysis of Time-Shared Computer Systems*. Cambridge, Mass.: M.I.T. Press, 1967.
[11] L. E. Schrage, "The queue M/G/1 with feedback to lower priority queues," *Management Sci.*, ser. A., vol. 13, pp. 466–474, 1967.