Reprinted from the 1966 IEEE Region Six CONFERENCE RECORD

THEORY OF QUEUES APPLIED TO TIME-SHARED COMPUTER SYSTEMS[*]

Leonard Kleinrock
Department of Engineering
University of California, Los Angeles

ABSTRACT

Time-shared computer (or processing) facilities are treated as stochastic queueing systems under priority service disciplines and the performance measure of these systems is taken to be the average time spent in the system. Results are presented for models in which time-shared computer usage is obtained by giving each request a fixed quantum, Q, of time on the processor, after which the request is placed at the end of a queue of other requests; the queue of requests is constantly cycled, giving each user Q sec on the machine per cycle. Results for the case for which Q → 0 (a processor-shared model) are then presented. A general time-shared facility is then considered in which priority groups are introduced. Specifically, the p^{th} priority group is given g_pQ seconds in the processor each time around. Letting Q → 0 we then get results for priority processor-shared system. These disciplines are compared to the first come first served disciplines. The systems considered provide the two basic features desired in any time-shared system, namely, rapid service for short jobs, and the virtual appearance of a (fractional capacity) processor available on a full-time basis.

No charge is made for swap time, thus providing results for "ideal" systems. The results hold only for Poisson arrivals and geometric (or exponential) service time distributions.

I INTRODUCTION

Interest in time-shared computing systems has been growing at an increasing rate in recent years. A number of such systems have been cropping up in various places throughout the country (see References [1]-[5]). The motivation for such interest is toward encouraging the interaction between the user (programmer) and the computer itself. Furthermore, it is recognized that the availability of computers must be increased so rapidly that we may soon find it expedient to offer computational and processing capacity as a "public utility." A natural way to do this is to provide the public with access to computers on a time-shared basis (not unlike the telephone company's use of graded trunk lines), thus providing a high efficiency for the user as well as for the computer facility.

Time-shared systems are often designed with the intent of appearing to a user as his personal processor (where, ideally, he is unaware of the presence of any other users). Of course, no such ideal system can guarantee a full-capacity full-time machine to any user (in the time-shared mode), but rather, they offer a fractional-capacity "full-time" machine to each user. In the ideal case, at any time, the fraction of the total capacity offered to any user will just be[†] the inverse of the number of users currently requesting service (i.e., we assume an harmonic variation of individual capacity with number of users).

Unfortunately, very little work has been carried out in analyzing the behavior of time-shared systems from a mathematical viewpoint. In this paper we proceed in that direction.

In Section II, we define three models of time-shared systems.

II QUEUEING MODELS OF TIME-SHARED FACILITIES

The Round-Robin Model

Our point of departure is the discrete time model of a time-shared processor studied by Kleinrock [6]. In this model, it is assumed that time is quantized with segments each Q seconds in length. At the end of each time interval, a new unit (or job) 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 (i.e., the required processing time) of a newly arriving unit is chosen independently from a geometric distribution such that for $0 \leq \sigma <1$

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

where s_n is the probability that a unit's service time is exactly n time intervals long (i.e., that its service time is nQ seconds).

[*] Research sponsored by Applied Mathematic Division, Office of Research, United States Air Force Grant No. AFOSR-700-65.

[†] This is generalized in our priority model described in Section II.

The procedure for servicing is as follows: a newly arriving unit joins the end of the queue, and waits in 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 one unit of service upon it (i.e., it services this job for exactly Q seconds). At the end of this time interval, the unit leaves the system if its service (processing) is finished; if not, it joins the end of the queue with its service partially completed, as shown in Figure 1. Obviously, a unit whose processing requirement is nQ time units long will be forced to join the queue n times in all before its service is completed.

The Processor-Shared Model (no priorities)

If we assume zero swap-time, we may consider the case of a round-robin system in which Q → 0. We must be careful in taking this limit since the service time, nQ also goes to zero in this case and our model loses all meaning. Consequently, let us agree to keep the average service time constant as Q → 0. This involves changing σ, the decay rate in Equation (1) such that $\sigma \to 1$ as $Q \to 0$. Specifically, we have that

$$n = \sum_{n=1}^{\infty} n\, s_n = \frac{1}{1-\sigma}$$

and, defining

$$\frac{1}{\mu C} = \text{average service requirement (in seconds)}$$

we get

$$\frac{1}{\mu C} = \frac{Q}{1-\sigma} = \text{constant as } Q \to 0 \text{ and } \sigma \to 1$$

or

$$\sigma = 1 - \mu C Q \tag{2}$$

Thus, the limiting operation we consider is where $Q \to 0$ and $\sigma \to 1$ in the manner expressed in Equation (2). The result of this limit is that the required service ℓ, (in operations) is exponentially distributed with parameter μ, viz.,

$$p(\ell) = \mu e^{-\mu \ell} \tag{3}$$

where ℓ is the length of the job.

We have chosen to assume that the length ℓ of a job is given in number of operations instead of in seconds, thus making the user requirement independent of the machine on which it is serviced. We then define, for any processor, a quantity

C = capacity of a processor in operations (say additions) per second.

The service time for a job then becomes ℓ/C seconds, with a mean service time of $1/\mu C$ seconds.

The arrival mechanism in the limit then becomes Poisson with an average arrival rate of λ customers per second. This model reduces to a system in which a user is processed at a rate C/k operations per second when there are k users sharing a computer of capacity C. This processing rate varies as new users enter and old ones leave the system. We are here assuming an harmonic variation of individual processing rate with number of customers. See Figure 2.

The Priority Processor-Shared Model

This is a generalization of the processor-shared system considered above. Here, we assume that the input traffic is broken up into P separate priority groups, where p^{th} group has a Bernoulli arrival pattern at an average rate of λ_p customers per second and a geometrically distributed service requirement whose mean is $1/(1-\sigma_p)$ operations. For the Q → 0 case, we give a member of the p^{th} priority group $g_p Q$ seconds of service each time he cycles around the queue (see Figure 3).

For Q → 0 (holding fixed $1/\mu_p C = Q/(1-\sigma_p)$ this model then reduces to a processor-shared model with a priority structure wherein a member from group p receives at time t a fraction f_p, where

$$f_p = \frac{g_p}{\sum_{i=1}^{P} g_i n_i} \tag{4}$$

of the total processing capacity C (here n_i is the number of customers from priority group i present in the system at time t). We note that we then have, for the p^{th} group, Poisson arrivals (λ_p per second) and exponential service with an average of $1/\mu_p C$ seconds. The non-priority processor-shared model considered earlier is the special case $g_p = 1$ for all p.

The interest of this model is to give preferential service to certain of the groups of users, where, for convenience, we may consider that the higher the value of p, the higher is considered the priority of that group. In such a case, we may assume that g_p is a monotonically increasing function of p (although we do not need this for the subsequent development).

In Figure 4 we show a diagram of the priority processor-shared system.

We observe that the two processor-shared models are ideal in the sense that swap-time is assumed to be zero and in that customers are given immediate use of the processor (albeit only a fractional capacity $f_p C$).

III RESULTS FOR TIME-SHARED SYSTEMS†

The Round-Robin System

This system has already been studied (see [6]). We present the results of that analysis here.

THEOREM 1: The expected value, T_n, of the total time‡ spent in the round-robin system for a job whose service time is nQ seconds, is

$$T_n = \frac{nQ}{1-\rho} - \frac{\lambda Q^2}{1-\rho}\left[1 + \frac{(1-\sigma\alpha)(1-\alpha^{n-1})}{(1-\sigma)^2(1-\rho)}\right] \tag{5}$$

where

$$\alpha = \sigma + \lambda Q \tag{6}$$

† Proofs for Theorems 1 and 2 may be found in Reference [6]. Proofs for the remaining theorems will be published shortly by the author.

‡ T_n is the sum of the time spent in the queue and the time spent in the service facility.

$$\rho = \frac{\lambda Q}{1-\sigma} \tag{7}$$

Furthermore, the expected number, E_r, of customers in the system is given by

$$E_r = \frac{\rho\sigma}{1-\rho} \tag{8}$$

THEOREM 2: The expected value, T_n', of the total time spent in the strict first-come first-served system[†] for a unit whose service time is nQ seconds is

$$T_n' = \frac{QE_r}{1-\sigma} + nQ \tag{9}$$

where E_r is defined in Equation (8).

In Reference [6] it is shown that a good approximation to T_n is

$$T_n = nQE_r + nQ \tag{10}$$

When we compare Equations (9) and (10), we see that for units which require a number of service intervals less (greater) than $1/(1-\sigma)$, the round-robin waiting time is less (greater) than the strict first-come first-served system. One notes, however, that the average number of service intervals, $\bar{n}$, is exactly $1/(1-\sigma)$. Thus, for this approximate solution, the crossover point for waiting time is at the mean number of service intervals. This effect is observable in Figures 5-7 in Section IV.

The Processor Shared System

This model considers the limit of the round-robin model in which $Q \to 0$ and $\sigma = 1 - \mu CQ$, giving a Poisson arrival mechanism with an average of λ units arriving per second and an exponential service distribution with an average of $1/\mu$ operations per customer. We have the following:

THEOREM 3: The expected value $T(\ell)$ of the total time spent in the processor-shared system for a customer requiring ℓ operations, is

$$T(\ell) = \frac{\ell/C}{1-\rho} \tag{11}$$

where

$$\rho = \lambda/\mu C \tag{12}$$

C = capacity of the processor in operations per second.

The expected number, E, of customers in the system is

$$E = \frac{\rho}{1-\rho} \tag{13}$$

In Section IV we compare these results with that of the round-robin model.

The Priority Processor-Shared System

In this system, we have P priority groups with Poisson arrivals at an average rate of λ_p per second and an exponentially distributed service requirement with a mean of $1/\mu_p$ operations $(p = 1, 2, \ldots, P)$. For a processor of capacity C operations per second, we assign a customer from the p^{th} priority group a capacity f_pC when there are n_i type i customers in the system; f_p is given by Equation (4), viz.,

$$f_p = \frac{g_p}{\sum_{i=1}^{P} g_i n_i} \tag{4}$$

For such a system, we have the following Theorem.

THEOREM 4: The expected value $T_p(\ell)$ of the total time spent in the priority processor-shared system for a customer from priority group p who requires ℓ operations is

$$T_p(\ell) = \frac{\ell}{C}\left[1 + \sum_{i=1}^{P} \frac{g_i\rho_i}{g_p(1-\rho)}\right] \tag{14}$$

the expected number, E_p, of type p customers in the system is

$$E_p = \frac{\rho_p}{1-\rho}\left[1 + \sum_{i=1}^{P}\left(\frac{g_i}{g_p} - 1\right)\rho_i\right] \tag{15}$$

where

$$\rho_p = \frac{\lambda_p}{\mu_p C}$$

and

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

and where $g_p > 0 \quad p = 1, 2, \ldots, P$.

In the following section, we compare this priority processor-shared model to the other two models studied. For completeness, we also consider a strict first-come first-served system with the same input and service requirements as in our priority model. To this end, we have

THEOREM 5: The first-come first-served system with a priority input yields, for customers with ℓ required operations, a total expected time in system as follows

$$T(\ell) = \frac{\ell}{C} + \frac{\rho/\mu C}{1-\rho} \tag{16}$$

where

$$\frac{1}{\mu C} = \frac{\rho}{\sum_{p=1}^{P} \lambda_p} \tag{17}$$

We note that, for P = 1, we have the (non-priority) processor-shared system.

IV DISCUSSION, EXAMPLES, AND COMPARISON OF THE SYSTEMS

Having considered three models of time-shared systems, we now wish to compare their performance among themselves as well as with the first served systems. The basis of comparison will be the average conditional

[†] This is our reference system and corresponds to the more usual case where a unit receives its complete processing requirement the first time it enters service.

additional delay experienced a customer (conditioned on his required processing as well as on his priority). We define the additional delay as the difference between the time such a customer spends in the time-shared system and the time he would spend in the system if no other customers were present (in a first-come first-served model, this is merely his time in queue), i.e., let

$W_p(\ell)$ = the average additional delay experienced by a customer from priority group p who requires ℓ operations in service (obvious analogous definition for $W_p(n)$ and $W(n)$ in the $Q > 0$ case).

We have[†]

$$W_p(\ell) = T_p(\ell) - \ell/C \tag{18}$$

In the most general model, we wish to display curves of $W_p(\ell)$ as a function of ℓ and as a function of ρ with p as a parameter. Furthermore we choose to plot

$$\frac{1-\sigma_p}{\sigma_p Q} \; W_p(n)$$

rather than $W_p(n)$ for purposes of a convenient normalization, which, in the case for $Q \to 0$ becomes $\mu_p C\, W_p(\ell)$. Below we present these curves for various examples.

The Round-Robin System

In Figures 5-7, curves[‡] of $(1-\sigma)/(\sigma Q)\, W_n \equiv k\, W_n$ are plotted to show the general behavior of the round-robin structure for the late arrival system. On each graph, (circled) points corresponding to the first-come first-served case have also been included. The normalization $(1-\sigma)/(\sigma Q)$ used is such that for the first-come first-served case, we obtain the curve $\rho/(1-\rho)$ which is a function only of ρ.

Figures 5-7 indicate the accuracy of the approximation discussed above in which the crossover point for waiting times is at the mean number of service intervals, $1/(1-\sigma)$. In Figures 5 and 6 there is no noticeable difference (on the scale used) between the first-come first-served points, and the curve for $n = 1/(1-\sigma)$; moreover, in Figure 7 the points fall between the curves for $n = 1$ and $n = 2$, since $1/(1-\sigma) = 1.25$.

In Figure 8, we plot kW_n as a function of n for $\rho = 1/2$, $\sigma = 4/5$. In all these curves (Figures 5-8) we observe that by introducing the round-robin system, one manipulates the relative waiting time for different jobs and thus imposes a method of time-sharing which gives preferential treatment to short jobs.

The Processor Shared System

In Figures 9 and 10 below, we plot $\mu CW(\ell)$ as a function at ρ (for various $\mu\ell$) and as a function[‡] of $\mu\ell$ (for various ρ) respectively.

In Figure 10, the circles indicate the values of $\mu CW(\ell)$ for the strict first-come first-served system (see Theorem 5). Again we see the preferential treatment given to shorter jobs, and again we see that the "break-even" point for jobs is the average job length ($\mu\ell = 1$).

The Priority Processor-Shared Model

For these curves, we let $\mu_p = \mu$, $\lambda_p = \lambda/P$, $P = 5$ for $p = 1, 2, \ldots, 5$. In Figures 11-13 we show $\mu CW_p(\ell)$ as a function of ρ for various p and for $\mu\ell = 1$. Figure 11 is for $g_p = p^2$; Figure 12 is for $g_p = p$; and Figure 13 is for $g_p = \log_2(p+1)$. In each of these figures, the circles correspond to the strict first-come first-served system (which compares the treatment as a function of p for the two systems).

In Figures 14-16 we show $\mu CW_p(\ell)$ as a function of $\mu\ell$ for various p and for $\rho = 1/2$. Again $g_p = p^2$, $g_p = p$ and $g_p = \log_2(p+1)$ for Figures 14, 15 and 16 respectively. In each of these figures, the circles correspond to the behavior of a first-come first-served system (on these axes, it is a constant additional delay, independent of $\mu\ell$).

In both processor-shared models, $W_p(\ell)$ approaches zero as $\rho \to 0$ for all ℓ and p.

In all of the curves presented, we see that the effect of introducing a time-sharing discipline is to reduce the average waiting time for customers with "short" service (processing) requirements at the expense of those customers with "longer" service requirements. For the non-priority cases (i.e., the first two models studied) we observe that customers with service (processing) requirements less (greater) than the average requirement spend, on the average, less (greater) time in the system, compared to a strict first-come first-served system.

In the priority processor-shared system, we see a similar trend (i.e., short jobs wait less than long jobs) and in addition, we give preferential treatment (shorter waiting) to certain select high priority groups. The effect now is that for job lengths below some critical value (dependent upon p, the priority group) a customer does better (waits less) in the time-shared system than in a first-come first-served system. This critical length is monotonically increasing with p. The degree and manner in which the different priority groups receive treatment depends upon the function g_p and may be varied over a considerable range of relative performance.

CONCLUSION

In this paper, we have considered serveral models of time-shared processing systems. These models provide the basic features desired in such systems, namely, rapid service for short jobs, and the virtual appearance of a (fractional capacity) processor available on a full-time basis.

The most general model, the priority processor-shared system, not only provides the above features, but also allows the population of customers to be divided into priority classes where the higher priority groups receive preferential treatment compared to the lower priority groups.

The assumption of zero swap-time results in models which provide the best possible performance of such time-shared systems. Comparison of these systems with the strict first-come first-served systems showed the relative improvement (or deterioration) of performance as a

[†] Obviously, for $Q > 0$ we have $W_p(n) = T_p(n) - nQ$.

[‡] These are the same curves as in Kleinrock [6]. In these curves, ρ was varied by fixing σ and varying λQ (recall $\rho = \lambda Q/(1-\sigma)$.

[†] $\mu\ell = \dfrac{\ell}{1/\mu}$ is the length of a job normalized with respect to its average length.

function of service requirement and priority group.

REFERENCES

1. R.M. Fano, "The MAC System: The Computer Utility Approach," IEEE Spectrum, Vol. 2, No. 1, pp. 56-64, January 1965.

2. W.W. Lichtenberger and M.W. Pirtle, "A Facility For Experimentation in Man-Machine Interactions," Proc. Fall Joint Computer Conference, Vol. 27, Part I, pp. 589-598, 1965.

3. J.W. Forgie, "A Time-and Memory-Shaving Executive Program for Quick Response On-Line Applications," Proc. Fall Joint Computer Conference, Vol. 27, Part 1, pp. 599-609, 1965.

4. J. McCarthy, "Time-Sharing Computer Systems," in Management and The Computer of the Future, M. Greenberger, Ed., Cambridge, Mass., the MIT Press, pp. 221-236, 1962.

5. J.I. Schwartz, E.G. Coffman, and C. Weissman, "A General Purpose Time-Sharing System," Proc. Spring Joint Computer Conference, pp. 335-344, 1962.

6. L. Kleinrock, "Analysis of A Time-Shared Processor," Naval Research Logistics Quarterly, Vol. 11, No. 10, pp. 59-73, March 1964.

List of Figures

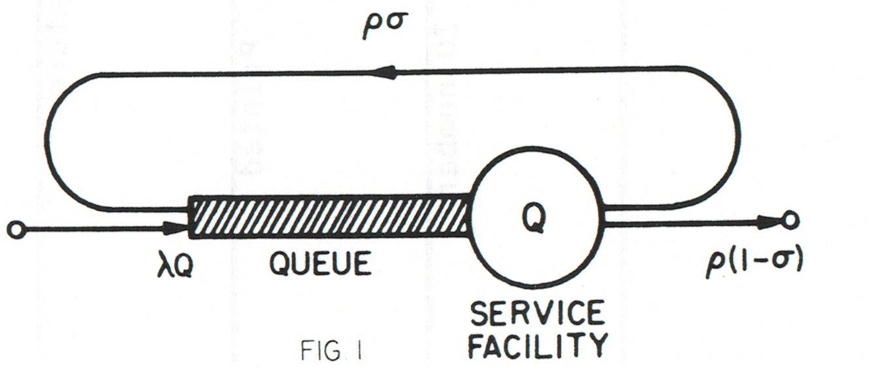

FIG 1

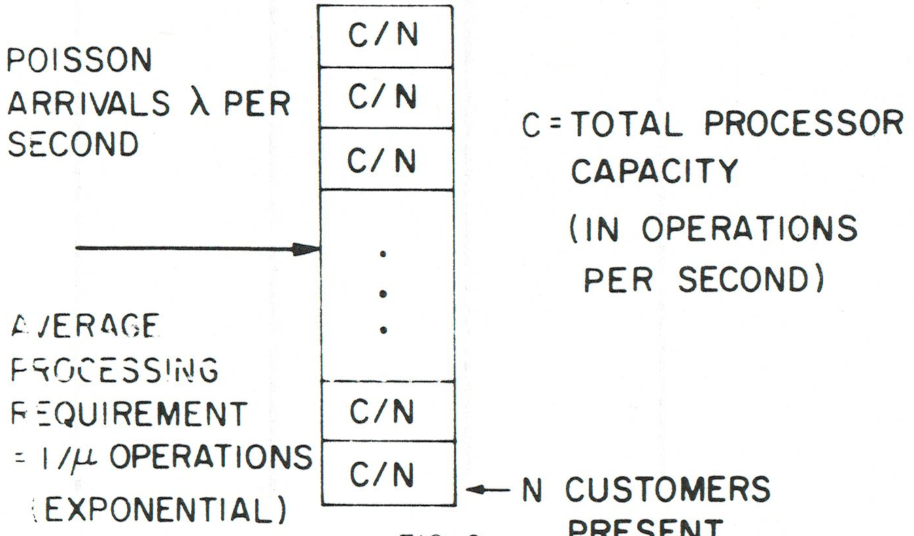

FIG. 2

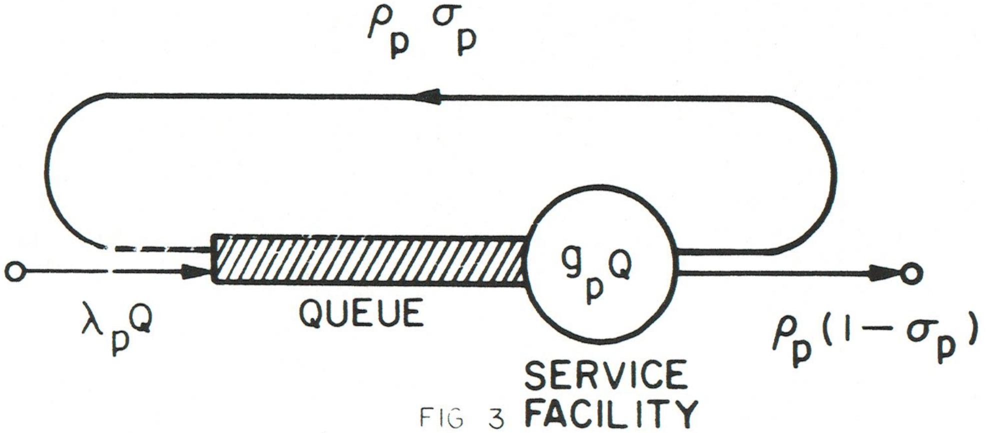

FIG 3

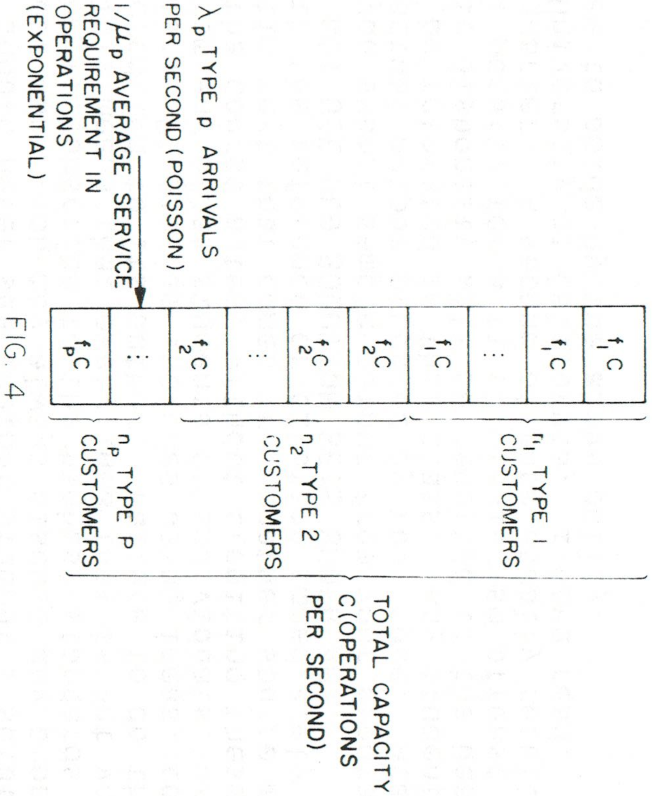

FIG. 4

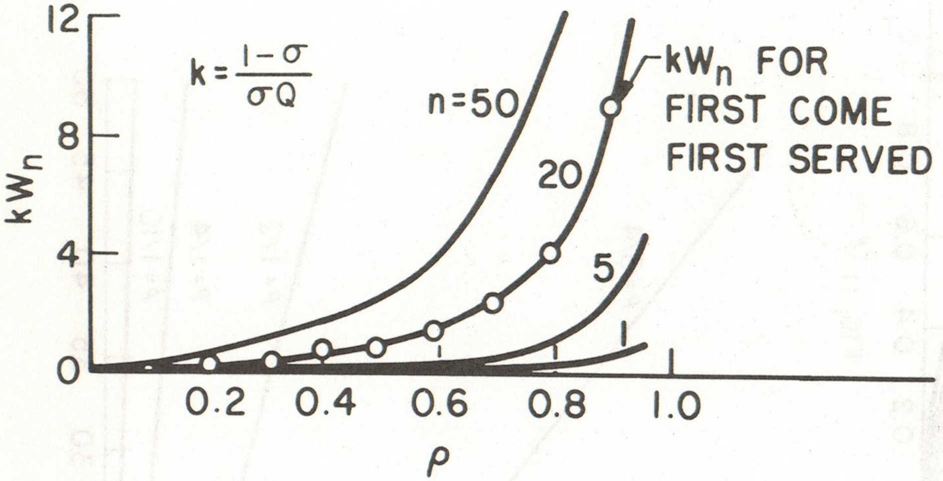

FIG. 5

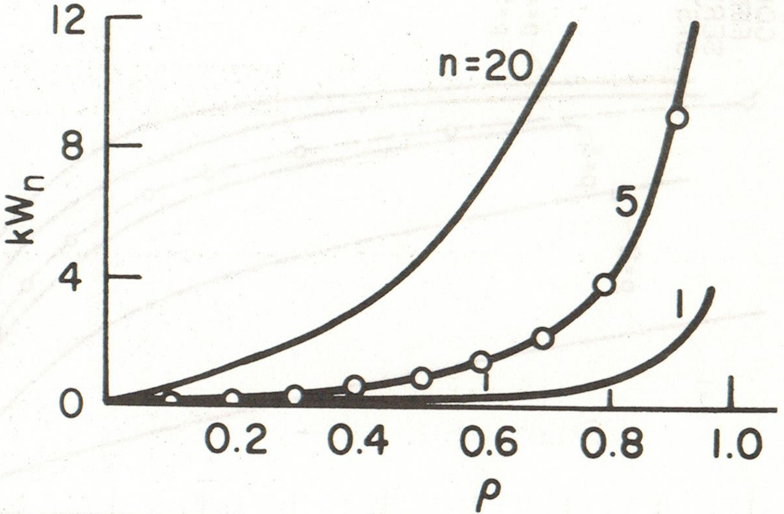

FIG. 6

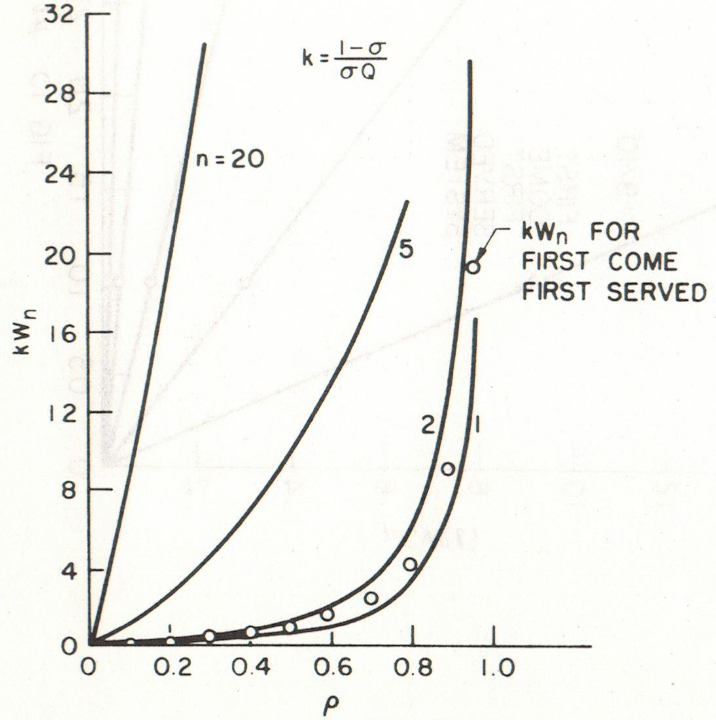

FIG. 7

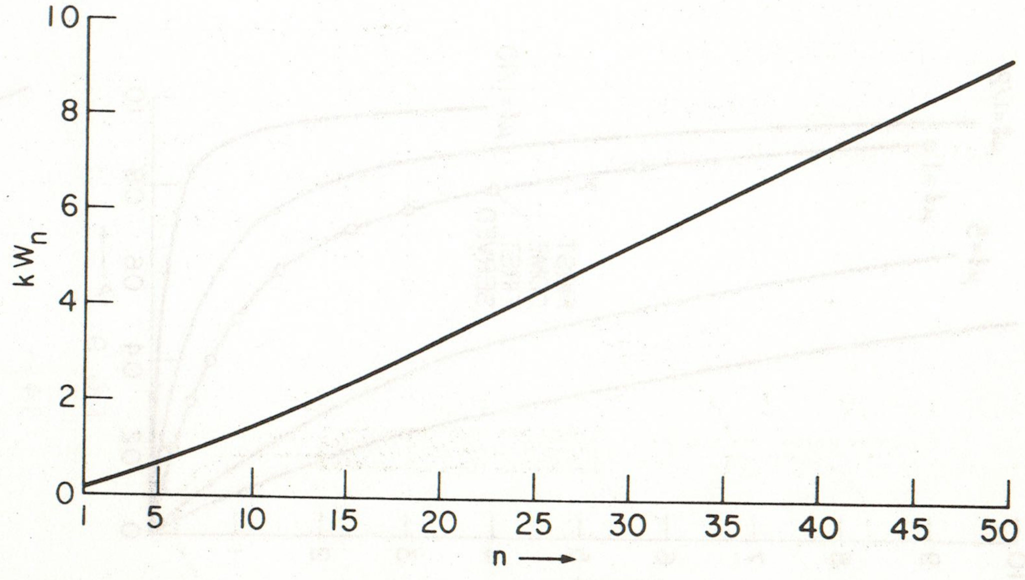

FIG. 8

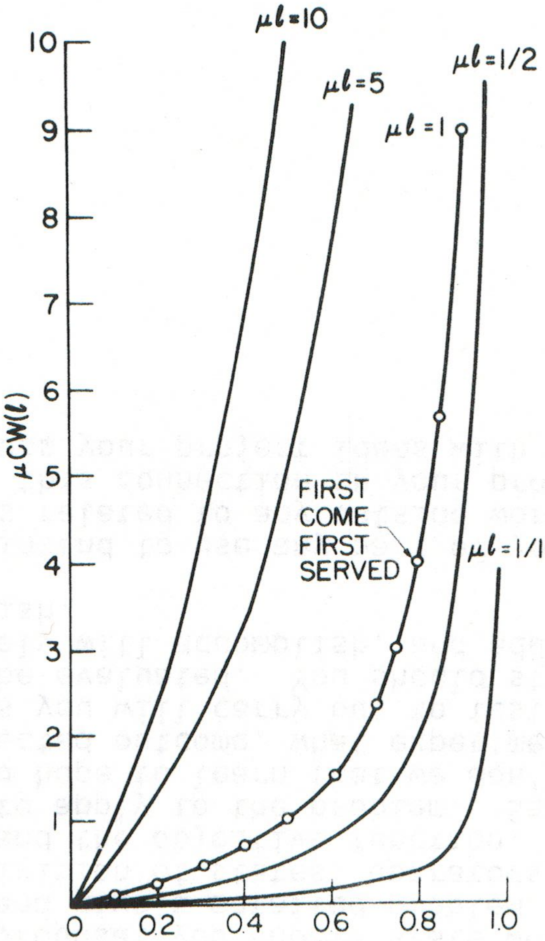

FIG. 9

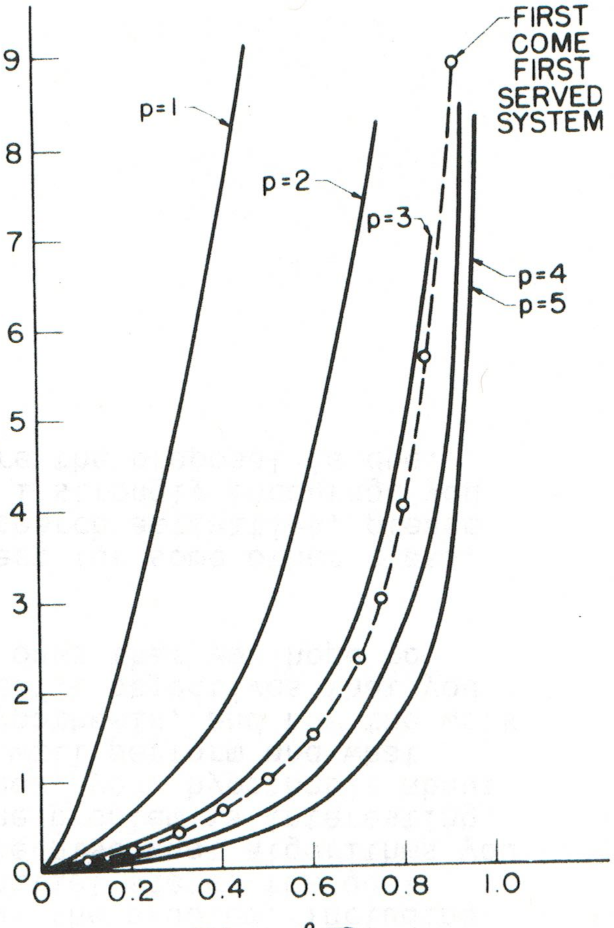

FIG. 11

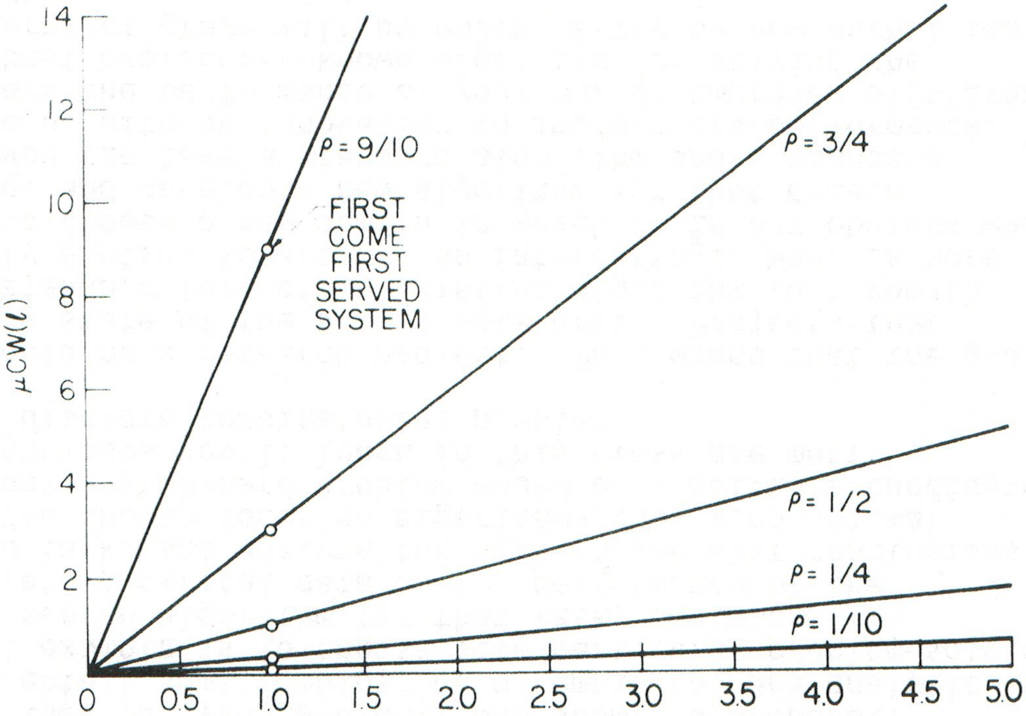

FIG. 10

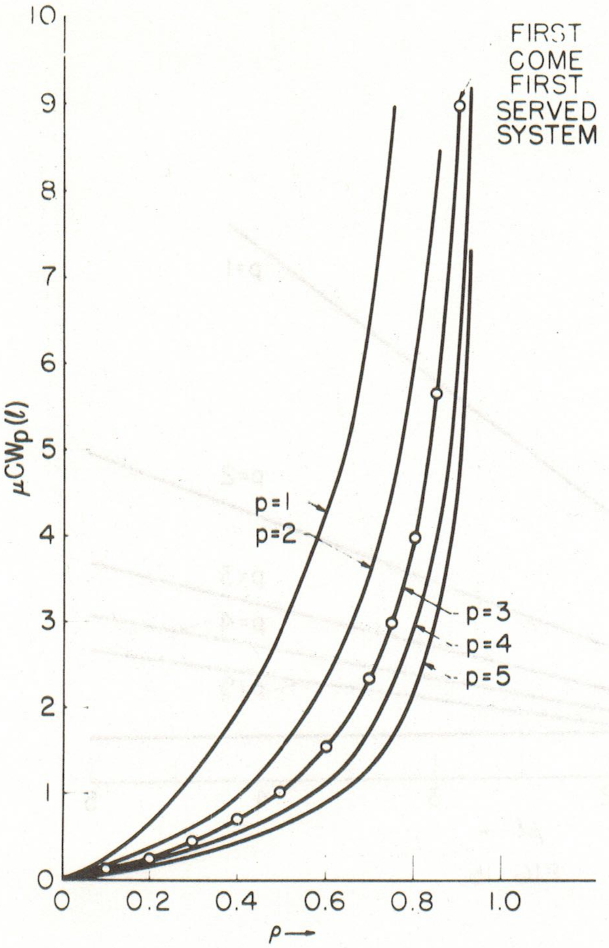

FIG. 12

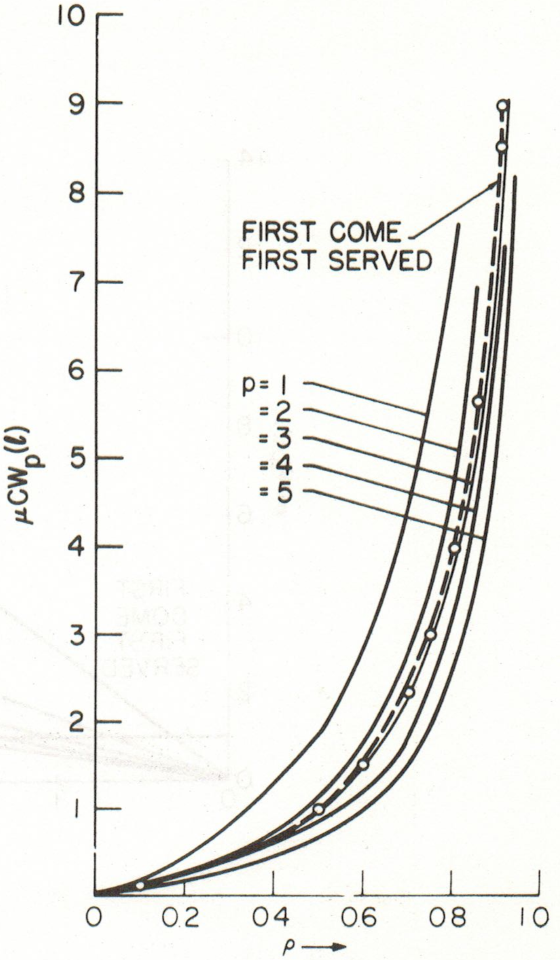

FIG. 13

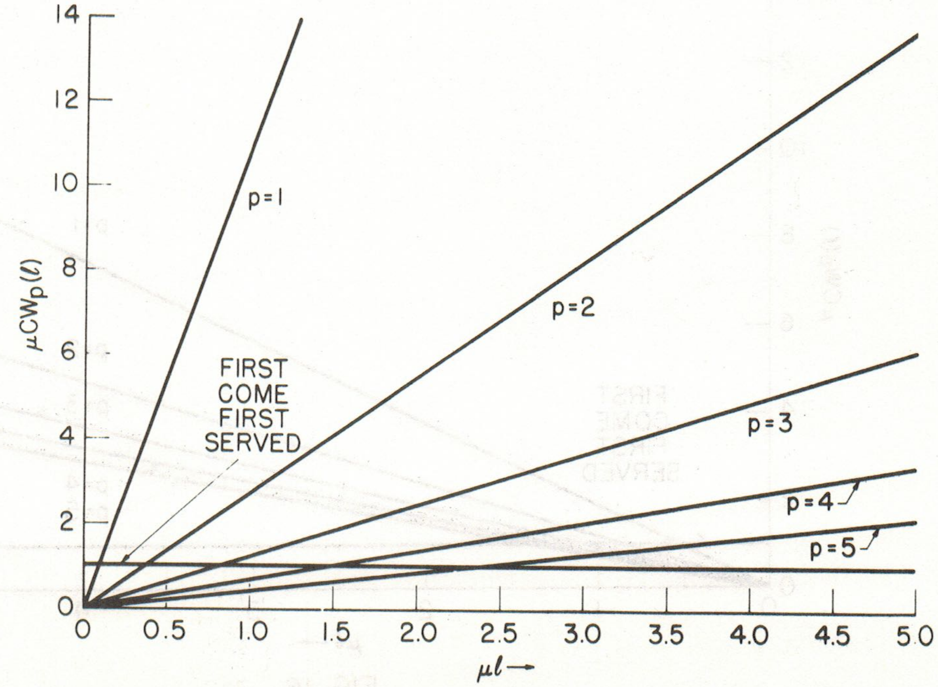

FIG. 14

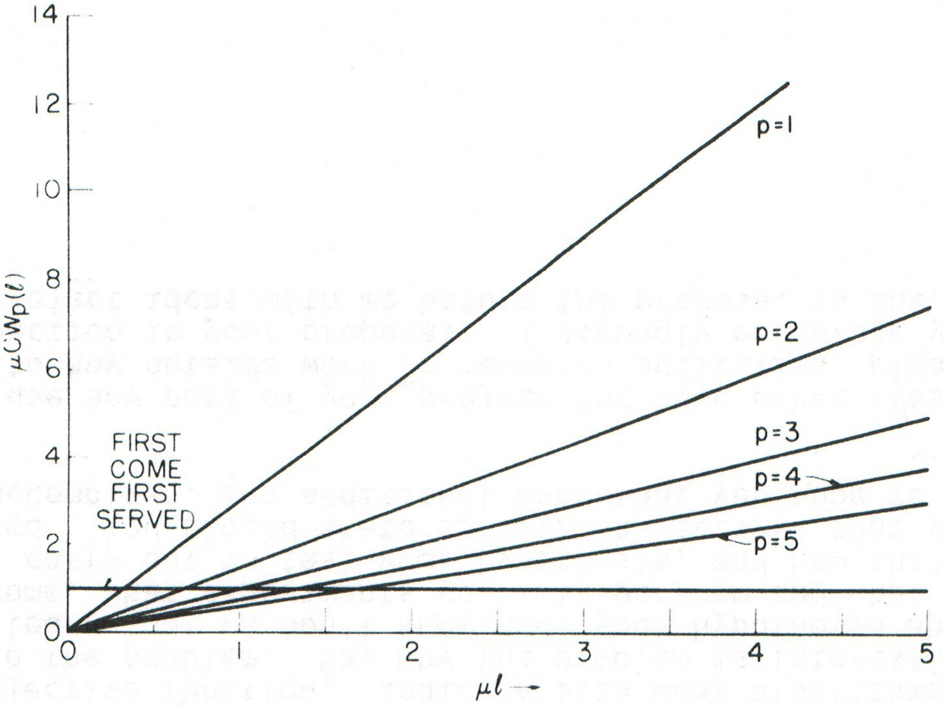
14
12
10
8
6
4
2
0
$\mu CW_p(\ell)$
0
1
2
3
4
5
$\mu \ell \rightarrow$
p=1
p=2
p=3
p=4
p=5
FIRST
COME
FIRST
SERVED

FIG. 15

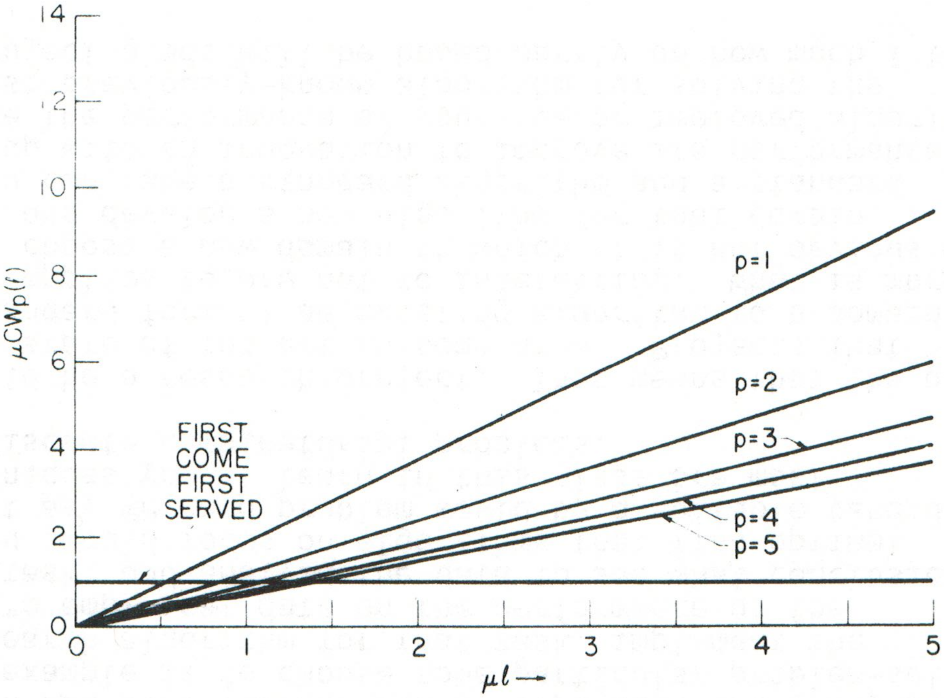
14
12
10
8
6
4
2
0
$\mu CW_p(\ell)$
0
1
2
3
4
5
$\mu \ell \rightarrow$
p=1
p=2
p=3
p=4
p=5
FIRST
COME
FIRST
SERVED

FIG. 16