#247

REPRINTED FROM

NAVAL RESEARCH LOGISTICS QUARTERLY

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

VOL. 12, NO. 2 JUNE 1965

A CONSERVATION LAW FOR A WIDE CLASS OF QUEUEING DISCIPLINES*

Leonard Kleinrock†

Department of Engineering
University of California
Los Angeles, California

ABSTRACT

A large class of queueing disciplines is defined for Poisson arrival statistics. For this class, a Conservation Law is proven which constrains the allowed variation in the average waiting times. Specifically, defining ρ_p as the product of the average arrival rate and the average service time for customers from the p^{th} priority group (where the priority system is any queue discipline included in the defined class) and W_p as their average waiting time (in queue), the Conservation Law states that $\Sigma \rho_p W_p$ is invariant over the set of queue disciplines in the class.

INTRODUCTION

When one considers the many classes of queue disciplines which have been analyzed (for example: last come first served, Wishart [10]; random service, Vaulot [9]; head of the line, Cobham [2]; delay dependent, Kleinrock [3]; etc.) and compares these to the first come first served discipline, one suspects that some measure of the average waiting time in all these systems should remain constant. In fact, it is quite reasonable to expect such an invariance based on the simple physical argument that some customers are given preferential treatment, and so need not wait as long as they would in a first come first served system; consequently, lower priority customers are forced to wait some additional time.

Indeed, we find that there is a law of conservation which holds for queueing systems subject to a large class of disciplines. For this class, the Conservation Law says that for a fixed set of arrival and service statistics, a particular weighted sum of the waiting times is a constant independent of queue discipline.

THE MODEL

A sufficient set of restrictions to define the class under consideration is as follows:

1. All customers (units) remain in the system until completely serviced (i.e., no defections);

2. There is a single service facility which is always busy if there are any units in the system;

---

*Material similar to this appears in a book by Leonard Kleinrock in the Lincoln Laboratory Publication Series entitled, Communication Nets; Stochastic Message Flow and Delay (McGraw-Hill Book Co., N.Y., 1964.

†This work was done while the author was employed at Lincoln Laboratory (operated with support from the U.S. Army, Navy, and Air Force), Massachusetts Institute of Technology, Lexington, Massachusetts.

3. Pre-emption is allowed only if the service time distributions are exponential, and the pre-emption is of the pre-emptive resume type;

4. Arrival statistics are all Poisson; service statistics are arbitrary; and arrival and service statistics are all independent of each other.

It is assumed throughout that the systems under consideration are in the steady-state equilibrium. In general, this is equivalent to requiring that the system has been operating for a long time, and that $\rho < 1$ where ρ, as usual, is the product of the average arrival rate of units and their expected service time [see (4)]; however, in some of the priority systems studied, it is possible to have $\rho \geq 1$ and still obtain a steady-state type solution for some of the higher priority units. For a full discussion of this aspect of the problem, the reader is referred to Phipps [7].

Specifically, we define a queue discipline as a system in which an entering unit is assigned a set of parameters (either at random or based on some property of the unit) which determine its relative position in the queue. This position will vary as a function of time due to the appearance of units of higher priority in the queue. At any time t, the priority of a particular unit is calculated as a function of the assigned parameters; the higher the value obtained by this function, the higher the priority. That is, the notation used is such that a unit with priority q_2 is given preferential treatment over a unit with priority q_1, where $q_2 > q_1$. Whenever a tie for the highest priority occurs, the tie is broken by a pre-determined rule (such as first come first served, random selection, and so on).

Consider a total of P different priority classes. Units from priority class p (p = 1, 2, . . ., P) arrive in a Poisson stream at rate λ_p units per second; each unit from priority class p has a total required service time selected independently from an arbitrary distribution, with mean $1/\mu_p$. We define

$$\lambda = \sum_{p=1}^{P} \lambda_p, \tag{1}$$

$$\frac{1}{\mu} = \sum_{p=1}^{P} \lambda_p/(\lambda \mu_p), \tag{2}$$

$$\rho_p = \lambda_p/\mu_p, \tag{3}$$

and

$$\rho = \lambda/\mu = \sum_{p=1}^{P} \rho_p. \tag{4}$$

We further define

W_p = expected value of the time spent in the queue for a unit with assigned parameter p.

THE CONSERVATION LAW

THEOREM 1:* For any queue discipline and any given arrival and service time parameters subject to restrictions 1-4 above,

$$\sum_{p=1}^{P} \rho_p W_p = \text{constant with respect to variation of the queue discipline.} \tag{5}$$

where P represents the total number of groups to be distinguished in the traffic. In particular,

$$\sum_{p=1}^{P} \rho_p W_p = \begin{cases} [\rho/(1-\rho)]V_1 & 0 \le \rho < 1 \\ \infty & \rho \ge 1 \end{cases}, \tag{6}$$

where

$$V_1 = (1/2) \sum_{p=1}^{P} \lambda_p E\left(t_p^2\right), \tag{7}$$

and $E\left(t_p^2\right)$ = second moment of service time distribution for group p.

V_1 may be interpreted as the expected time required to complete service on the unit found in service upon entry, for a first come first served system. That is, convert the system at hand to one in which the same arrival and service time distributions apply, but where the entire priority and pre-emptive structure is removed and the system therefore operates on a first come first served basis. Thus, V_1 is itself independent of the particular queue discipline chosen.

CONCLUSIONS

Note that the Conservation Law constrains the allowed variation in the W_p for any discipline within the wide class considered. If we form the sum

$$\sum_{p=1}^{P} (\lambda_p/\lambda) W_p \tag{8}$$

(which weights the expected waiting time of the p^{th} priority group by its relative arrival rate λ_p/λ), the Conservation Law says that this sum is a constant in the case where all μ_p are equal. This sum (if multiplied by λ) represents the average number of units in the queue (see the appendix). If we form the time-averaged waiting time†

*See the appendix for proof of this theorem. Along with the proof, we state and prove two related corollaries.

†Physically, we may think of this average as the following. Let us sample the system at random points in time; each time we sample, we record the time spent in the queue by the unit currently being serviced. The average value of this set of numbers is the average we are referring to.

$$(9) \qquad \sum_{p=1}^{P} (\lambda_p/\lambda\mu_p) W_p$$

(which weights the W_p not only by λ_p/λ, but also by $1/\mu_p$, the average service time of a p type unit), then the Conservation Law says that this average is a constant.

In conclusion, we state that the Conservation Law probably holds for a more inclusive class of queue disciplines than that described by restrictions 1-4. Indeed certain disciplines with non-Poisson arrival distributions have been investigated (see for example, Kleinrock [4]) and have been shown to obey the Conservation Law.

APPENDIX

We make extensive use of a well-known result in queueing theory in this appendix. The result was conjectured by many researchers, and recently, a formal proof of its validity was published by Little [6]. Roughly stated, the result says that the expected number, E(n), of units in a queueing system which has reached equilibrium, is equal to the product of input rate, λ of these units to the system, and the expected value, τ, of the time spent by these units in the system, i.e.,

$$(10) \qquad E(n) = \lambda\tau.$$

Certain weak restrictions are placed upon the queueing process, but these need not concern us since all the systems with which we deal satisfy these conditions.

The definition of system in this equality is left unspecified, and so, we may choose to define it as the queue itself, in which case we use the notation $\tau = W$; or we may choose to define it as the system which includes both the queue and the service facility, in which case we use the notation $\tau = T$. In addition, we may choose to separate the units in the system into a set of subgroups, in which case, the equality above holds for each subgroup separately (i.e., labeling the p^{th} subgroup by the subscript, p, we have $E(n_p) = \lambda_p \tau_p$ where τ_p may take the form W_p or T_p, depending upon the choice of the definition of the system).

PROOF OF THE CONSERVATION LAW

Let us define U(t) as the total unfinished work* present in the system at time t. In particular, U(t) represents the time that it would take to empty the system of all units present at time t, if no new units were allowed to enter the system after time t. A typical section of U(t) might look like the graph shown in Fig. 1.

The instants t_i are the times of arrival (independent and Poisson) of new units to the system, each unit having its service time, v_i chosen independently from some distribution. The U(t) function decreases at a steady rate of 1 sec/sec as long as U(t) is positive; it jumps by v_i at the times t_i, and once having reached zero, it remains there until the next unit's

*Benes [1] defines a function W(t) similar to U(t), which he calls the virtual waiting time, which is the time a customer would have to wait for service if he arrived at time t in a first come first served system. U(t) is distinct from W(t) in that it does not, in general, represent a customer's waiting, but rather, represents the backlog of work from the service facility's point of view. When service is given in order of arrival, then W(t) and U(t) are identical.

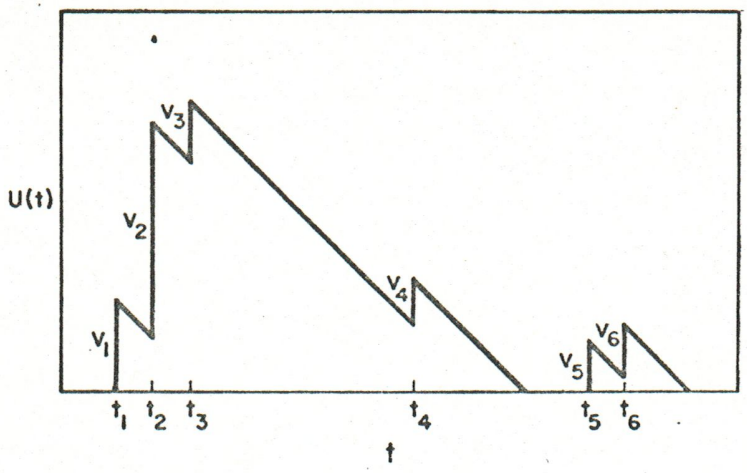

Figure 1 - Total unfinished work, U(t), in the system

arrival. Now, it is clear, that the following limit is well-defined (and exists whenever $\rho < 1$ for the system under consideration):

$$\overline{W} = \lim_{T \to \infty} \frac{1}{T} \int_0^T U(t)\, dt .$$

Thus $\overline{W}$ is defined as the time average of U(t).

Let us restrict the class of queueing systems that we consider to those which satisfy the conditions 1-4. In particular, these first three restrictions produce a U(t) function which is a function only of the v_i and the t_i. This is true since the only times that U(t) increases is at those times (t_i) when a new unit enters the system bringing with it a required service time v_i. The third assumption guarantees that pre-emptive disciplines introduce no new work into the system. Aside from the time (t_i) when U(t) jumps, it must decrease at a rate of 1 sec/sec (as long as it is positive) since the second assumption forces the service facility to be busy working whenever any unit is in the system; as long as U(t) is positive some unit must still be in the system. The U(t) function cannot jump downward discontinuously since this would correspond to the premature departure of an incompletely serviced unit and this is prohibited by the first assumption. As far as U(t) is concerned, the order in which units are serviced is immaterial since the total unfinished work is what U(t) measures. Thus under these restrictions, it is clear that no matter what discipline is used (priority, pre-emption, or what have you), only the set t_i and v_i determine the form of U(t), and as long as the same set of t_i and v_i are involved, the function U(t) will be the same. It is further obvious, that no matter which U(t) function turns up, as long as the same statistics are used for the t_i and v_i, the expected value $\overline{W}$, of the unfinished work will be the same.

One recognizes that the expected value of the waiting time (in queue only) for a unit in a strict first come first serve discipline is just $\overline{W}$ (i.e., the waiting time is exactly equal to the unfinished work in a Poisson arrival first come first served system). Now, in view of the independence of $\overline{W}$ to the particular discipline used, we proceed to calculate $\overline{W}$ for a strict first come first served discipline; the calculation consists of deriving an expression for the expected value of

the waiting time (in queue) since we have seen that this value is just $\overline{W}$ itself. We consider first, the case $0 \leq \rho < 1$.

Accordingly, let us consider the value of U(t) at an arbitrary instant of time. Let there be n_p type p units present in the queue at this time; also, let t_{ip} represent the time which is yet to be spent in service by the i^{th} unit ($i = 1, 2, \ldots, n_p$) of type p. Further, let t_0 be the time required to complete service on the unit found in service at this time. Thus, U(t) may be written as

$$U(t) = t_0 + \sum_{p=1}^{P} \sum_{i=1}^{n_p} t_{ip}. \tag{11}$$

We have separated the units in the system into P classes. This is done in anticipation of applying the result of this derivation to priority systems, etc., which have P classes of units. Now t_0, t_{ip}, n_p are all random variables. Let us next form the expected value* on both sides of (11):

$$\overline{W} = V_1 + \sum_{p=1}^{P} \sum_{n_p=0}^{\infty} r(n_p) \sum_{i=1}^{n_p} E(t_{ip}) \tag{12}$$

where clearly,† $E(t_0) = V_1$ and $r(n_p)$ is the probability that n_p type p units are present in the queue. We define

$$E(t_{ip}) = 1/\mu_p,$$

where all service times for type p units are chosen independently from the same distribution (not necessarily exponential)‡ whose mean is $1/\mu_p$. Thus (12) becomes

$$\overline{W} = V_1 + \sum_{p=1}^{P} (1/\mu_p) \sum_{n_p=0}^{\infty} n_p r(n_p).$$

Now, from (10) we recognize that

$$\sum_{n_p=0}^{\infty} n_p r(n_p) = \lambda_p W_p.$$

Thus, we arrive at the following general form for $\overline{W}$:

*Note that (11) is capable of yielding more relationships of the type stated in (6). These may be obtained by first raising (11) to the n^{th} power and then taking expected values.

†Equation (7) which gives an explicit expression for V_1, has been derived by a number of authors; for example, a simple derivation may be found in Saaty [8, Sec. 11-2.1a].

‡In the case of pre-emption, we insist upon an exponential distribution of service time (see assumption 3) which, due to the memoryless property of the exponential distribution, allows us to say that the expected time remaining for any pre-emptied unit is still $1/\mu_p$.

$$\overline{W} = V_1 + \sum_{p=1}^{P} \rho_p W_p . \tag{13}$$

Let us now evaluate $\overline{W}$ by considering a strict first come first served discipline with Poisson input traffic; this implies that all waiting times, W_p are equal, and in particular, $W_p = \overline{W}$ for all p since U(t) represents the virtual waiting time in the first come first served case. Thus, we convert (13) to

$$\overline{W} = V_1/(1-\rho) . \tag{14}$$

Substituting the value of $\overline{W}$, as given by (14), into (13), we obtain

$$\sum_{p=1}^{P} \rho_p W_p = [\rho/(1-\rho)] V_1 ,$$

which establishes (6) for $0 \leq \rho < 1$.

For the case $\rho \geq 1$ we need only recognize that the input traffic rate exceeds* the service rate in which case we see immediately that at least one of the W_p (where $\rho_p > 0$) grows without bound. Of course, in such a case, we have no steady-state solution. This completes the proof of the Conservation Law.

It is convenient to digress at this point in order to illustrate a simple method of establishing the result,

$$V_1 = \sum_{p=1}^{P} \rho_p/\mu_p$$

for exponentially distributed service times. Let us define T_p as equal to $W_p + (1/\mu_p)$. Also define $(1/\mu_p')$ as the expected value of the additional time required by a unit of type p, given that this unit was still in the system at an arbitrary instant of time. Accordingly, the expected value of the unfinished work is

$$\overline{W} = \sum_{p=1}^{P} \lambda_p T_p/\mu_p' , \tag{15}$$

where, once again, we have used (10). Taking advantage of the memoryless property of exponential distributions, we come to the conclusion that

$$1/\mu_p' = 1/\mu_p .$$

Using this, as well as the substitution $T_p = W_p + (1/\mu_p)$ in (15), we find that

*For $\rho \to 1$, we note that $[\rho/(1-\rho)]\, V_1$ approaches ∞. The limiting case for $\rho = 1$ is discussed fully by Lindley [5].

$$\overline{W} = \sum_{p=1}^{P} (\rho_p/\mu_p) + \sum_{p=1}^{P} \rho_p W_p .$$

Comparing this to (13), we conclude that, for exponential service times,

$$V_1 = \sum_{p=1}^{P} \rho_p/\mu_p .$$

COROLLARIES TO THE CONSERVATION LAW, AND THEIR PROOF

There exist priority disciplines for which $\rho \geq 1$ and for which a subclass of the priority groups obtains a bounded steady-state solution for W_p. In particular, this is true for the head of the line discipline studied by Cobham [2]. If we consider such systems with $0 \leq \rho$, we expect that some of the W_p may grow without bound; let us label this set with the indices $p = 1, 2, \ldots, j-1$. For $p = j, j+1, \ldots, P$ we expect* bounded W_p. The Conservation Law holds, of course, but we wonder what conservation constraints on the waiting time may exist for those groups with $p \geq j$. We express these constraints in the following two corollaries.

COROLLARY 1: For $0 \leq \rho$ and a head of the line priority discipline with no preemption, and under restrictions 1, 2, and 4 above,

$$\sum_{p=j}^{P} \rho_p W_p = [s_j/(1-s_j)]\,(V_j + V_j') , \tag{16}$$

where

$$V_j = 1/2 \sum_{p=j}^{P} \lambda_p E\left(t_p^2\right), \tag{17}$$

$$V_j' = (f/2)\,\lambda_{j-1} E\left(t_{j-1}^2\right), \tag{18}$$

$$j = \text{smallest positive integer such that } \sum_{p=j}^{P} \rho_p < 1 \tag{19}$$

$$s_j = \sum_{p=j}^{P} \rho_p$$

and

$$f = \begin{cases} 0 & \rho < 1 \\ (1-s_j)/\rho_{j-1} & \rho \geq 1 \end{cases} . \tag{20}$$

*Once again, the reader is referred to Phipps [7].

$$\sum_{p=j}^{P} \rho_p W_p = [s_j/(1-s_j)]\,(V_j + V_j').$$

Due to the Poisson input statistics, we may apply the result that Cobham [2] and Phipps [7] obtained,* i.e.,

$$V_j = (1/2) \int_0^\infty t^2 \sum_{p=j}^{P} \lambda_p \, dF_p(t) = (1/2) \sum_{p=j}^{P} \lambda_p E\left(t_p^2\right)$$

and

$$V_j' = (f/2) \int_0^\infty t^2 \lambda_{j-1} \, dF_{j-1}(t) = (f/2)\, \lambda_{j-1} E\left(t_{j-1}^2\right)$$

Also, we notice that $f\rho_{j-1}$, the fraction of time that type $(j-1)$ units utilize the service facility, may be calculated as

$$f\rho_{j-1} = 1 - s_j \qquad \text{for } \rho \geq 1,$$

and so

$$f = (1-s_j)/\rho_{j-1} \qquad \text{for } \rho \geq 1 \text{ (or } j > 1);$$

and for completeness, we define

$$f = 0 \qquad \text{for } \rho < 1 \ (j = 1).$$

With these substitutions, we note that for the case $\rho < 1$, we obtain the same result as given in (6) which, of course, we must. This completes the proof of Corollary 1.

COROLLARY 2: For a head of the line priority discipline with pre-emptive resume, exponential service time distributions, and under restrictions 1, 2, and 4 as expressed above,

$$\sum_{p=j}^{P} \rho_p W_p = [s_j/(1-s_j)]\, V_j . \tag{24}$$

PROOF: We now show how an equation similar to (16) may be obtained for a pre-emptive resume situation with exponentially distributed service times. Clearly, (21) still holds. In order to evaluate U_j, we now use the same trick as for the nonpre-emption case (i.e., form all priority groups into two groups — the first group consisting of classes $j, j+1, \ldots,$ P and all others being in the second group) except we allow members of the first group to pre-empt units from the second group. Then we see that, for $p \geq j$,

*The cumulative distribution function for the service time of the p^{th} priority group is denoted by $F_p(t)$.

$$W_p = U_j ,$$

and so, (21) becomes

$$U_j = V_j + \sum_{p=j}^{P} \rho_p U_j$$

or

$$U_j = V_j/(1 - s_j) . \tag{25}$$

Substituting (25) into (21) yields

$$V_j/(1 - s_j) = V_j + \sum_{p=j}^{P} \rho_p W_p$$

or

$$\sum_{p=j}^{P} \rho_p W_p = [s_j/(1 - s_j)] V_j .$$

Once again, V_j is as given previously. Note also that for $j = 1$, (24) reduces to (6). This completes the proof of Corollary 2.

We note here that in the case of exponentially distributed service times, we define

$$V_j = \sum_{p=j}^{P} \rho_p/\mu_p .$$

REFERENCES

[1] Benes, V. E., General Stochastic Process in the Theory of Queues (Addison-Wesley, 1963).

[2] Cobham, A., "Priority Assignment in Waiting Line Problems," Operations Research, 2, 70-76 (1954).

[3] Kleinrock, L., "A Delay Dependent Queue Discipline," N.R.L.Q., 11, No. 4, 329-341 (Dec. 1964).

[4] Kleinrock, L., "Analysis of a Time-Shared Processor," N.R.L.Q., 11, No. 1, 59-73 (Mar. 1964).

[5] Lindley, D. V., "The Theory of Queues with a Single Server," Proc. Cambridge Phil. Soc., 48, 277-289 (1952).

[6] Little, J. D. C., "A Proof for the Queueing Formula $L = \lambda W$," Operations Research, 9, 383-387 (1961).

[7] Phipps, T.E., Jr., "Machine Repair as a Priority Waiting-Line Problem," Operations Research, 4, 76-85 (1956).

[8] Saaty, T. L., Elements of Queueing Theory with Applications (McGraw-Hill Book Co. Inc., New York, 1961).

[9] Vaulot, A. E., "Delais d'attente des appels telephoniques, traites au hasard," Compt. rend. 222, 268-269 (1946).

[10] Wishart, D. M. G., "Queueing Systems in Which the Discipline is Last-come, First-served," Operations Research, 8, No. 5, 591-599 (1960).

* * *

* U.S. GOVERNMENT PRINTING OFFICE : 1966 O—209-335