Operations Research

Machine Repair as a Priority Waiting-Line Problem

Thomas E. Phipps, Jr.,

MACHINE REPAIR AS A PRIORITY WAITING-LINE PROBLEM

THOMAS E PHIPPS, JR.

Operations Evaluation Group, Massachusetts Institute of Technology, P O Box 2176, Potomac Station, Alexandria, Virginia

(Received July 29, 1955)

A paper by Cobham[4] in this JOURNAL, dealing with the assignment of priorities in waiting-line problems, is taken as the basis for treatment of a particular variety of machine-repair problem in which shortest jobs, rather than first arrivals, receive highest priority Cobham's results for the single channel case are found to be easily applicable to this type of priority assignment These results take a particularly simple form when the priority-labeling index is allowed to assume continuous, rather than discrete, values They apply to any case in which a one-channel facility repairs random failures from a very large population of machines

THE CUSTOMARY, or 'Swedish telephone,' approach[1, 2] to machine maintenance treats machines as being entitled to the same rights as telephone subscribers in respect to service in the strict order of arrival Consideration reveals, however, that no company employing large numbers of similar machines would be well advised to allow its maintenance shop to repair them on the basis of such a work-priority policy An optimum priority system would assign highest priority to the *shortest job* in the waiting line, rather than to the job that has been kept waiting the longest This has been established analytically by Cobham[3] for the case of 'exponential holding times,' but seems evident intuitively for any type of repair time distribution The company is presumably concerned with keeping as many of its machines operable as much of the time as possible, such a purpose is better served by a repair-priority system based on job duration than on waiting duration

The subject of priority assignment in waiting-line problems, for general priority systems, has been opened up by a pioneering paper of Cobham's[4 5 6] In his article Cobham demonstrated that (although probability distributions connected with priority waiting-line problems, except in simple cases, continue to defy analysis because of ingrained non-Markovian probability correlations) certain *expectation values* characteristic of the waiting process can be easily calculated He derived explicit expressions for the expected waiting time of a unit of given priority, for an arbitrary servicing time distribution in the case of a single-channel system (single repair man or cooperative servicing group) and for an exponential servicing time distribution in the case of multiple channels Cobham pointed out[4] that his results have application in the field of machine maintenance, but he has not further explored the possibilities in published work

In the present paper we shall do this It is found possible to make direct use of Cobham's single-channel results The generality of Cobham's single-channel treatment, unfortunately not characteristic of his multiple-channel analysis, is entirely necessary for dealing with the machine maintenance problem, as here construed Therefore, the interesting case in which several repairmen are available for simultaneous servicing of more than one machine cannot be attacked by the methods of the present paper

From a practical viewpoint the results obtained are limited in applicability by the assumption of an infinite population of operating machines (i e , of potential repair jobs) Our results will serve, however, as a guide to further thinking about priority assignment in machine maintenance, and as a basis for approximate quantitative study of problems involving very large numbers of machines

REVIEW OF COBHAM'S FORMULAS

COBHAM'S BASIC ASSUMPTIONS AND NOTATION will be used A summary of his results follows, stated in terminology appropriate to the machine-repair problem

The 'company' is assumed to own an infinite number of machines, subject to random (Poisson law) breakdown A steady state is considered in which the number of operable machines is infinite In this steady state new repair jobs of various magnitudes are generated at an average arrival rate λ No specialized assumption is made about the distribution of repair times Repair of random failures alone is considered, necessary checks and routine preventive maintenance of a periodic, predictable nature are presumed to be accomplished by some additional facility not connected with the problem

The company possesses a single-channel repair facility (i e , only one machine can be repaired at a time) Each repair job is assigned a numerical priority index p, which in Cobham's formulation takes on integral values ranging between unity and a positive integer r A job with priority p_1 is worked on before any with priority p_2, if $p_1 < p_2$ Jobs of equal priority are worked on in the order of their arrival Jobs of priority p arrive in Poisson law fashion at an average rate λ_p, evidently, $\lambda = \sum_1^r \lambda_p$ The average arrival rate of jobs actually worked on by the maintenance facility is

$$\lambda^* = \sum_1^r f_p \lambda_p, \tag{1}$$

where f_p is the probability that a job of priority p receives service from the maintenance facility This quantity is introduced in the interest of unifying Cobham's treatments of the saturated[6] and unsaturated[4] cases, defined below, it is evaluated in a later paragraph

The probability that a job of priority p requires a repair time less than t is denoted by $F_p(t)$ The average cumulative job repair time distribution for units of all priorities is

$$F(t)=\frac{1}{\lambda}\sum_1^r \lambda_p F_p(t)$$

The corresponding repair time distribution for units actually worked on by the repair facility is

$$F^*(t)=\frac{1}{\lambda^*}\sum_1^r f_p \lambda_p F_p(t) \tag{2}$$

Of course, $$\int_0^\infty dF_p(t)=\int_0^\infty dF(t)=\int_0^\infty dF^*(t)=1$$

The average repair time for units of the pth priority level is

$$\frac{1}{\mu_p}=\int_0^\infty t\, dF_p(t) \tag{3}$$

The average repair rate μ^* for jobs worked on by the facility is seen to be

$$\frac{1}{\mu^*}=\left(\sum_1^r \frac{f_p \lambda_p}{\mu_p}\right)\Big/\left(\sum_1^r f_p \lambda_p\right) \tag{4}$$

If the repair facility is adequate to handle all work required of it, $f_p=1$ for all p, and we shall speak of the system as 'unsaturated,' or as operating 'below saturation' In this case the average repair rate μ is given by

$$\frac{1}{\mu}=\frac{1}{\mu^*}=\frac{1}{\lambda}\sum_1^r \frac{\lambda_p}{\mu_p} \qquad \text{(Unsaturated case)}$$

Since the repair rate capability must exceed the job arrival rate for unsaturated activity, i e , $\mu>\lambda$, the condition for the obtaining of this case is

$$\sum_1^r \frac{\lambda_p}{\mu_p}<1 \qquad \text{(Unsaturated case)} \tag{5}$$

If this inequality is not satisfied, $f_p<1$ for some p, and the system will be said to be 'saturated,' implying the growth of an infinite waiting line

$$\sum_1^r \frac{\lambda_p}{\mu_p}\geq 1 \qquad \text{(Saturated case)} \tag{6}$$

In this case the average repair rate of jobs admitted to the maintenance facility must equal the average arrival rate of such jobs, $\mu^*=\lambda^*$ From equations (1) and (4), this implies that

$$\sum_1^r \frac{f_p \lambda_p}{\mu_p}=1, \qquad \text{(Saturated case)} \tag{7}$$

for a steady state This relation provides a determining condition on the f_p, together with the evident requirements that $0 \leq f_p \leq 1$ for all p, and that no f_{p_1} can be less than unity if any $f_{p_2} > 0$ for $p_2 > p_1$ (That is, if any low priority units are worked on, all higher priority units must certainly be serviced) Thus, in the saturated case equation (7) determines an integer s and a fraction m, $0 < m \leq 1$, such that

$$f_p = \begin{cases} 1 & \text{for} \quad p < s, \\ m & \text{for} \quad p = s, \\ 0 & \text{for} \quad p > s \end{cases} \qquad \text{(Saturated case)} \quad (8)$$

As previously stated, in the alternative case,

$$f_p = 1 \quad \text{for all } p \qquad \text{(Unsaturated case)} \quad (9)$$

The determining equation for s and m is

$$\sum_1^{s-1} \frac{\lambda_p}{\mu_p} + m \frac{\lambda_s}{\mu_s} = 1 \qquad (10)$$

Cobham's final result[4] for the expected waiting time W_p of a job of priority p, expressed in his equation (3), may be written

$$W_p = W_0 \bigg/ \left[\left(1 - \sum_1^{p-1} \frac{f_k \lambda_k}{\mu_k} \right) \left(1 - \sum_1^{p} \frac{f_k \lambda_k}{\mu_k} \right) \right] \qquad (11)$$

We have taken the liberty of inserting f-factors in the denominator for consistency with the rest of our development These alter nothing in the unsaturated case and provide a necessary correction[6] in the saturated case The constant W_0 is evaluated[6] as

$$W_0 = \tfrac{1}{2} \lambda^* \int_0^\infty t^2 \, dF^*(t) \qquad (12)$$

It is convenient to introduce the concept of an *active waiting line*, each member of which is certain to receive eventual service from the facility The expected number of jobs in the active waiting line is

$$N = \sum_1^r f_p \lambda_p W_p \qquad (13)$$

In contrast, saturated activity produces a category of low-priority units that never receive service from the facility in question These units we shall call *permanent duds*, by analogy with the case of aircraft repair, in which aircraft too seriously damaged for local repair are sometimes so termed Permanent duds will be thought of as distinct from jobs entering the active waiting line, and subject to altogether different treatment.

Inspection of the denominator in equation (11) casts light on the nature of the permanent dud phenomenon In the unsaturated case [equation (5) satisfied] this denominator cannot vanish This circumstance implies finite expected waiting times for jobs of all priorities, with resulting unit probability that any job demanding service will eventually receive it [cf equation (9)] In the case of saturation [equation (6) satisfied] the denominator in equation (11) vanishes for any $p \geq s$ [cf equations (7) and (8)] This fact implies an infinite expected wating time for any job with such a priority index Consequently there is zero probability (or at best a probability m in the case $p=s$) that the job can be serviced during any finite interval of observation Jobs of priority $p>s$ thus automatically become permanent duds, jobs of priority s may be thought of as having their fate decided by the toss of a skew coin, with probability m of being admitted to the active waiting line, probability $(1-m)$ of being consigned to the permanent dud category

Similarly, in the unsaturated case the expected number N of jobs in the waiting line is finite, because each of the W_p is finite [see equation (13)] But in the case of saturation there is a critical priority index s such that f_s is finite and W_s, as given by equation (11), is infinite Hence, in this case the length of the active waiting line is infinite Moreover, since permanent duds are continually being generated, and since the process, to be considered an equilibrium one, must have been in progress for an infinite time, the permanent dud population is also infinite, and is growing roughly linearly with time The process of saturated activity, as hitherto described, is therefore one of quasi-equilibrium only To achieve a true equilibrium process, it is convenient to modify the model to the extent of supposing that each permanent dud is repaired or replaced after a finite time by some means external to the problem This is realistic, since no company would allow its stock of machines to deteriorate indefinitely without taking special action

The fact that the active waiting line is of infinite length in the case of saturation is consistent with our previous observation that $f_p=0$ for $p>s$, for the presence of an infinite waiting line of higher priority jobs implies zero probability of ever being able to service a low priority job Similarly the fact that W_p, as given by equation (11), is finite for $p<s$ implies $f_p=1$ for such p-values, since a finite expected wait for service means unit probability of eventual service For the critical priority level $p=s$ an infinite wait W_s is to be expected, but a fraction m of the jobs of this level are assigned to the active waiting line and may expect to wait finite times in excess of W_{s-1} Although each of these jobs of priority s in the active waiting line receives eventual service (with unit probability) the number of them is infinite The unlimited character of the active waiting line derives solely from units of the critical priority level.

MACHINE REPAIR

IN APPLYING THE FORMULAS just reviewed to the problem of machine maintenance, it is important to take note of the assumptions characteristic of the model Particular attention is called to the assumption, inherited from telephone theory, that 'calls are not interrupted', that is, once the maintenance establishment is committed to the repair of a particular machine the repair is completed, even though a job of higher priority may enter the waiting line meanwhile As a question of optimum repair policy, the issue of whether or not a job *should* be interrupted in such circumstances is open to debate The answer cannot be given in general, but depends on the nature of the job There are certain types of repair work, requiring great concentration or involving much dismantling and covering the floor with nuts and bolts, which certainly should not be interrupted, once undertaken In other cases it would be inadvisable to continue work on a long job if one or more short ones became available It is not our purpose here to explore all interesting variations of the machine-maintenance problem, but to pursue the particular one suggested by Cobham's work

In the machine-repair problem, highest priority is assigned to the shortest job in the waiting line, and any job takes priority over any longer job It is presupposed that the repair facility has resources of experience that enable it to estimate the duration of repair jobs accurately in advance, at least to the extent of being able to order the jobs by duration This seems generally true in practice, although there may be exceptions Each job as it enters the waiting line for repair is labelled with a priority index t, which is its estimated repair time Under the simplifying assumption of perfect estimation, t is also the actual repair time, when and if repair is undertaken

It will be noted that small values of t correspond to high priorities, so that we have in t a priority-index parameter which behaves exactly like Cobham's priority index p, except that it assumes continuous, rather than discrete values This circumstance enables us to take over Cobham's results verbatim, with t substituted for p and sums over p replaced by integrals with respect to t (If t, as a time *estimate*, is allowed only discrete values, Cobham's results apply directly However, the continuous formulation is simpler and more elegant)

The analogue of Cobham's λ_p is $\lambda_t\, dt$, the arrival rate of jobs having durations between t and $t+dt$ Evidently $\lambda_t\, dt=\lambda\, dF(t)$ The function f_p, given by equation (8), has as its continuous case analogue a step function $f(t)$, viz ,

$$f(t)=\begin{cases}1 & \text{for} \quad t<\tau, \\ 0 & \text{for} \quad t>\tau\end{cases} \qquad \text{(Saturated case)} \quad (14)$$

The critical job duration τ, analogous to the critical priority level s, is determined by the means discussed below.

The analogue of $1/\mu_p$, the average servicing time for units of the p priority level is the average repair time for jobs of duration t, which, of course, is just equal to t. Similarly, $F_{t'}(t)$, the analogue of Cobham's $F_p(t)$, is the cumulative job repair time distribution for units of priority t'. That is, it is the probability that a job of duration t' will have a repair time less than t. As such, it is a step function,

$$F_{t'}(t)=\begin{cases}0 & \text{for} \quad t'>t,\\ 1 & \text{for} \quad t'<t\end{cases} \tag{15}$$

The continuous analogue of equation (1), above, is

$$\lambda^*=\int_0^\infty f(t)\,\lambda_t\,dt=\lambda\int_0^\tau dF(t)=\lambda\,F(\tau) \tag{16}$$

The analogue of equation (2) is

$$\begin{aligned}F^*(t)&=\frac{1}{\lambda^*}\int_0^\infty f(t')\,\lambda_{t'}\,dt'\,F_{t'}(t)\\ &=\frac{1}{\lambda^*}\int_0^t f(t')\,\lambda_{t'}\,dt'=\frac{\lambda}{\lambda^*}\int_0^t f(t')\,dF(t')\end{aligned} \tag{17}$$

For $t\leq\tau$, this implies that

$$F^*(t)=F(t)/F(\tau) \tag{18}$$

The analogue of equation (3) is

$$\begin{aligned}\frac{1}{\mu^*}&=\left(\int_0^\infty f(t)\,\lambda_t\,dt\;t\right)\Big/\left(\int_0^\infty f(t)\,\lambda_t\,dt\right)=\left(\int_0^\tau t\,dF(t)\right)\Big/\left(\int_0^\tau dF(t)\right)\\ &=\frac{1}{F(\tau)}\int_0^\tau t\,dF(t)\end{aligned} \tag{19}$$

The case of non-saturation is obtained from these formulas by replacing τ by ∞ and noting that $F(\infty)=1$.

Conditions defining the saturated and unsaturated cases are obtained by analogy with equations (5) and (6), *viz*.,

$$\int_0^\infty \lambda_t\,dt\cdot t=\lambda\int_0^\infty t\,dF(t)<1, \qquad \text{(Unsaturated case)} \tag{20}$$

and

$$\lambda\int_0^\infty t\,dF(t)\geq 1 \qquad \text{(Saturated case)} \tag{21}$$

The critical job duration τ is determined by analogy with equation (7) or (10), viz,

$$\int_0^\infty f(t)\,\lambda_t\,dt\ t=\int_0^\tau \lambda_t\,dt\ t=\lambda\int_0^\tau t\,dF(t)=1. \qquad (22)$$

If $dF(t)$ behaves in such a way that ambiguity arises, it is to be understood that τ has the *smallest* value satisfying equation (22) (In fact, there are certain mathematical restrictions on $dF(t)$ that we have not stated For practical purposes it is simplest to rule out infinite singularities of this function, although the formal treatment could be modified to admit a plurality of such)

From equations (16), (19), and (22) the reader will readily confirm that $\mu^*=\lambda^*$ in the case of saturation

Both summations appearing in the denominator of equation (11) become in the continuous limit

$$\sum_1^p \frac{f_k\lambda_k}{\mu_k}\rightarrow\int_0^t f(t')\,\lambda_{t'}\,dt'\ t'=\lambda\int_0^t f(t')\,t'\,dF(t')=\lambda\int_0^t t'\,dF(t') \quad \text{for} \quad t\leq\tau$$

Consequently our final result for the expected waiting time of a job of duration t is

$$\begin{aligned} W_t&=W_0\Big/\left(1-\lambda\int_0^t t'\,dF(t')\right)^2 && \text{for} \quad t\leq\tau \\ W_t&=\infty && \text{for} \quad t>\tau \end{aligned} \qquad (23)$$

Making use of the relation $\lambda^*\,dF^*(t)=f(t)\,\lambda\,dF(t)$, which follows from equation (17), we evaluate the constant W_0 [cf equation (12)] as

$$W_0=\tfrac{1}{2}\lambda\int_0^\infty f(t)\,t^2\,dF(t)=\tfrac{1}{2}\lambda\int_0^\tau t^2\,dF(t). \qquad (24)$$

The active waiting line consists of jobs of duration less than τ, all other jobs are permanent duds The analogue of equation (13) is

$$N=\int_0^\infty f(t)\,\lambda_t\,dt\,W_t=W_0\int_0^\tau\left[dF(t)\Big/\left(1-\lambda\int_0^t t'\,dF(t')\right)^2\right] \qquad (25)$$

This integral is improper for the case of saturation, so that the active waiting line, as before, is infinite It acquires this character of infiniteness by an accumulation of job durations having τ as a least upper bound Thus, for any $\epsilon>0$ the number of jobs of duration less than or equal to $(\tau-\epsilon)$ in the active waiting line is finite In the case of non-saturation the integral in equation (25) is finite, so that there is a finite waiting line All other qualitative features of the discrete priority index problem carry over to the continuous formulation The greater formal simplicity of the latter will be noted

It is of interest to apply the general formulas just derived in a specific example

Exponential Repair Times

The cumulative distribution function of repair times is $F(t)=1-e^{-\mu t}$ From equations (20) and (21) it is seen that saturation occurs if $\lambda \geq \mu$,

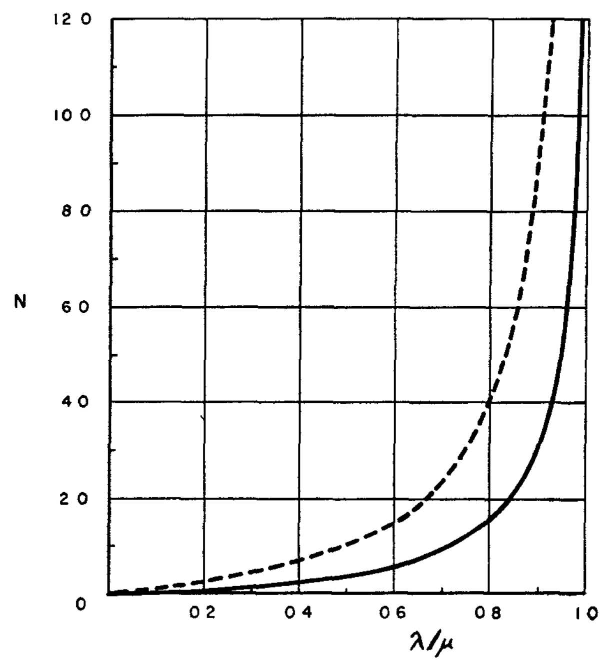

Fig. 1. Expected number N of units in the waiting line, as a function of 'utilization factor,' λ/μ, for exponential servicing times, unsaturated system The dotted curve is the conventional result for jobs treated on a 'first come, first served' basis, the solid curve applies to the case in which shortest jobs are given highest priority

nonsaturation if $\lambda < \mu$ From equation (22) the critical job duration τ in the case of saturation is given by the solution of $(\lambda/\mu)\,[1-e^{-\mu\tau}(1+\mu\tau)]=1$ In this case the average repair rate μ^* of jobs in the active waiting line is given, with the help of equation (16), by $\mu^*=\lambda^*=\lambda\ F(\tau)=\lambda\,(1-e^{-\mu\tau})$ In the case of non-saturation $\mu^*=\mu$

The constant W_0, given by equation (24), is evaluated to be

$$W_0=\lambda/\mu^2, \qquad \text{(Unsaturated case)}$$

$$W_0=(1/\mu)[1-\lambda\tau\, e^{-\mu\tau}] \qquad \text{(Saturated case)}$$

From equation (23) the expected waiting time for a job of duration t is

$$W_t = W_0/\{1-(\lambda/\mu)[1-e^{-\mu t}(1+\mu t)]\}^2$$

In the saturated case it is necessary to make the stipulation that $t \leq \tau$ in the expression just given, for t-values greater than τ, W_t is infinite

The expected number of jobs in the waiting line in the unsaturated case is given by equation (25),

$$N = \frac{\lambda^2}{\mu}\int_0^\infty \frac{e^{-\mu t}\,dt}{\{1-(\lambda/\mu)[1-e^{-\mu t}(1+\mu t)]\}^2}$$

N is infinite in the saturated case

The quantity N, calculated numerically from this formula, is graphed in Fig 1 (solid curve) Also shown (dotted curve) is the expected length of the waiting line for priority assignment on a 'first come, first served' basis The latter is given by the well-known formula, $N = (\lambda/\mu)/[1-(\lambda/\mu)]$ Exponential holding times and an unsaturated system are assumed Comparison of the two curves shows the effectiveness of an optimum priority system in reducing average waiting line lengths

SUMMARY

THE GENERAL EXPECTATION-VALUE FORMULAS developed by Cobham have, as he predicted, shown themselves useful in dealing with a certain variety of machine-maintenance problem In making this application we found that it simplified the principal formulas to introduce a continuous, rather than a discrete, priority index parameter

The methods here employed permit the description of a servicing facility operating far beyond its natural capacity, through the institution of a high-priority 'active waiting line' and a separate category of 'permanent duds,' which are denied service Although the model considered is limited in scope, it is believed that these concepts are of fairly general applicability

The writer wishes to acknowledge profitable discussions of the subject of this paper with Mr H W Kreiner, Mr C F Kent, and Dr J M Danskin, of the Operations Evaluation Group The assistance of Mr R E Weinstein in computing the Fig 1 is also gratefully acknowledged

REFERENCES

1 FELLER, *Probability Theory and Its Applications*, vol 1, Wiley, New York, 1950
2 C PALM, *Industritidningen Norden* **75,** 75–80, 90–94, 119–123 (1947)
3 A COBHAM, Internal Memorandum of the Operations Evaluation Group, Washington, D C (unpublished)
4 A COBHAM, *J Opns Res Soc Am* **2,** 70 (1954)
5 J L HOLLEY, *J Opns Res Soc Am* **2,** 341 (1954)
6 A COBHAM, *J Opns Res Soc Am* **3,** 547 (1955)