JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 76B, Nos. 3 and 4, July–December 1972

Scheduling a Time-Shared Server to Minimize Aggregate Delay*

A. J. Goldman

Institute for Basic Standards, National Bureau of Standards, Washington, D.C. 20234

(July 21, 1972)

A simplified analysis is given of a problem situation, previously treated in the literature, which pertains to the delay-minimizing allocation of servicing times among N incoming streams requiring "processing" of some kind by a single "server" (e.g., a time-shared computer). The original problem is generalized to permit different "weights" for the delays suffered by different streams.

Key words: Computer systems; optimization; scheduling; time-sharing; traffic control.

1. Introduction

A 1967 paper by Rangarajan and Oliver [1][1] contains a formulation and analysis of the two problems described below, which pertain to the allocation of servicing times among N incoming streams requiring "processing" of some kind by a single "server." The server might for example be a switching point or a congestion point (e.g., a tunnel entrance) in a transport network, in which case "processing" an item (vehicle) simply means letting it through. Or, the server might be a computer handling reservations from several ticket offices, or exercising real-time control over vehicle movements on several network links, or performing some other tasks on a time-shared basis.

The streams are treated as continuous flows. During each service *cycle*, of duration T, the server handles stream 1 for time G_1, switches (with associated known switch-over or "dead" time) to handle stream 2 for time G_2, etc. The arrivals in each stream are assumed nonrandom, with a known uniform rate (possibly different for different streams). The server's processing rate, when serving a particular stream, is also assumed nonrandom and constant (possibly different for different streams). Each G_i is constrained to be at least large enough so that no queue remains in the ith stream when one of that stream's service periods ends.

The two problems formulated and analyzed are these:

PROBLEM 1: For given cycle time T, what allocation $G_1, G_2, \ldots, G_N$ of service times among the various streams is optimal, in the sense of minimizing total waiting time per cycle?

PROBLEM 2: What value of the cycle time T will minimize average waiting time?

Subsequently Horn [2] showed that the more general case, in which all streams are served equally often (possibly *more* than once) per cycle, can be reduced to PROBLEM 1. This provides additional reason for offering an alternative analysis, which is more self-contained and (at least to the writer) simpler than that of reference [1]. In addition, a mild generalization will be introduced by permitting the penalties for delay to be different for different streams.

AMS Subject Classification Number: Primary 9310.
* Research supported by the Northeast Corridor Transportation Project (Dept. of Transportation) and the NBS Center for Computer Science and Technology.
[1] Figures in brackets indicate the literature references at the end of this paper.

2. Formulation

The data for PROBLEM 1 are

T = cycle time,
N = number of streams,
a_i = arrival rate for ith stream,
s_i = service rate limit when processing ith stream ($s_i > a_i$),
d_i = dead time in switching from ith stream to next one,
p_i = penalty factor for delays to ith stream.

Note that our (a_i, s_i, d_i) are the $(\lambda_i, \mu_i, \tau_i)$ of reference [1], which in effect assumes all $p_i = 1$.

Under the assumption of first-in-first-out service within each stream, the waiting time per cycle for the ith stream is found as in reference [1] to be

$$W_i[T, G_i] = \{a_i(T-G_i)^2 + a_i^2(T-G_i)^2/(s_i - a_i)\}/2$$

$$= a_i s_i (T-G_i)^2/2(s_i - a_i). \tag{1}$$

(The factor 1/2 was omitted from the analogous equation in reference [1]; also our (1) differs from that formula by a factor T because we work with total rather than time-averaged delay.) Thus the function to be minimized is

$$W^0 = \sum_1^N p_i a_i s_i (T-G_i)^2/2(s_i - a_i). \tag{2}$$

The condition, that each stream have its queue disappear before its service period ends, is expressed by

$$G_i(s_i - a_i) \geqslant (T-G_i)a_i \qquad \text{or equivalently} \qquad G_i s_i \geqslant T a_i,$$

which is equivalent to

$$(T-G_i)s_i \leqslant T(s_i - a_i). \tag{3}$$

The remaining constraint on the G_i's is the obvious identity which can be expressed, in terms of total service time and total dead time

$$G = \sum_1^N G_i \qquad \text{and} \qquad D = \sum_1^N d_i,$$

in the form

$$G + D = T. \tag{4}$$

We simplify by introducing the new variables

$$x_i = (T-G_i)/T,$$

and also

$$W = 2W^0/T^2 \tag{5}$$

as the new minimand, equivalent to the previous one since T is fixed for PROBLEM 1. Furthermore, let

$$b_i = (s_i - a_i)/s_i > 0,$$

$$c_i = p_i a_i / b_i = p_i a_i s_i / (s_i - a_i) > 0,$$

$$q_i = 1/c_i = b_i/p_i a_i > 0,$$

$$B = N - 1 + (D/T).$$

Then from (2) and (5), we see that PROBLEM 1 requires the minimization of

$$W(T) = W = \sum_1^N c_i x_i^2 \tag{6}$$

subject to the conditions (3), which are equivalent to

$$0 \leq x_i \leq b_i, \tag{7}$$

and to condition (4), which is equivalent to

$$\sum_1^N x_i = B. \tag{8}$$

From (7) and (8) we obtain the condition

$$B \leq \sum_1^N b_i, \tag{9}$$

which is both necessary and sufficient for the consistency of the constraints, and is assumed to hold in what follows.

3. Solution of PROBLEM 1

Since the problem requires minimizing a continuous strictly convex function over the closed bounded subset of x-space defined by (7) and (8), there must exist a unique relative minimum which is in fact the unique absolute minimum. Hence we need only derive enough necessary conditions, for a local minimum, to single out just one point in x-space.

The streams will be numbered (in analogy with p. 76 of ref. [1]), so that

$$p_1 a_1 = b_1 c_1 \geq p_2 a_2 = b_2 c_2 \geq \ldots \geq p_N a_N = b_N c_N > 0. \tag{10}$$

Observe first that at a local minimum,

$$c_i x_i < c_j x_j \qquad \text{implies } x_j = 0 \qquad \text{or } x_i = b_i,$$

for otherwise we could further decrease the objective function (6) without violating the constraints (7) and (8), by decreasing x_j and increasing x_i by the same sufficiently small positive quantity. Since $x_j = 0$ in this situation would lead to a contradiction of the condition $x_i \geq 0$, we in fact have

$$c_i x_i < c_j x_j \qquad \text{implies } x_i = b_i. \tag{11}$$

In analogy with eq (4a) of reference [1], let r be the smallest index for which $x_r = b_r$ in the locally optimal solution under consideration. (If $x_i < b_i$ for $i = 1, 2, \ldots, N$, then take $r = N+1$.) We next show that

$$x_i = b_i \qquad \text{if} \qquad i \geq r, \tag{12}$$

i.e., that streams 1 through $r-1$ are precisely those served longer than needed to eliminate their queues. That $x_i < b_i$ for $i < r$, follows from the definition of r. To rule out the possibility that $x_i < b_i$ for some $i > r$, note that $b_i c_i \leqslant b_r c_r$, so that

$$c_i x_i < b_i c_i \leqslant b_r c_r = c_r x_r,$$

which by (11) implies $x_i = b_i$, a contradiction.

In particular, the solution is fully determined (each $x_i = b_i$) if $r=1$, which by (12) and (9) can occur iff $fB = \sum_1^N b_i$. Thus in what follows we temporarily assume $B < \sum_1^N b_i$, so that $r > 1$.

Next, (11) and (12) imply the existence of some $K > 0$ such that

$$c_i x_i = K \qquad \text{for all } i < r,$$

or equivalently

$$x_i = q_i K \qquad \text{for all } i < r. \tag{13}$$

It follows from (8), (12), and (13) that

$$B = K \sum_1^{r-1} q_i + \sum_r^N b_i,$$

implying

$$K = \frac{(B - \sum_r^N b_i)}{\sum_1^{r-1} q_i}. \tag{14}$$

From (13) and the fact that $x_{r-1} < b_{r-1}$, we have

$$K < b_{r-1} c_{r-1} = p_{r-1} a_{r-1}. \tag{15}$$

If $r \leqslant N$, then it follows from $x_{r-1} < b_{r-1}$ and (11) . . . with $i = r-1$ and $j = r$ . . . that

$$K = c_{r-1} K q_{r-1} = c_{r-1} x_{r-1} \geqslant c_r x_r = c_r b_r (r \leqslant N). \tag{16}$$

We next dispose of the case $r = N+1$. By (8) and (13), if $r = N+1$ then

$$\begin{aligned} x_i &= \frac{q_i B}{\sum_1^N q_j} \qquad \text{(all } i\text{)}, \\ K &= \frac{B}{\sum_1^N q_i}. \end{aligned} \tag{17}$$

Using (15), we see from (18) that $r = N+1$ implies

$$B < b_N c_N \sum_1^N q_i. \tag{19}$$

Conversely if $r < N+1$, then (14) and (16) would both hold, yielding

$$B=K\sum_{1}^{r-1} q_i+\sum_{r}^{N} b_i \geqslant b_r c_r \sum_{1}^{r-1} q_i+\sum_{r}^{N} (b_i c_i) q_i \geqslant b_N c_N \sum_{1}^{N} q_i,$$

contradicting (19). So (19) is a necessary and sufficient condition for $r=N+1$.

Suppose now that $1<r<N+1$. Using (14), (15), and (16), we have

$$b_r c_r \sum_{1}^{r-1} q_i + \sum_{r}^{N} b_i \leqslant B < b_{r-1} c_{r-1} \sum_{1}^{r-1} q_i + \sum_{r}^{N} b_i$$

as the test for determining r. With r known $(1<r<N+1)$, the optimal solution is given by (12), (13), and (14). Since $b_{r-1}c_{r-1}q_{r-1}=b_{r-1}$, the test can be rewritten

$$B_r \leqslant B < B_{r-1}, \tag{20}$$

in terms of the quantities

$$B_k=b_k c_k \sum_{1}^{k-1} q_i+\sum_{k}^{N} b_i. \tag{21}$$

With the convention $B_{N+1}=0$, the test remains valid when $r=N+1$, according to the discussion surrounding (19). And with the convention $B_0=\infty$, it remains valid for $r=1$ as well (necessarily with $B=B_1$). The test is satisfied for at *most* one value of r since $B_{N+1}<B_N$ and for $1<k<N+1$,

$$B_{k-1}-B_k=(b_{k-1}c_{k-1}-b_k c_k)\sum_{1}^{k-1} q_i \geqslant 0;$$

it is satisfied for *at least* one value or r since $B_{N+1}<B\leqslant B_1$.

We conclude this section by summarizing the solution process, in terms of the problem data (assuming the ordering (10)):

Step 1: Calculate the total dead time per cycle, D.

Step 2: Calculate $B=N-1+(D/T)$.

Step 3: Calculate the quantities $b_i=(s_i-a_i)/s_i$ and their sum B_1.

Step 4: If $B>B_1$, then stop; the problem is infeasible. If $B=B_1$, the optimal solution is $G_i=Ta_i/s_i$ for all i. If $B<B_1$, continue.

Step 5: Beginning with B_1 and with $Q_0=0$, calculate quantities Q_1, B_2, Q_2, B_3, etc. by the formulas

$$q_k=\frac{b_k}{p_k a_k}$$

$$Q_k=Q_{k-1}+q_k,$$

$$B_k=B_{k-1}-(p_{k-1}a_{k-1}-p_k a_k)Q_{k-1}.$$

Stop as soon as $B_k\leqslant B$ is attained, set $r=k$, and go to Step 6. If $B<B_N$ is encountered, the optimal solution is

$$G_i=T-(TB/Q_N)q_i$$

for all i.

Step 6: Calculate

$$K=\frac{\left(B-\sum_{r}^{N} b_i\right)}{Q_{r-1}}.$$

The optimal solution is given by

$$G_i = T - TKq_i \qquad (i < r),$$

$$G_i = Ta_i/s_i \qquad (i \geqslant r).$$

4. Solution of Problem 2

Recall the relation

$$D/T = B - N + 1 \tag{22}$$

between B and T, which yields

$$dB/dT = -D/T^2. \tag{23}$$

The decreasing sequence $\{B_k\}_1^N$ defined by (21) yields, through (22), an increasing sequence $\{T_k\}_1^N$ of break-points in "T-space." The feasibility condition $B \leqslant B_1$ is equivalent to $T \geqslant T_1$, and the interval $B_k \leqslant B < B_{k-1}$ on which $r = k$ corresponds to the interval $T_{k-1} < T \leqslant T_k$.

Let $W_{\min}(T)$ be the minimized value of $W(T)$, as determined in section 3. Then by (5), we have the expression

$$W^0_{\min}(T) = T^2 W_{\min}(T)/2$$

for the minimum delay per cycle, so that

$$V^0(T) = TW_{\min}(T) \tag{24}$$

is twice the minimized *time-averaged* delay per cycle. Thus our objective in PROBLEM 2 is to choose T, subject to $T \geqslant T_1$, so as to minimize $V^0(T)$.

First consider the behavior of $V^0(T)$ on the interval (T_N, ∞) corresponding to the range $B < B_N$. By (17) and (6),

$$W_{\min}(T) = (B/Q_N)^2 \sum_1^N c_i q_i^2,$$

so that (24) yields

$$V^0(T) = (B^2 T) \times (\text{pos. const.}). \tag{25}$$

Using (23), we have

$$(d/dT)(B^2T) = B^2 + 2BT(dB/dT) = B^2 - 2B(D/T)$$

$$= B[2(N-1) - B] > B[2(N-1)) - B_N],$$

and since

$$B_N \leqslant B_1 = \sum_1^N b_i \leqslant N < 2(N-1)$$

(assuming of course that $N > 1$), it follows that (T_N, ∞) is an interval on which $V^0(T)$ is increasing, hence *not* an interval on which the minimum of $V^0(T)$ can occur.

Next, consider the behavior of $V^0(T)$ on the interval (T_{r-1}, T_r). Minimizing $V^0(T)$ over this interval is equivalent to minimizing

$$V_r(T) = V^0(T) \sum_1^{r-1} q_i. \tag{26}$$

Using (12) and (13) to substitute the optimal solution to PROBLEM 1 into (6), we obtain

$$W_{\min}(T) = K^2 \sum_1^{r-1} c_i q_i^2 + \sum_r^N c_i b_i^2$$

$$= K^2 \sum_1^{r-1} q_i + \sum_r^N c_i b_i^2,$$

which by (14) can be rewritten

$$W_{\min}(T) = \left(B - \sum_r^N b_i\right)^2 \left(\sum_1^{r-1} q_i\right)^{-1} + \sum_r^N c_i b_i^2. \tag{27}$$

It follows from (24) and (26) that

$$V_r(T) = T\left(B - \sum_r^N b_i\right)^2 + T\left(\sum_r^N c_i b_i^2\right)\left(\sum_1^{r-1} q_i\right).$$

This formula, (22) and (23) yield

$$dV_r/dT = \left(B - \sum_r^N b_i\right)\left(2N - 2 - \sum_r^N b_i - B\right) + \left(\sum_r^N c_i b_i^2\right)\left(\sum_1^{r-1} q_i\right), \tag{28}$$

$$d^2V_r/dT^2 = 2D^2/T^3 > 0. \tag{29}$$

Suppose in particular that $r \geqslant 3$. It will be shown that

$$dV_r/dT \geqslant 0 \text{ (right derivative at } T = T_{r-1}), \tag{30}$$

which by (29) implies that (T_{r-1}, T_r) is an interval on which $V_r(T)$ and hence $V^\circ(T)$ is increasing, hence *not* an interval on which the minimum of $V^\circ(T)$ can occur.

By (21) and (28), the expression in (30) whose sign is to be determined is

$$\left(B_{r-1} - \sum_r^N b_i\right)\left(2N - 2 - \sum_r^N b_i - B_{r-1}\right) + \left(\sum_r^N c_i b_i^2\right)\left(\sum_1^{r-1} q_i\right)$$

$$= \left(\sum_1^{r-1} q_i\right)\left\{b_{r-1}c_{r-1}\left(2N - 2 - 2\sum_r^N b_i - b_{r-1}c_{r-1}\sum_1^{r-1} q_i\right) + \sum_r^N c_i b_i^2\right\}.$$

This has the same sign as

$$D_{r-1} = 2N - 2 - 2\sum_r^N b_i - b_{r-1}c_{r-1}\sum_1^{r-1} q_i + (b_{r-1}c_{r-1})^{-1}\sum_r^N c_i b_i^2.$$

Since

$$b_{r-1}c_{r-1}\sum_1^{r-1} q_i \leqslant \sum_1^{r-1} b_i c_i q_i = \sum_1^{r-1} b_i,$$

we have

$$D_{r-1} \geqslant 2N - 2 - 2\sum_r^N b_i - \sum_1^{r-1} b_i + (b_{r-1}c_{r-1})^{-1}\sum_r^N c_i b_i^2.$$

Because each $b_i < 1$, the two subtracted terms in the right-hand side total less than

$$2(N - r + 1) + (r - 1) = 2N - r + 1,$$

which is no greater than $2N - 2$ for $r \geqslant 3$. Thus $D_{r-1} > 0$ for $r \geqslant 3$, verifying (30).

We have shown that the minimum of $V^\circ(T)$ over (T_1, ∞) is given by its minimum over $[T_1, T_2]$. That is, as noted in reference [1], in an optimal solution one has $r=1$ or $r=2$, so that for all but at most one stream one has $G_i/T=a_i/s_i$, i.e., all "slack time" (if there is any) is concentrated in the period allotted to a single stream.

The minimum over $[T_1, T_2]$ is determined as follows. Using (27) with $r=1$ and $r=2$, it is readily verified that $W_{\min}(T)$ is right-hand continuous at T_1. Thus the problem is equivalent to that of minimizing $V_2(T)$ over $[T_1, T_2]$.

By (29), the minimum will occur at T_1 if (30) applies there, and by (28) this condition reads

$$b_1\left(2N-2-b_1-2\sum_2^N b_1\right)+q_1\sum_2^N c_i b_i^2 \geqslant 0,$$

or equivalently

$$p_1 a_1(2N-2-2B_1+b_1)+\sum_2^N p_i a_i b_i \geqslant 0\cdot \tag{31}$$

If (31) does *not* hold, then $dV_2/dT=0$ occurs at a unique value T^*, where $T^*>T_1$, and the optimum occurs at T^* or T_2 according as $T^*\leqslant T_2$ or $T^*>T_2$. Specifically, from (28) and (22) we have

$$\left(D/T^*+N-1-\sum_2^r b_i\right)\left(N-1-D/T^*-\sum_2^r b_i\right)+\left(\sum_2^N c_i b_i^2\right)q_1=0,$$

or equivalently

$$\left(N-1-\sum_2^N b_i\right)^2-(D/T^*)^2+\left(\sum_2^N c_i b_i^2\right)q_1=0,$$

yielding

$$T^*=D\left\{\left(N-1-\sum_2^N b_i\right)^2+q_1\sum_2^N c_i b_i^2\right\}^{-1/2}\cdot \tag{32}$$

The solution process for PROBLEM 2 can be summarized as follows, assuming the ordering (10):

Step 1: Calculate the total dead time per cycle, D.

Step 2: Calculate the quantities $b_i=(s_i-a_i)/s_i$, their sum B_1, and the quantity

$$B_2=B_1-(b_1-p_2a_2b_1/p_1a_1)\cdot$$

If $B_1\leqslant N-1$, stop; the system is infeasible.

Step 3: If (31) holds, set $T=D/[B_1-(N-1)]$ and $G_i=Ta_i/s_i$ for all i.

Step 4: Otherwise, calculate $T_2=D/[B_2-(N-1)]$ and

$$T^*=D\left\{N-1-B_1+b_1)^2+(b_1/p_1a_1)\sum_2^N p_i a_i b_i\right\}^{-1/2}.$$

If $T^*>T_2$, set $T=T_2$ and

$$G_1=T_2(1+B_1-B_2-b_1).$$

If $T^*\leqslant T_2$, set $T=T^*$ and

$$G_1=T^*[B_1-b_1-(N-2)]-D.$$

In both cases, set $G_i=Ta_i/s_i$ for $i>1$.

5. References

[1] Rangarajan, R., and Oliver, R. M., Allocations of servicing periods that minimize average delay for N time-shared traffic streams, Transp. Sci. **1**, 74–80 (1967).

[2] Horn, W. A., Allocating service periods to minimize delay time, J. Res. Nat. Bur. Stand. (U.S.), **72B** (Math. Sci.), No. 3, 215–227 (1968).

(Paper 76B3&4–365)