r during 1959 and 1960 of
d in the second and third
nsed version was prepared
of International Studies,
olume of proceedings from

REDUCING LETTER DELAYS IN POST OFFICES†

Robert M. Oliver

University of California, Berkeley

and

Arych H. Samuel

Stanford Research Institute, Menlo Park, Calif.

(Received July 31, 1961)

This paper reports a number of mathematical models and experiments that have been designed for the analysis and evaluation of delays of first-class letter mail in a post office. The flow pattern of mail consists of a number of serial and parallel processing stages. A letter takes a particular path through this flow network, which depends on its final destination; consequently, the delay of letter mail depends on its address as well as the inventories of other mail and the processing rates met enroute. While mail flow into a post office may contain many random elements, it is generally the case that input rates are predictable and strongly time-dependent. Scheduling policies must take into account the peak flows that temporarily exceed available processing rates and, in addition, must observe certain specified restrictions on the cost of processing, sorting, and storage operations. The effect of various transportation facilities between processing stages and from one post office to another must also be considered. The mathematical analysis deals with the minimization of letter delay through a network of processing and storage stages where there are capacity restrictions on individual and/or serial and parallel stage combinations. Analytical and graphical procedures are developed and numerical results are reported. The paper also reports a series of full-scale experiments performed at one of the larger United States Post Offices where theoretical procedures and decision rules were applied and tested. Delay reductions for first-class letter mail are believed to be of the order of 25 per cent.

MATHEMATICAL models which describe postal operations seem to have been given little attention in scientific literature. While this paper is primarily concerned with a formal description and analysis of mail sorting operations, it also includes results that were obtained in the course of experiments at one of the larger United States Post Offices.

Summary and Objectives

This paper includes a discussion of flow problems that arise within the confines of mail-sorting and classifying operations, i.e., intra-post office

† Much of this research was performed by the authors while employed by Broadview Research Corporation, Burlingame, California. A fuller account of credits and acknowledgments is given in reference 1: "Reducing Letter Delays in Post Offices," Research Report 11, Operations Research Center, University of California, Berkeley (1961).

rather than inter-post office. Even then we will find that most of the mathematical models are motivated by, but not necessarily restricted, to the flow, scheduling, and storage aspects of first-class letter mail.

Although we do not include any major discussion of fully-automated mail recognition and processing equipment, much of the discussion and analysis, and especially that portion which deals with the effects of fixed dispatch times, is applicable to automated systems. It will be seen that the major portion of average letter delay is often caused by fixed dispatch times and that the effects of automatic high-speed sorting and processing equipment can be predicted by studying the infinite sorting rate cases of our mathematical formulas.

In a post office two objectives are of fundamental importance: to decrease costs and to avoid delays. Other goals can also be formulated, such as a reduction of the number of letters lost. But in the dual world of goals and restrictions, we chose to make the primary aim that of reducing average letter delay subject to restrictions on total cost of system operation. It is sometimes possible to reduce both delays and costs by introducing new operating rules and design criteria, but it should be pointed out that if both aims are pursued they must eventually become incompatible.

In recent years, the study of costs in a post office has been systematized to a great extent by modern cost accounting methods. By comparison, the effort to reduce delays has been less systematized, and operating rules can be found that have actually led to increases in delays. The expressed interest of postal departments in reducing letter delays and our personal interest in constructing mathematical models that could be formulated, solved, interpreted, and tested are the major reasons for emphasizing the analysis of letter delays.

The Sorting and Storage Problems

While post office personnel generally reserve the word 'delay' for excessive or needless delay above some nominal amount, we shall very simply refer to the delay of a letter as the difference between its arrival and its departure time.

This delay is a function of many variables: the arrival time, the inhomogeneous nature of the mail stream, storage capacities, the numbers and processing rates of men and machinery, the number of distinct classifications into which mail must be sorted and, perhaps most important of all, the times at which mail can leave a post office.

In general and sometimes vague ways, restrictions may be imposed on these same sorting and processing rates, budgets for manpower and machinery, acceptable working conditions, capacities of storage facilities, and flow capacities of mechanisms that transport mail from place to place within a post office.

Mail may arrive at a post office in bulk form or it may be packaged in bundles of various sizes and shapes. Almost all of it passes through a predetermined number of serial processing stages; in the early stages the processing operations prepare the mail for manual or automatic sorting schemes.

In these early stages mail is first dumped and weighed. These are followed by culling or sieving operations where non-first-class mail and other bulky items are rejected from the main stream. Letters are faced, i.e., they are oriented, for the stamp-cancelling operation; by this time

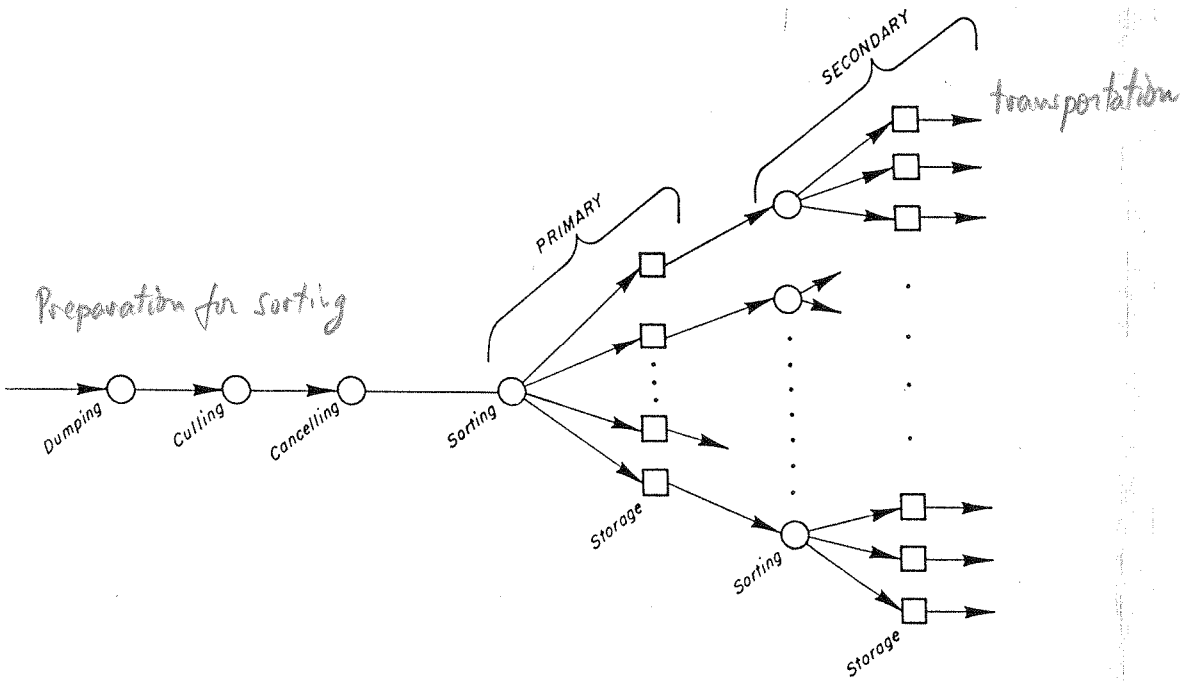


Fig. 1. Flow diagram.

some simple sorting operations may have been performed to separate high and low priorities such as regular and airmail. A schematic flow diagram is shown in Fig. 1 while Figs. 18 and 19 are diagrams of actual flow patterns in a United States Post Office.

The first major sorting operations come in the *Primary*, an area where letters are sorted into distinct categories or branches. These branches may correspond to states, regions, city zones or other types of geographical areas. The mail stream that flows through any particular *Primary* branch may undergo still further subdivision in a *Secondary*; this process of classification and branching may continue through *Tertiary* and other sorts.

At the end of the sorting process, letters in each branch are collected and packaged in a form suitable for transportation to successive post

offices or perhaps the ultimate addressee of the mail. The time that elapses while a letter proceeds through a post office is made up of essentially three parts: the time in service (sorting, dumping, culling, facing, etc.), the time waiting in queue for this service, and the time spent by a letter in storage waiting for transportation service. While the first two types of wait need little explanation, the storage delays are not obvious to the casual observer of a mail-processing system.

Before describing storage delays in some detail, it is worthwhile to consider the problems that arise when serial sorting and processing stages are encountered. An analysis of flow through predominantly serial mail-processing stages would appear to be an extension of aspects of the classical theory of stationary queues. There are several important exceptions. In the first place, the average mail input rates are not constant over time but are very sharply peaked as a result of many late afternoon business and private mailings. Variations in mail flows are not so much due to random fluctuations about a known mean rate as they are time-variations in the mean rate itself. Peak mail input rates propagate rapidly through successive processing operations; at first, these flows are easy to see and measure. In later stages of the sorting process, the peaks spread out and they may be difficult to locate and measure.

A major contributor to letter delay within a post office is the shape of the input flow rate. In fact, average arrival rates of letter mail may rise, peak, and fall off within a matter of a few hours. At the present time approximately 70 per cent of all letter mail enters a post office within a four-hour period.

In the theory of stationary, stochastic queues, arriving units are delayed even though the average servicing rate is greater than the average demand for service. These delays are due to the unpredictability of arrivals and services. There is a positive probability that very high arrival rates will occur over relatively short periods of time; since pre-servicing of items is not allowed, a queue can build up and, on the average, remain greater than zero. In a post office, on the other hand, long queues of mail are usually the result of an input rate that, although larger than the short-term processing capacity, is predictable from day to day.

In our studies of postal sorting and processing operations, emphasis is given to those situations where average arrival rates are predictable but greater than average servicing rates for part of the time. A more general treatment should certainly include stochastic effects. However, in response to the scheduling and allocation questions that arise in the serial production process, our analysis of the operational problems centered around deterministic queuing and storage models. Allowances for the relatively minor stochastic elements of mail flow have been made in the experiments

by introducing small corrections in the theoretical decision rules. One of the first decision problems that arises is that of scheduling manpower and machines to sort the peak flows as they propagate through the many processing stages.

There are several places where storage stages interrupt the main stream flows. One of the more obvious locations is at the end of the processing and sorting operations. When the mail is sorted in the Secondary, it is generally put into bags or pouches. The mail then waits in storage for the departure of busses, trains, ships, or planes that carry the mail to another post office. In the event that the address of a letter is local, a postman usually makes final delivery after a long storage interval.

The times at which mail inventories are released from storage stages are called 'dispatch times.' If a post office is located at or close to a major transportation facility, such as an airport or railroad terminal, the dispatch times are but little earlier than the departure times; if distant, the dispatch times must reflect the time needed to haul the mail to these major transportation centers.

Storage stages are evident throughout the Primary and Secondary sorting areas. These storage areas exist because: (*i*) the bulk handling of accumulated inventories is often less expensive than individual handling of letters and (*ii*) facilities for conveying letters from one sorting stage to the next may be temporarily unavailable or capacity-restricted.

Unfortunately the storage processes create many delay problems; hence optimal storage and release rules must be sought in addition to the processing and sorting rate assignments discussed above.

Mathematical Problems

The mathematical scheduling and storage problems discussed in this paper are related to a general theory of the flow of information and material over large networks. Where the objective function being optimized is linear and where flows are constrained by linear inequalities, one can resort to the theory and numerical algorithms of linear programming. Unfortunately, however, the criterion function and the restrictions upon the decision variables are not always linear ones; hence, one may have to call upon the calculus of variations.

A close look at the physical and mathematical problems of mail flow leads one to the conclusion that several inequality constraints may be operative at any one time and that the classical theory of the calculus of variations will not, by itself, suffice to establish optimal decisions as to how one should sort, store, and process mail.

The flow and scheduling problems can be formulated mathematically as the minimization of letter delay subject to inequality constraints on

various integral and derivative functions of the processing and sorting rates. One finds intervals where the decision variables, i.e., sorting rates, lie on the edge of the constraint region followed by intervals where the Euler equations are satisfied within the interior of the region.

Contents of Paper and Notation

In addition to this introductory section, the paper is divided into three parts. The second section studies the scheduling problems in a network that consists of a number of serial and parallel processing and sorting stages. The basic equations of letter delay are developed and a number of schedules that minimize average letter delay are obtained and illustrated.

In the third section attention is given to the effects of storage and some of the new scheduling and dispatching problems that naturally arise.

In the fourth section experimental tests are described and interpreted. Modifications of the theoretical solutions give rise to a set of scheduling and dispatch rules, which are then applied in a number of full-scale experiments at a large United States Post Office.

The final section is a summary of the major objectives achieved by this research.

Notation will be defined as it is introduced; however, it may help the reader to have a concise list of notation to which he can refer. The flow rate of mail as a function of time is shown by $\lambda(t)$. The sorting or processing rates will be denoted by $v(t)$ with a subscript $0, 1, \cdots, j$ referring to the jth stage. Cumulative flows are indicated by a capital letter; for example, $V_i(t)$ is the integral from 0 to t of $v_i(t)$. $\tau_i(t)$ denotes the delay of a letter entering the ith stage at time t. D is the average delay of a large group of letters. The small letters c and k, refer to upper bound restrictions on cost and capacity processing rates. The capital letter T will be used, with or without subscripts, to denote a dispatch time, i.e., the time at which contents of storage are released into a flow stream. The small letters r_i, s_i, t_i will refer to particular points in time. For example, the letters are often used to denote those times when capacity processing rates begin or end. The parameters α, β refer to fractions, less than one, which the flow of a branch bears to the major flow stream of mail.

MODELS OF SERIAL PROCESSING OPERATIONS

The Delay of a Letter

Consider first of all, the flow characteristics of two serial processing stages in Fig. 2. All of the mail that enters the system at a rate $\lambda(t)$ is

added to the queue, if any, at stage 1. The input rate to stage 2 is always the output rate of stage 1. We let $v_1(t)$ and $v_2(t)$ be the processing rates and $\tau_1(t)$ and $\tau_2(t)$ be the delays of letters that enter either stage 1 or 2 at time t.

A letter entering stage 1 at time t enters the second stage queue at time $t+\tau_1(t)$. A letter that enters stage 2 at t leaves at $t+\tau_2(t)$ and hence the total delay, $\tau(t)$, of a letter entering the two-stage system at time t is just the sum of both delays,

$$\tau(t)=\tau_1(t)+\tau_2[t+\tau_1(t)]. \tag{1}$$

These delays are functions of the processing rates, $v_1(t)$ and $v_2(t)$, and the shape of the input rate, $\lambda(t)$.

We may assume that letters are processed serially. All letters preceding the letter that entered at t must be processed by time $t+\tau_1(t)$. Hence, the solution for $\tau_1(t)$ equates cumulative flows through stage 1,

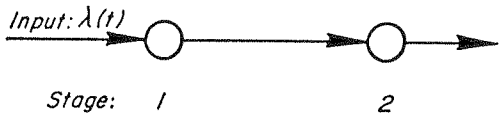


Fig. 2. Two serial processing stages.

$$\int_0^t \lambda(s)\,ds=\int_0^{t+\tau_1(t)} v_1(s)\,ds, \tag{2}$$

and similarly for the delay $\tau_2(t)$ at stage 2.

If we let $\Lambda(t)$, $V_1(t)$, $V_2(t)$ be the cumulative flows, (2) can be written,

$$\Lambda(t)=V_1[t+\tau_1(t)], \tag{3a}$$

$$V_1(t)=V_2[t+\tau_2(t)]. \tag{3b}$$

The solution for $\tau(t)$ is found by substituting equations (1) and (3a) into (3b) to get,

$$\Lambda(t)=V_2[t+\tau(t)]. \tag{4}$$

The continuous nature of the mail stream and the first-come first-serve priority rule make the calculation of the delay of a letter a simple one. If the processing rate at a stage were constant, the delay of a letter entering at time t would simply be the unprocessed inventory (if any) divided by this constant sorting rate. On the other hand, if neither input rates nor processing rates are constant, the best we can do is find an implicit solution for the delay of a letter entering the queue at time t.

If, for example, the processing rate at stage 1 is unrestricted but $\lambda(t)$

is greater than the constant processing rate, k, at stage 2 for a short period of time, the delay of a letter in the two-stage network is:

$$\tau(t) = k^{-1}[\Lambda(t) - \Lambda(s_1)] - (t - s_1), \qquad (t \in S) \quad (5)$$

where $S = (s_1, s_2)$ is the interval where inventories and letter delays are positive. Figure 4 is a plot of individual letter delay as a function of its arrival time when the input flows of Fig. 3 and the indicated values of the constant processing rate, k, are used. Case (I) corresponds to an early sharp peak, Case (II) to a late and flatter peak in the input rate. We

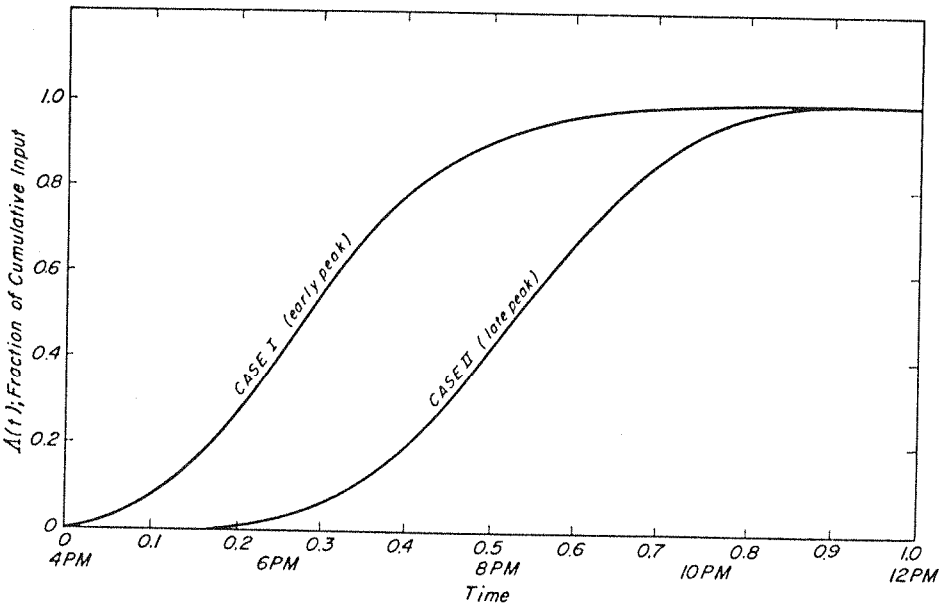


Fig. 3. Cumulative input as a function of time.

will frequently refer to these two cases in our mathematical models; specifically, many of our numerical results are derived from the cumulative flow curves of Fig. 3.

Average Letter Delay

Average letter delay, D, is found by integrating the product of the arrival rate of letters with the delay of an individual letter and dividing this expression by the total volume of letters. Two normalizations improve notation and greatly reduce the number of algebraic manipulations. The first one normalizes the time scale to (0, 1). The second normalization requires that the total volume of letters equal one, i.e., $\Lambda(0) = 0$, $\Lambda(1) = 1$†. Since the delay of all letters arriving between t and $t+dt$

† With these normalizations the dimensionless values of D (average delay), $\tau(t)$ (the delay of a letter), need only be multiplied by the actual length of the interval to obtain the real value of D, $\tau(t)$.

is $\tau(t)$ we can write the average delay as

$$D=\int_0^1 \lambda(t)\tau(t)\,dt. \tag{6}$$

If processing rates and inventories are nonnegative at each step,

$$v_1(t),\ v_2(t)\geqq 0, \tag{7a}$$

$$\Lambda(t)\geqq V_1(t)\geqq V_2(t), \tag{7b}$$

the solutions of equations (3) and (4) lead to nonnegative letter delays and $D\geqq 0$. If processing rates are unrestricted from above, we can easily satisfy the equality restrictions in equation (7b). The solutions for

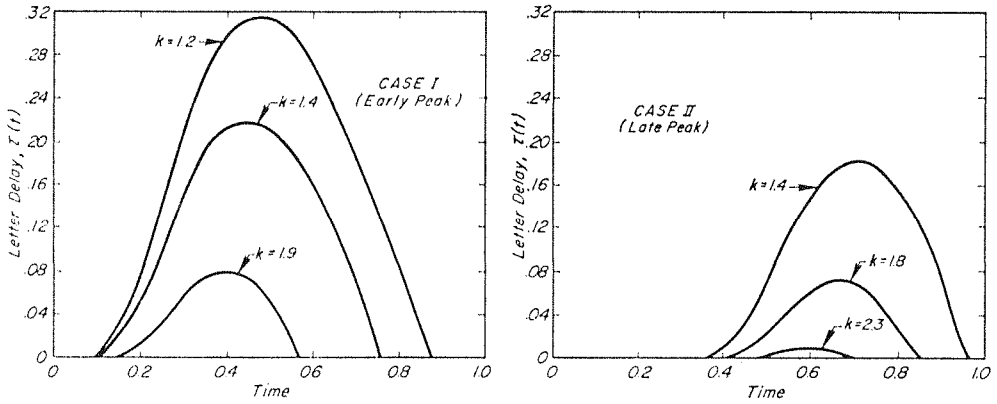


Fig. 4. Letter delay as a function of time (constant processing rate, k).

$\tau_1(t)$, $\tau_2(t)$, and $\tau(t)$ in equations (2), (3), and (4) can then be made zero; hence the average delay of letters traveling through the two-stage system of Fig. 2 can also be made zero. However, restrictions of the sum of processing rates,

$$v_1(t)+v_2(t)\leqq k, \tag{8}$$

and, more generally, of the form

$$v_1(t)+av_2(t)\leqq k(t) \tag{9}$$

play an important part when the total sorting rate at time t is capacity restricted. Manpower pools, though interchangeable with different efficiencies between processing stages, may be restricted in total size by the availability of skilled personnel.†

† We are indebted to one of the referees for pointing out that we do not treat those cases where one individual is more efficient than all other individuals at the same stage. It happens that we did not find this to be the case, perhaps because of the uniformity imposed by civil service examinations.

In addition, there may be capacity restrictions on individual stages, such as the maximum seating capacity around tables where mail is faced, culled, and sorted. Another example is the maximum flow rate of a conveyor. Even semi-automatic and fully-automatic machines will be limited in their sorting rates. In these cases an upper bound can be placed on the processing rates:

$$v_1(t) \leqq c_1, \qquad v_2(t) \leqq c_2.$$

Before we obtain solutions that minimize average letter delay under these, and other, more severe restrictions, we want to sketch briefly the intuitive procedures that had been adopted by postal personnel prior to the research. We do this before the mathematical treatment because, once the solutions are obtained, they appear obvious in retrospect. They were not initially obvious to either author and moreover, the intuitive procedures used by postal personnel were neither uniform nor clearly formulated. If any policy was representative of the majority of rules used by foremen to schedule processing rates, it was that one should at all times match processing rates with input rates at each stage and at the same time keep letter delays in the first stage of the process as small as possible. As the peak mail flows through the first stage subsided, a larger fraction of the processing capacity was then assigned to the second stage.

The minimization of D in equation (6) is a simple variational problem: find that (feasible) processing rate assignment $v_1(t)$, $v_2(t)$ that makes D a minimum while satisfying the inequality constraints of (7) and (8). In the absence of derivatives of $v(t)$ in the expression for delay of a letter, only two analytical difficulties may arise. First of all the delay of a letter is an implicit function of the sorting rate and, with certain exceptions, one can make no explicit substitution of processing rates or cumulative flows into equation (6). Secondly, the inequality constraints (7) and (8) will lead to optimal processing rate solutions that, in one interval of time, will lie on the edge of one constraint; in a succeeding interval the sorting rate will switch to another constraint. While the switch-over process may be simple conceptually, it may be difficult to find algebraic expressions for the exact point of transition.

It seems intuitively clear that we increase processing rates at each stage whenever we can—that is to say, until one of the inequalities in equations (7) and (8) becomes a strict equality. It is fortunate that in the majority of our scheduling problems one need only divide the interval (0, 1) into two types of subintervals S and $\bar{S}$. In $\bar{S}$ capacity processing rates are not restrictive and cumulative input flows equal cumulative amounts processed: $v_1(t)+v_2(t)<k$ and $\Lambda(t)=V_1(t)=V_2(t)$. In S processing rates are restricted and equal to one another while cumulative input flows are greater than cumulative amounts processed: $\Lambda(t)>V_1(t)=V_2(t)$.

The Region $\bar{S}$

Assume that $\lambda(t)$ is
In equation (6) D is alw
delay of a letter entering
zero. If, in the optimal
is ever zero we will denot

$$\tau(t$$

where the star (*) denot
that if equation (10) hol

Since $v_1^*(t)+v_2^*(t) \leqq k$ fr

is a necessary condition
(11). That is to say, t
half the processing rate

The Region S

We now consider inte
at time t is greater than
sorting rates is a strict
average delay is obtain
equal.

If $\tau(t)$ is greater than
can build up in front o
actual location of the i
input flows exceed flows
one of the three followin
$V_2(t)$:

Since positive letter del
creasing the processing r
dition is that processing

Since letter delays in $\bar{S}$
written,

The Region $\bar{S}$

Assume that $\lambda(t)$ is always greater than zero in the interval (0, 1). In equation (6) D is always reduced by negative variations in $\tau(t)$, the delay of a letter entering at time t. The lowest feasible value of $\tau(t)$ is zero. If, in the optimal program, the delay of a letter arriving at time t is ever zero we will denote those regions in time by $\bar{S}$; that is

$$\tau(t)=0;\ \Lambda(t)=V_1^*(t)=V_2^*(t), \qquad (t\in\bar{S}) \quad (10)$$

where the star (*) denotes the optimal processing rates. It is also true that if equation (10) holds in $\bar{S}$ then

$$\lambda(t)=v_1^*(t)=v_2^*(t). \qquad (t\in\bar{S}) \quad (11)$$

Since $v_1^*(t)+v_2^*(t)\leqq k$ from equation (8), we find that

$$\lambda(t)\leqq k/2 \qquad (t\in\bar{S}) \quad (12)$$

is a necessary condition for the optimal solutions to be given by equation (11). That is to say, the input rate must be less than or equal to one-half the processing rate capacity.

The Region S

We now consider intervals S where $\tau(t)$, the delay of a letter entering at time t is greater than zero. We show that (*i*) the sum restriction on sorting rates is a strict equality in this region and (*ii*) that the minimum average delay is obtained when the processing rates at each stage are equal.

If $\tau(t)$ is greater than zero in the optimal program, an inventory of mail can build up in front of the first and/or second stage. Wherever the actual location of the inventory, we know that $\Lambda(t)>V_2^*(t)$, i.e., that input flows exceed flows processed at stage 2. In Fig. 5 we see that any one of the three following solutions for $V_1(t)$, $V_2(t)$ is feasible when $\Lambda(t)>V_2(t)$:

$$\Lambda(t)>V_1(t)=V_2(t), \qquad (13a)$$

$$\Lambda(t)>V_1(t)>V_2(t), \qquad [t\in S=(s_1, s_2)] \quad (13b)$$

$$\Lambda(t)=V_1(t)>V_2(t). \qquad (13c)$$

Since positive letter delays at each stage can always be reduced by increasing the processing rate at that stage, a necessary but insufficient condition is that processing rates be capacity constrained in S,

$$v_1(t)+v_2(t)=k. \qquad (t\in S) \quad (14)$$

Since letter delays in $\bar{S}$ are zero, variations in average letter delay can be written,

$$\delta D = \int_S \lambda(t)\delta\tau(t)\, dt, \tag{15}$$

where we use the notation $\delta x(t)$ to denote a variation in $x(t)$ holding t constant. In equation (4) the delay of a letter through two processing stages is related to the sorting rate at the second stage as an implicit solution of an integral equation. The variation in $\tau(t)$ due to variations in $v_2(t)$ is obtained by straightforward differentiation of equation (4) and

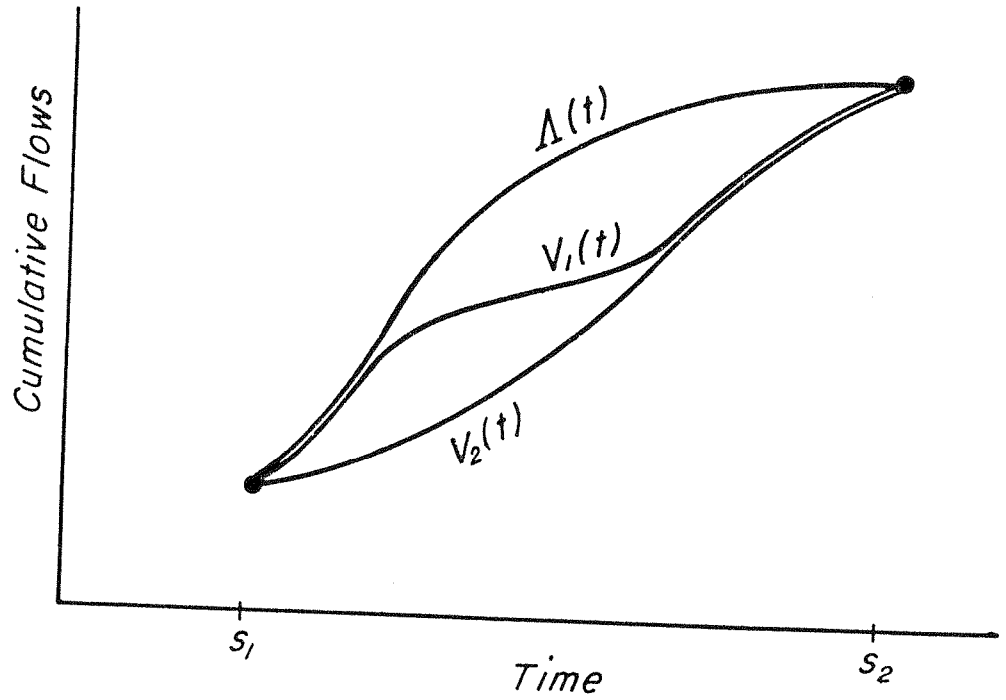


Fig. 5. Feasible cumulative processing rates.

use of the formula,

$$\delta\left\{\int_0^{y(t)} x(s)\, ds\right\} = \int_0^{y(t)} \delta x(s)\, ds + x[y(t)]\, \delta y(t). \tag{16}$$

We obtain

$$\delta\tau(t) = -\delta V_2[t+\tau(t)]/v_2[t+\tau(t)] = -\delta V_2(r)/v_2(r), \tag{17}\dagger$$

where r is the time of exit of a letter from the second processing stage. Equation (17) shows that variations in the processing rate in the interval $[0, t+\tau(t)]$ may affect the delay of the letter entering at time t; scheduling small processing rates over a long period of time before the arrival of a letter may be as effective as scheduling large processing rates just after its arrival. The derivative of equation (4) with respect to time is

† When $v_2(r)>0$, i.e., positive processing rates.

$$\lambda(t) = v_2[t+\tau(t)][1+(d\tau/dt)], \tag{18}$$

and if we substitute this result and equation (17) into (15) we obtain

$$\delta D = -\int_S \delta V_2(r)\, dr. \tag{19}$$

Hence, positive variations in cumulative flows at the second stage decrease D. Since positive variations in $V_2(t)$ equal negative variations in $V_1(t)$ we obtain an optimal program by decreasing $V_1(t)$ and increasing $V_2(t)$ until cumulative flows out of the first stage equal cumulative flows out of the second stage. Consequently, in S the optimal processing rates must also be equal and from (14) we obtain

$$v_1^*(t) = v_2^*(t) = k/2. \qquad (t \in S) \tag{20}$$

In the optimal program only (13a) holds; inventories and delays are concentrated at the first stage and there are no delays and no inventories at the second stage. Even when the capacity varies with time, as in equation (9), we again find that delays and inventories are only concentrated at the first stage. In this case

$$v_1^*(t) = v_2^*(t) = (1+a)^{-1}k(t). \qquad (t \in S) \tag{21}$$

Up to this point we have only found the optimal schedules within S. We have not yet found the optimal location of S; this follows by studying the variation of D at the end points of S.

Corner Conditions and Minimum Average Delay

In addition to the condition that sorting rates be equal, the boundary (transversality) conditions at the variable end points are

$$\lambda(t)\tau(t) = 0. \qquad (t = s_1,\ s_2) \tag{22}$$

In other words, if the input rate is positive the letter delay at the corner points must vanish. This requirement is identical to the equality of cumulative input flows and cumulative output flows from the second stage.

Let us assume that $\lambda(t)$ starts at zero, increases, attains a single maximum, and falls off to zero before the end of the interval (0, 1). If the maximum input rate is greater than one-half the available processing rate, and if the total processing capacity is greater than the total mail volume, the interval (0, 1) will be divided into three subintervals, $\bar{S} = (0,\ s_1)$, $S = (s_1,\ s_2)$, and $\bar{S} = (s_2,\ 1)$.

As $\lambda(t)$ increases from zero, t will reach the point s_1 where for the first time the capacity restriction of equation (8) becomes a strict equality; mathematically, s_1 is the smaller root of

$$\lambda(s_1) - \tfrac{1}{2}k = 0. \tag{23}$$

For the two-stage case this corresponds to the simple graphical solution (Fig. 6b) where the tangent of the cumulative input flows is first equal to $\frac{1}{2}k$. As $\lambda(t)$ continues to increase ($t > s_1$) the inventory of unprocessed mail in front of the first stage will also increase and reach a maxi-

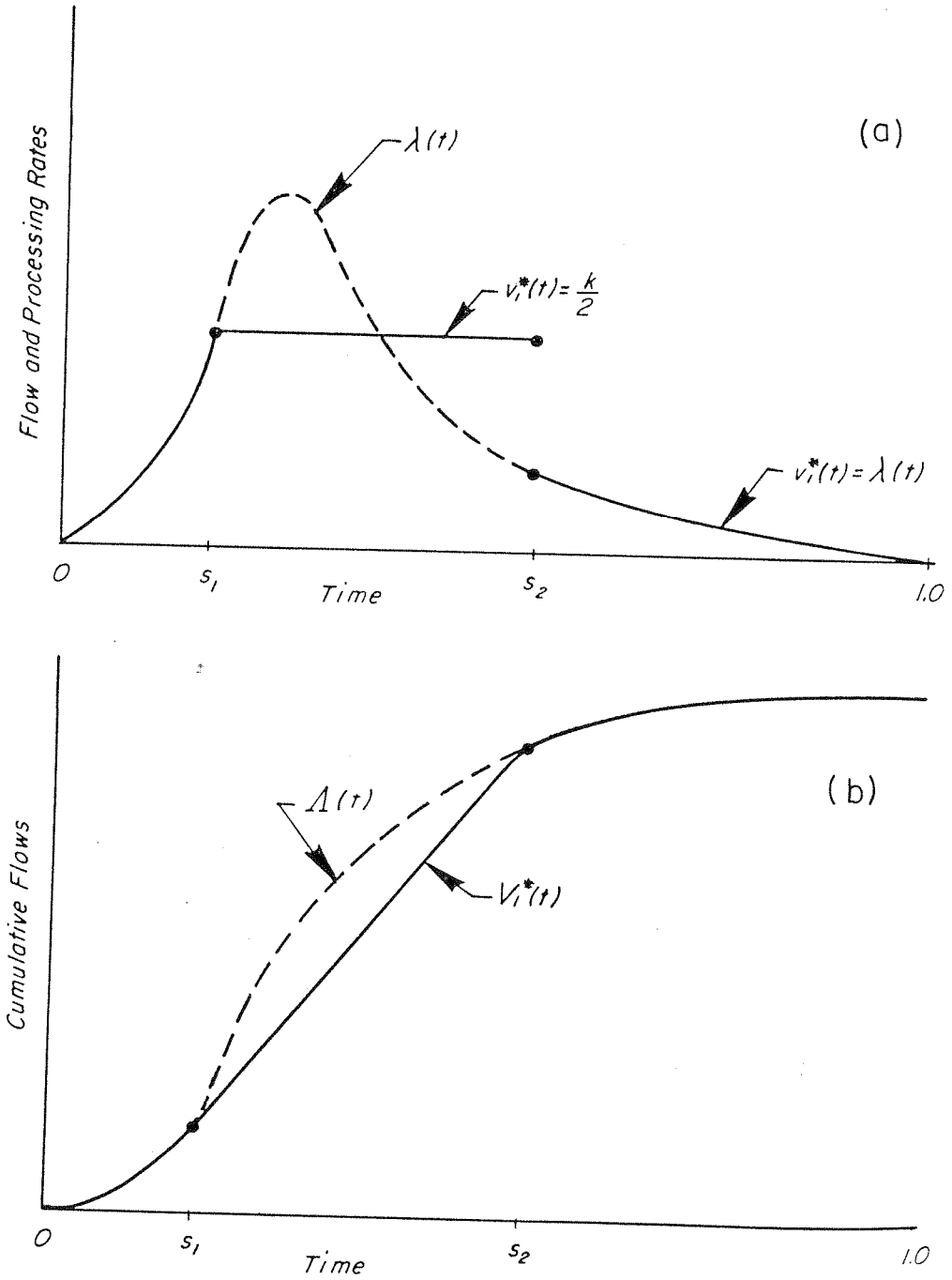


Fig. 6. Optimal processing rates as a function of time.

mum at a point in tim
inventory will always p
mum value and will b
in the capacity constra
the maximum processin

This expression and th
of a letter in the first
the optimal processing
are discontinuous at s_2,

$$v^*($$

With the optimal a
of letter mail will alwa
of each letter arriving
and zero for $t \in \bar{S}$. Usi
minimum average delay

$$D^* =$$

D^* can be expressed in
stitutes equations (5) a

$$D^* = \tfrac{1}{2}k$$

$$= \tfrac{1}{4}k$$

N Serial Stages

When there are N se
becomes,

Again, the values of a_i
manpower or machines
results that have been o
difficulty, to N stages.

and nonnegative invent
$V_i(t) \geqq V_{i+1}(t)$. We fi
variations in $V_N(t)$; he
last stage until either th

mum at a point in time that is the larger root of equation (23). The inventory will always peak after the input flow rate has reached its maximum value and will become zero when the flows that have accumulated in the capacity constrained interval (s_1, s_2) are completely processed by the maximum processing rates, i.e., when

$$\Lambda(s_2) - \Lambda(s_1) = \tfrac{1}{2} k(s_2 - s_1). \tag{24}$$

This expression and the solution for s_2 are obtained by setting the delay of a letter in the first stage equal to zero at the end-point s_2. Whereas the optimal processing rates are continuous at the corner-point s_1 they are discontinuous at s_2,

$$v^*(s_2^-) = \tfrac{1}{2} k; \qquad v^*(s_2^+) = \lambda(s_2). \tag{25}$$

With the optimal assignments of equations (11) and (20) the delays of letter mail will always be concentrated at the first stage and the delay of each letter arriving at time t will be given by equation (5) for $t \in S$ and zero for $t \in \bar{S}$. Using this optimal assignment of processing rates the minimum average delay becomes

$$D^* = \int_S \lambda(t)\tau(t)\, dt = \int_{s_1}^{s_2} \lambda(t)\tau_1(t)\, dt. \tag{26}$$

D^* can be expressed in terms of the corner-points s_1 and s_2 when one substitutes equations (5) and (24) into (26)

$$\begin{aligned} D^* &= \tfrac{1}{2} k(s_2 - s_1)^2 - \Lambda(s_2)(s_2 - s_1) + \int_{s_1}^{s_2} \Lambda(t)\, dt \\ &= \tfrac{1}{4} k(s_2^2 - s_1^2) - \int_{s_1}^{s_2} t\lambda(t)\, dt. \end{aligned} \tag{27}$$

N Serial Stages

When there are N serial stages the inequality restriction of equation (9) becomes,

$$\sum_i a_i v_i(t) \leqq k. \tag{28}$$

Again, the values of a_i not equal to unity indicate different efficiencies of manpower or machines that can be interchanged between stages. The results that have been obtained for the two-stage case generalize, without difficulty, to N stages. The total delay through N stages is the solution of

$$\Lambda(t) = V_N[t + \tau(t)] \tag{29}$$

and nonnegative inventory restrictions at the ith stage are of the form, $V_i(t) \geqq V_{i+1}(t)$. We find that the average delay decreases with positive variations in $V_N(t)$; hence, available sorting capacity is assigned to the last stage until either the nonnegative inventory restriction or the capacity

sorting restrictions of equation (28) are violated. Again we find that the time-axis can be subdivided into two intervals: $\bar{S}$ and S. In the former, the delay through the network is zero and the nonnegative restriction on inventories would be violated if processing rates were increased further. In the latter, letter delay is positive and the equality of (28) holds. If we consider the case where input flow rates increase from zero, peak, and fall off, delays and inventories (if any) are always concentrated at the first stage by making processing rates equal at each stage, i.e., the optimal program reads

$$\begin{aligned} v_i^*(t) &= k\left(\textstyle\sum_i a_i\right)^{-1} && (t \in S) \\ &= \lambda(t). && (t \in \bar{S}) \end{aligned} \tag{30}$$

The beginning and end points of the capacity constrained interval are obtained by substituting $k\left(\sum a_i\right)^{-1}$ for $\frac{1}{2}k$ in equations (23) and (24). The same substitution must be made in equation (27) for D^*.

When $\lambda(t)$ consists of multiple peaks there is no difficulty in extending these results. Intervals of S and $\bar{S}$ alternate. In the case where inventories have accumulated as a result of one peak in $\lambda(t)$ the capacity constrained interval may extend beyond the second, third, or later peak. On the other hand if processing rates are large enough, the inventories arising from the first peak will be reduced to zero before the second peak. The solutions of the corner-points and the minimum average delay are simple generalizations of our earlier results.

Initial Inventories

Until this point we have been concerned with the simple physical situation where initial inventories are zero at time zero. If we add the initial inventories I_1 and I_2 to stage 1 and 2 at time zero, the solutions for $\tau_1(t)$, $\tau_2(t)$, and $\tau(t)$ are obtained by adding I_1, I_2, and I_1+I_2 to the left-hand-sides of equations (3) and (4). Equation (7b) is replaced by

$$I_1+\Lambda(t) \geqq V_1(t) \geqq V_2(t)-I_2. \tag{31}$$

In an interval $\bar{S}$ where $\tau(t)=0$ the strict equalities hold and $v_1^*(t)=v_2^*(t)=\lambda(t)$. In an interval S where $\tau(t)>0$ we can no longer force cumulative amounts processed at stage 1 to equal cumulative amounts processed at stage 2. Nevertheless we still find that variations in average delay are proportional to negative variations in the cumulative processing rates at stage 2 when inventories at that stage are positive; hence one initially assigns all resources to the second stage. Once the inventories in the second stage are reduced to zero the optimal program is identical to the assignment of equation (20). To illustrate the optimal schedules in more detail we examine two cases: I and II. In the former, the single peak in

the input rate comes early
In Case II, initial inventor
In both Cases I and II

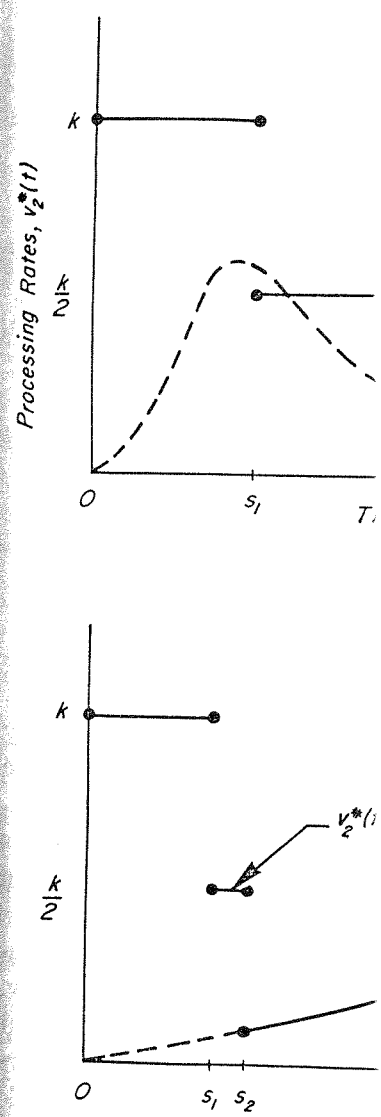


Fig. 7. P

processing rate k at stage 2.
stage is first reduced to zer
and the peak flow rate is early
root of equation (23); the ca
stages from s_1 to s_2 where s
for the left-hand-side of equa

ed. Again we find that the
s: $\bar{S}$ and S. In the former,
e nonnegative restriction on
ates were increased further,
equality of (28) holds. If
crease from zero, peak, and
always concentrated at the
each stage, i.e., the optimal

$$\begin{matrix}(t\in S)\\(t\in\bar{S})\end{matrix} \qquad (30)$$

y co: rained interval are
equations (23) and (24),
n (27) for D^*.
s no difficulty in extending
. In the case where in-
peak in $\lambda(t)$ the capacity
ond, third, or later peak.
e enough, the inventories
o before the second peak.
imum average delay are

the simple physical situ-
ro. If we add the initial
o, the solutions for $\tau_1(t)$,
I_1+ to the left-hand-
s repl ed by

$$\text{2.} \qquad (31)$$

hold and $v_1^*(t)=v_2^*(t)=$
longer force cumulative
e amounts processed at
ns in average delay are
tive processing rates at
ive; hence one initially
the inventories in the
ram is identical to the
imal schedules in more
mer, the single peak in

the input rate comes early and initial inventories at both stages are large. In Case II, initial inventories are small and $\lambda(t)$ has a late peak.

In both Cases I and II we start at time zero by assigning the maximum

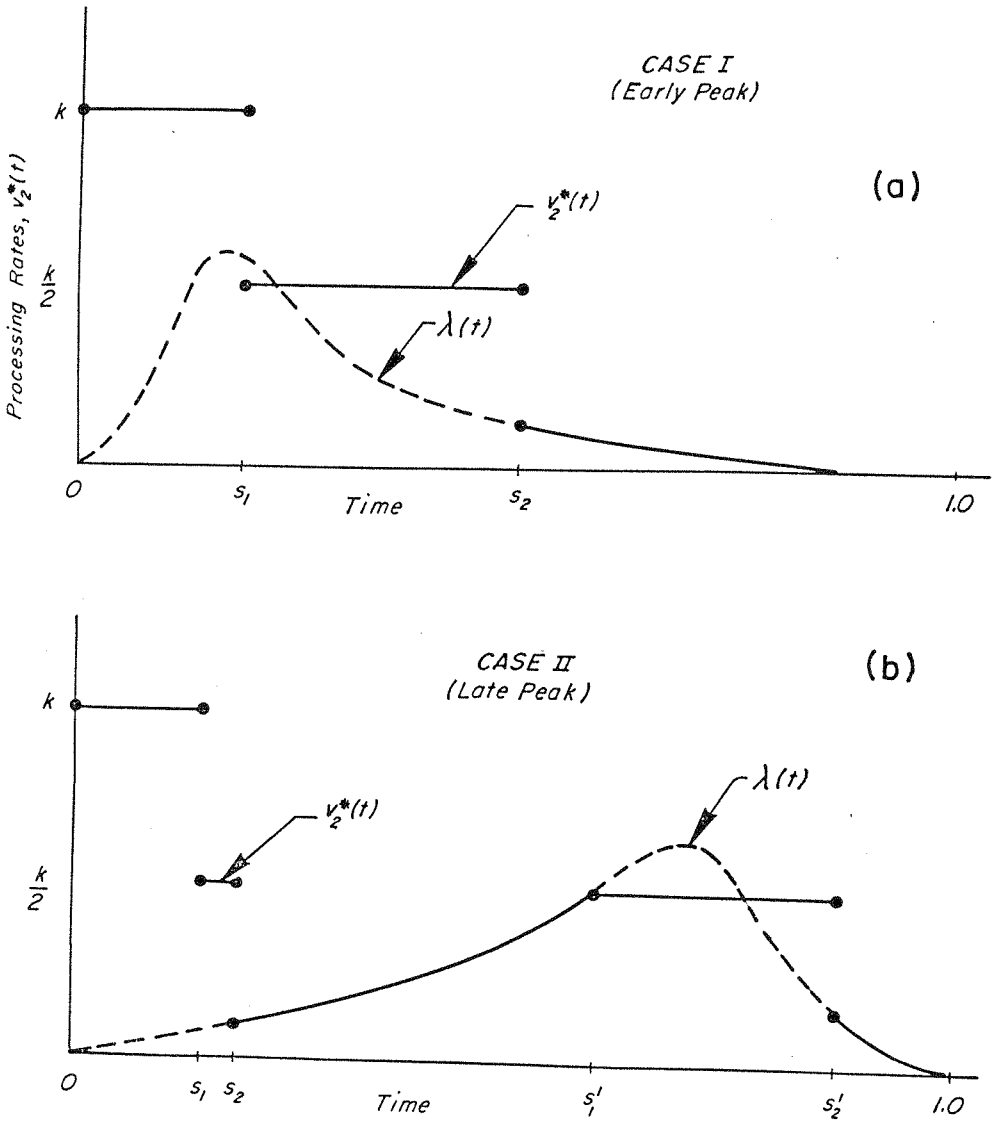


Fig. 7. Processing initial inventories.

processing rate k at stage 2. The time when the inventory in the second stage is first reduced to zero is $s_1=I_2/k$. If initial inventories are large and the peak flow rate is early this point in time will come after the smallest root of equation (23); the capacity sorting rate will be split between both stages from s_1 to s_2 where s_2 is now obtained by substituting $I_1+\Lambda(s_2)$ for the left-hand-side of equation (24) (Fig. 7a).

In Case II (Fig. 7b) the peak in the input rate comes after s_2, the time at which inventories in both stages have been reduced to zero. As a result capacity processing rates are in effect during $S' = (s_1', s_2')$ as well as $S = (s_1, s_2)$.

As one might expect, the N stage case is, in principle, no different than the solutions we have already outlined. The optimal assignments can be stated in simple terms:

Start at the last, Nth, stage and assign capacity processing rates to that stage until all inventory is processed. When the inventory in the Nth stage becomes zero, split the capacity processing rates such that flow out of the $(N-1)$st stage equals flow out of the Nth stage; i.e., maintain zero inventory at the Nth stage. Continue until inventory at stage $N-1$ is completely processed; split the capacity processing rates between the Nth, $(N-1)$st, and $(N-2)$nd stage. Continue in this fashion until inventories at all stages are reduced to zero. Once inventories are reduced to zero, the optimal assignments are,

$$v_1^* = \lambda(t) \qquad (t \in \bar{S})$$
$$= k\left(\sum_i a_i\right)^{-1}. \qquad (t \in S)$$

Sorting and Branching

Mail is processed through a number of stages. Most sorting operations are toward the end, and only a few simple sorts are made at the beginning. Physically, sorting differs from other processing operations, such as dumping and weighing, in that the letters are observed piece by piece and a routing decision is made. Depending on whether or not the sorter is aided by automatic equipment, the number of flow routes branching from a storage stage may vary from two to several thousand. Stages toward the end also include special features of storage and interstage transportation to be analyzed later.

Here we take up simple sorts with few branches and with the predominant delay in queue—the result of flows temporarily in excess of capacity. Binary sorts are important since they handle large volumes or separate out high priority mail such as special delivery. Other examples are letters with insufficient postage or oversized packages. Consider first the simple two-branch case (Fig. 8).

Denote the input rate by $\lambda(t)$; in practice it may be the output rate of a previous stage. The sorting rate is $v_0(t)$ at stage 0; a *constant* fraction α flows into stage 1 with a processing rate $v_1(t)$ and the remaining fraction $1 - \alpha$ flows into stage 2. The average delay of the three-stage system in Fig. 8 includes the delays of letters through stage 0 and stage 1 as well as those through the lower branch. We denote these branch delays by $\tau_\alpha(t)$ and $\tau_{1-\alpha}(t)$.

The average system delay is obtained by multiplying the fraction of letters that go through each branch by their respective delays,

$$D=\int_0^1 [\alpha\tau_\alpha(t)+(1-\alpha)\tau_{1-\alpha}(t)]\lambda(t)\,dt. \tag{32}$$

Nonnegative inventory restrictions apply, as before. Since fractions α and $1-\alpha$ flow into the two branches, these become

$$\Lambda(t)\geqq V_0(t), \tag{33a}$$

$$\alpha V_0(t)\geqq V_1(t), \tag{33b}$$

$$(1-\alpha)V_0(t)\geqq V_2(t). \tag{33c}$$

Basically two additional types of restrictions are encountered in postal operations. Either the sum of processing rates at stages 1 and 2 are

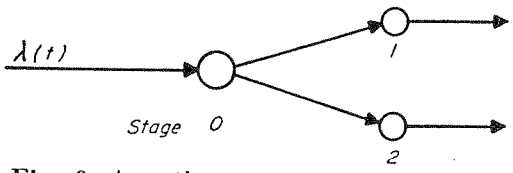


Fig. 8. A sorting stage with branching.

restricted and $v_0(t)$ is not restricted,

$$v_1(t)+av_2(t)\leqq k, \tag{34}$$

or the sum of processing rates at stages 1 and 2 and the sorting rate at stage 0 are capacity restricted,

$$v_0(t)+a_1\,v_1(t)+a_2\,v_2(t)\leqq k. \tag{35}$$

Consider the policies which minimize equation (32) subject to the nonnegative inventory restrictions of equation (33) and the restriction of equation (34) on processing rates. Again, the notation S and $\bar{S}$ refers to those intervals where letter delays are nonzero and zero respectively. The sorting rate at stage 0 always equals the arrival rate, $\lambda(t)$, and in $\bar{S}$ the processing rates at stages 1 and 2 are $\alpha\lambda(t)$ and $(1-\alpha)\lambda(t)$. The interval S begins when the sum of processing rates equals the capacity†

† Again we assume that $\lambda(t)$ starts from zero, has a single peak, and returns to zero. We also assume that $\Lambda(1)\leqq k$, i.e., that all flows can eventually be processed in the interval.

$$\lambda(s_1)-k(\alpha-a\alpha+a)^{-1}=0. \tag{36}$$

For s_2 to be the end point of a capacity constrained interval S, the total amount processed in S must equal the cumulative flow into stages 1

and 2 in that same interval. Since the sum of processing rates and the sum of cumulative flows must satisfy equation (34) and its integral, and since inventories must be zero at the corner points,

$$V_2(s_1)=(1-\alpha)\Lambda(s_1);\ V_2(s_2)=(1-\alpha)\Lambda(s_2). \tag{37}$$

From these requirements and equation (34) we get

$$\Lambda(s_2)-\Lambda(s_1)=k(s_2-s_1)(a-a\alpha+\alpha)^{-1}. \tag{38}$$

This solution for s_2 is independent of the actual assignment of processing rates at stages 1 or 2 and is only a function of s_1, α, a, k and the shape of the input flows $\Lambda(t)$. Since s_1 is the smaller root of equation (36), s_2 is only a function of a, α, k, and $\Lambda(t)$.

From the large set of schedules that satisfy equation (33) we have only considered a smaller set as candidates for an optimal scheduling policy. These are the schedules that make s_2-s_1 as small as possible by assigning capacity processing rates and reducing inventories to zero as quickly as possible. Even this assignment of capacity processing rates does not guarantee a minimum average delay schedule when the constant, a, differs from unity. Until we look at the effect of variations in $V_1(t)$ or $V_2(t)$ within S it is not clear what fraction of capacity processing rates should be assigned to stages 1 or 2. By methods we have already used it can be shown that if a is greater than 1 in equation (34), positive variations in the processing rates at stage 1 reduce the average delay of letter mail.† To minimize D we increase the processing rate at stage 1 until either equation (33b) becomes a strict equality or the processing rate at stage 2 becomes zero.

If k is larger than the maximum flow rate in the top branch, but less than $(a+\alpha-a\alpha)$ times the peak flow rate into stage 0, equation (33b) will be the inequality that becomes a strict equality in S. No delays will occur in the top branch and the delays in the bottom branch will be caused by the discrepancy between the flow rate, $(1-\alpha)\lambda(t)$ into the bottom branch and the processing rate $v_2^*(t)=[k-\alpha\lambda(t)]/a$. The optimal processing rate $v_2^*(t)$, is plotted in Fig. 9.

On the other hand if k is less than the peak flow rate into the top branch the optimal policies, $v_1^*(t)$ and $v_2^*(t)$, have a curious behavior in that the interval S will contain a subinterval R during which time the processing rate at stage 2 is zero and the delays in the top branch are positive, (Fig. 10). Figure 9 corresponds to the case where R is empty. For the case where R is not empty, let r_1 be the first time when the flow rate into stage 1 equals the capacity processing rate, i.e., the smaller real

† $a>1$ corresponds to the case where stage 1 is more 'efficient' than stage 2.

root of

The solution for r_2 is then
(24) and r_1 and r_2 for s_1
responding expressions
reference 1.

MODELS OF S

As WE mentioned in th
preceding section would

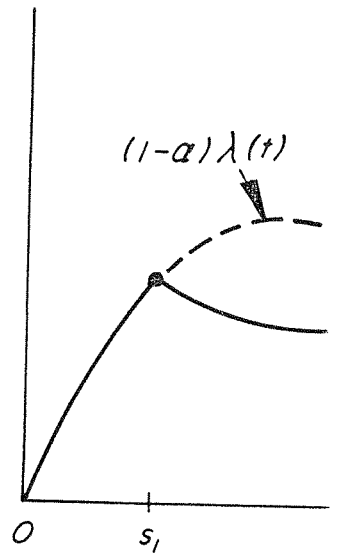


Fig.

but for the fact that stor
processing stage. The eff
terruptions to an otherwis

The reasons for mail s
first place a storage area
mon destination. Mail h
region) may have been pr
of the same post office.
portation and handling c
realized.

If a post office containe
there would be a need for
end of the process, i.e., aft

root of

$$\alpha\lambda(t)-k=0. \tag{39}$$

The solution for r_2 is then obtained by substituting k/α for $\frac{1}{2}k$ in equation (24) and r_1 and r_2 for s_1 and s_2 respectively. These solutions and the corresponding expressions for minimum average delay are summarized in reference 1.

MODELS OF SORTING AND STORAGE OPERATIONS

As we mentioned in the introduction, the mathematical models of the preceding section would describe intra-post-office scheduling problems

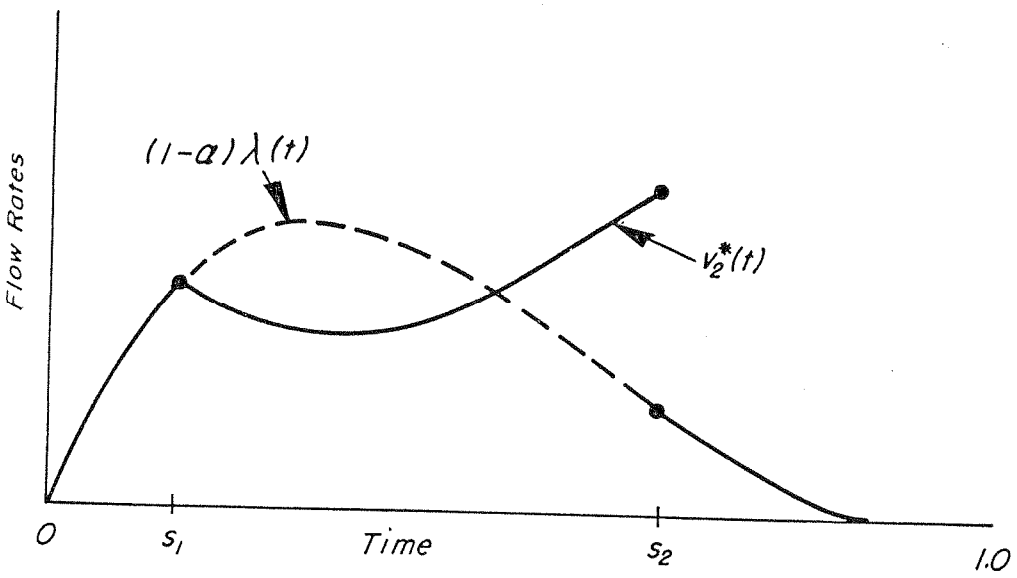


Fig. 9. Optimal processing rates.

but for the fact that storage stages often precede or follow a sorting or processing stage. The effect of a storage stage is that of providing interruptions to an otherwise continuous flow of letter mail.

The reasons for mail storage within a post office are twofold. In the first place a storage area provides a collection point for mail with a common destination. Mail having the same address (state, county, district, region) may have been processed and sorted in distinct and distant parts of the same post office. Secondly, storage areas reduce certain transportation and handling costs where economies in bulk service can be realized.

If a post office contained only Primary sorting areas it is doubtful that there would be a need for storage stages other than those located at the end of the process, i.e., after the major sorting operations. With a Sec-

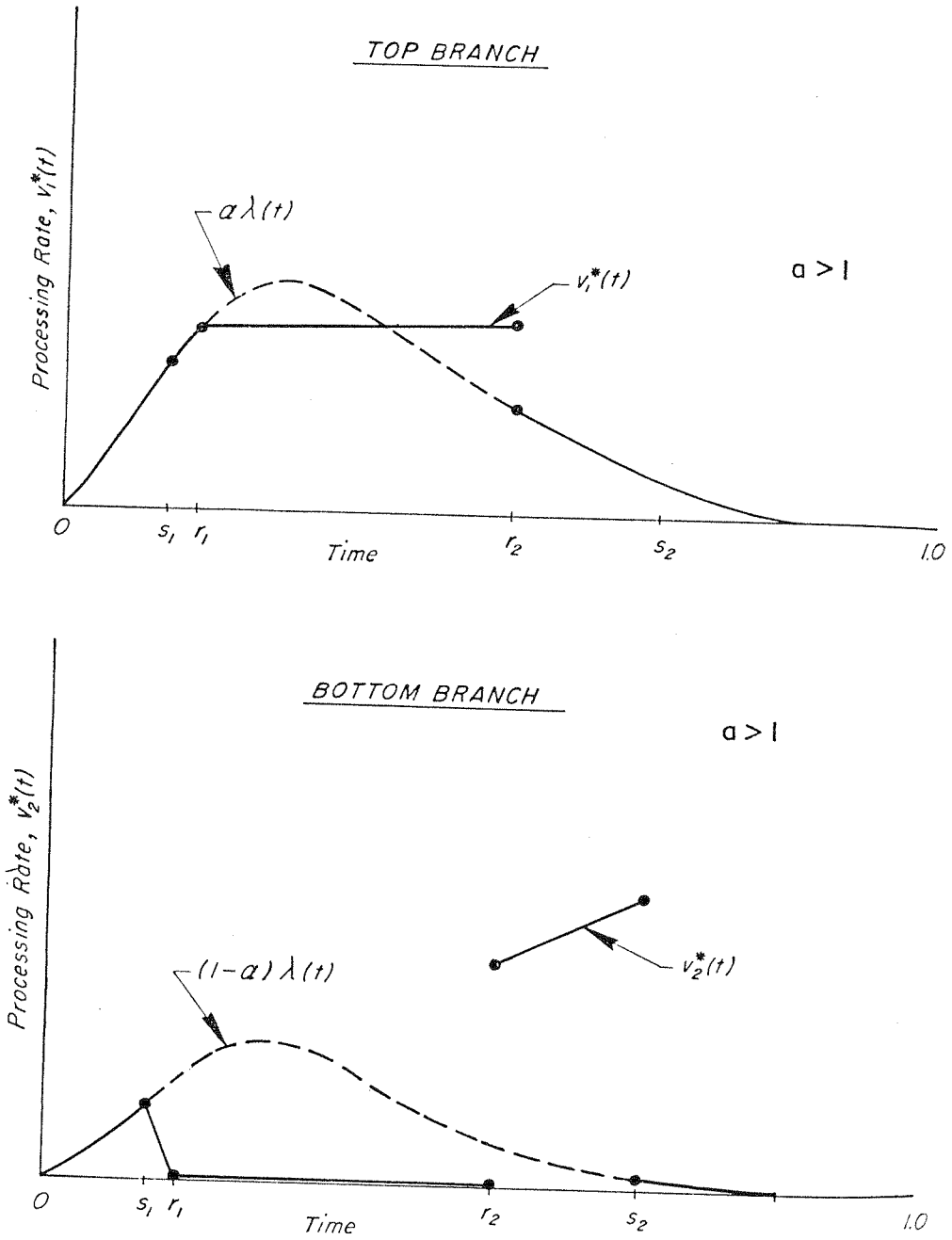


Fig. 10. Optimal processing rates.

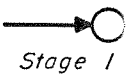
ondary or Tertiary as well as a Primary sorting area, the need for additional storage areas becomes increasingly important.

As we mentioned earlier, benefits in cost are almost always offset by increased delays. Although it is difficult if not impossible to equate costs

and delays it may be quit
policies subject to a fixed c

In many industrial and
the storage facility provid
infrequently, can meet dem
not being re-stocked. A
resembles a saw-tooth; the
flow into the storage facili
flow from the system as den

In contrast to this proc
tinually being fed by the
the mailing public or the
contents of storage that are
mail carriers or other facili
In a sense, the shape of the s



Proces

Fig. 11. A s

increase of inventories is f
tories.

One of the functions of
dispatch times of trains, bu
determines optimal re-orde
storage more closely resemb
than they do the storage o
major distinction from eithe
ment objectives: in the case
may be one of them. Na
rules may differ from those

The Storage Process

It is convenient to think
serially connected pair. N
postal operations but, fortun
that accurately describe the

In Fig. 11, a processing
into the system at a rate λ(
into the storage area of st

and delays it may be quite simple to find optimal storage and release policies subject to a fixed cost or operating budget.

In many industrial and military production and storage operations, the storage facility provides a reservoir of items that, though ordered infrequently, can meet demand in those intervals of time when items are not being re-stocked. A plot of inventory as a function of time often resembles a saw-tooth; the leading edge of the saw-tooth represents sudden flow into the storage facility and is generally followed by more gradual flow from the system as demands occur and inventory is released.

In contrast to this process, storage areas within a post office are continually being fed by the 'production' process, i.e., the input flows from the mailing public or the output flows of sorting stages. It is now the contents of storage that are released infrequently at 'dispatch' times when mail carriers or other facilities are made available to transport the mail. In a sense, the shape of the saw-tooth is reversed in time so that the gradual

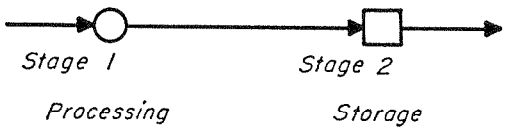


Fig. 11. A serial processing and storage stage.

increase of inventories is followed by a fairly sudden reduction of inventories.

One of the functions of a post office should be that of calculating the dispatch times of trains, busses, and other mail carriers just as a retailer determines optimal re-order policies. While the characteristics of mail storage more closely resemble the flow of water into hydroelectric facilities than they do the storage of items that are produced upon request, the major distinction from either of these processes lies in the choice of management objectives: in the case of mail the reduction of average letter delay may be one of them. Naturally, the optimal mail storage and release rules may differ from those that arise with minimum cost objectives.

The Storage Process

It is convenient to think of a processing stage and a storage stage as a serially connected pair. Not only is this flow pattern evident in many postal operations but, fortunately, mathematical models can be constructed that accurately describe the effect of storage on letter delays.

In Fig. 11, a processing operation takes place at stage 1. Mail flows into the system at a rate $\lambda(t)$, is processed at a rate $v(t)$, and then flows into the storage area of stage 2. Accumulated inventories are released

or 'dispatched' from storage at selected points in time. The question we consider is how to pick these dispatch times so that average letter delay is made small.

With very high processing rates, a letter enters the system and flows quickly through stage 1, proceeds to stage 2, where it then waits for a dispatch. Because of the high processing rates, the wait in queue at stage 1 is small in comparison to the wait for a dispatch from stage 2. Moreover, the amount of mail that leaves the system at dispatch time is just equal to the cumulative flow that has entered the system in the interval preceding this dispatch time.

If, on the other hand, processing rates are small, mail waits in the queue of stage 1 and in stage 2. A smaller fraction of processed mail makes a dispatch at time T because an amount less than cumulative flow into the system, $\Lambda(T)$, is processed by the dispatch time.

Individual Letter Delays

Let us now assume that two dispatches of stage 2 inventories are being scheduled; that is to say, the contents of stage 2 are released at two times in the interval (0, 1). The intermediate dispatch is scheduled at time $T<1$ and the final one at time $t=1$. The latter ensures eventual dispatch of all mail. The total delay of a letter is again made up of two parts, the delays in stages 1 and/or 2.

The total delay, $\tau(t)$, is the sum of $\tau_1(t)$ in stage 1 and the delay of a letter that enters the storage stage at time $t+\tau_1(t)$. Although the total delay is still given by

$$\tau(t)=\tau_1(t)+\tau_2[t+\tau_1(t)], \tag{40}$$

the shape of $\tau(t)$ generally differs from that of a letter delayed at a processing stage. Figure 12 is a plot of stage 1 and stage 2 letter delays as a function of time when an intermediate dispatch is located at T. The interval $S=(s_1, s_2)$ corresponds to the time when mail inventories appear in the queue of stage 1. It is immaterial whether the delay $\tau_1(t)$ is due to processing rate restrictions on a number of stages (in a large network) preceding stage 1 or whether it is due to the capacity restrictions on flow through a single stage; what is important, however, is that we be able to calculate the fraction of letters that do not make the intermediate dispatch at T as a result of the positive queue delays in stage 1.

Even when processing rates at stage 1 are less than input rates in an interval $S=(s_1,s_2)$, a letter which enters early in the interval (0, 1) may have little or no wait in queue at stage 1 but then waits in storage for the first dispatch at time T. Hence, $\tau(t)=\tau_2(t)$ and is shown in Fig. 12 as the difference between the saw-tooth and the time axis. As one examines delays of those letters that arrive later in time, one arrives at the region

S where stage 1 delays
storage.

If arrival time plus
the letter leaves the sy
is greater than T the l
the end of the interval.
group is $T-t$ and in the

At some point in tim
stage 1 by the dispatch

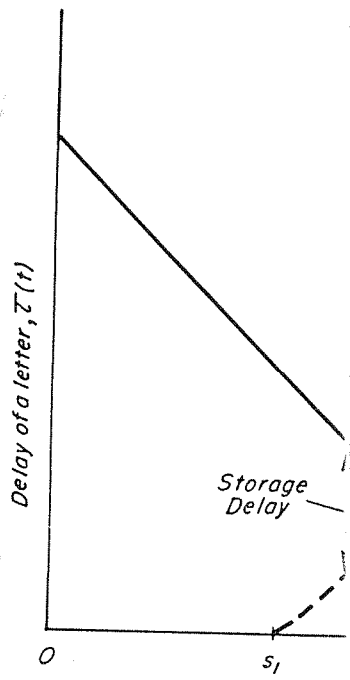


Fig. 12. Letter

wait in queue at stage 1;
intermediate dispatch at
equation (40), which ad
stage 1 and equates this s

With constant capacity
explicit solution for $\tau_1(t)$ i

$$\tau_1(t)$$

where k is the constant
substitute equation (42)

† We also point out to th
at each of N stages if all lette

S where stage 1 delays occur and a letter must wait in queue as well as in storage.

If arrival time plus stage 1 delay is less than the dispatch time, T, the letter leaves the system at time T; if arrival time plus stage 1 delay is greater than T the letter leaves the system on the second dispatch at the end of the interval. Hence the total delay for a letter in the former group is $T-t$ and in the latter, $1-t$.

At some point in time, t_1, a letter arrives, and is just processed through stage 1 by the dispatch time, T. The entire delay of this letter is due to a

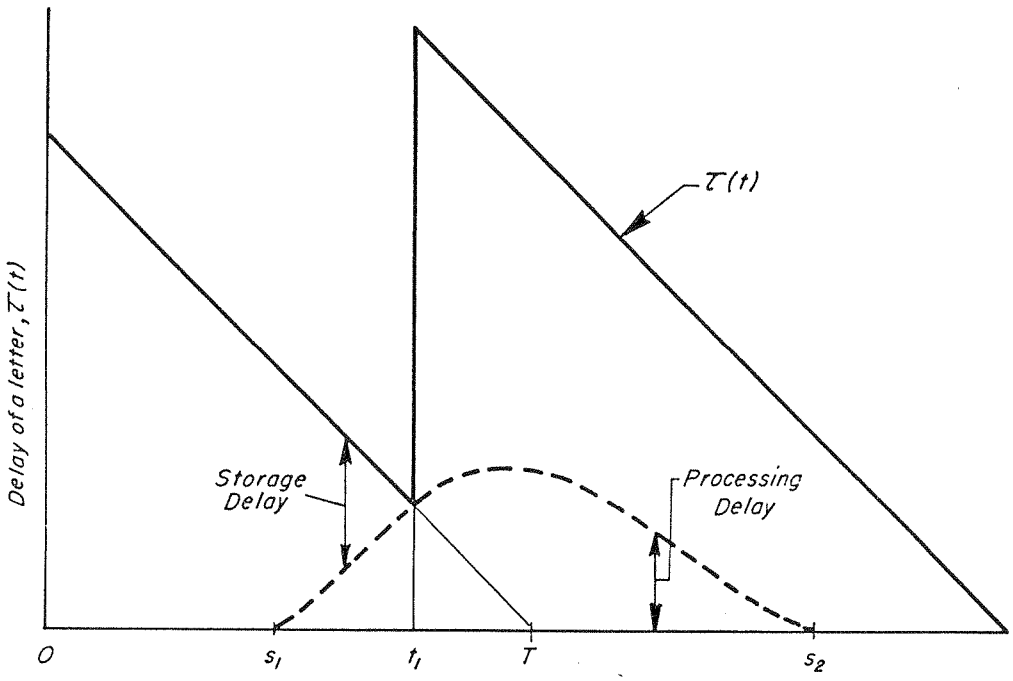


Fig. 12. Letter delay with an intermediate dispatch at T.

wait in queue at stage 1; any letter that arrives later does not make the intermediate dispatch at time T. Hence, t_1 is simply the solution of equation (40), which adds the arrival time of the letter to its delay at stage 1 and equates this sum to the intermediate dispatch time,

$$t_1+\tau_1(t_1)=T. \tag{41}$$

With constant capacity restrictions on processing rates at stage 1 the explicit solution for $\tau_1(t)$ is

$$\tau_1(t)=[\Lambda(t)-\Lambda(s_1)]/k-(t-s_1), \qquad (t\in S) \tag{42}$$

where k is the constant value for $v_1(t)$ in S.† It is a simple matter to substitute equation (42) into the expression for $\tau_1(t_1)$ in equation (41)

† We also point out to the reader that k might be the constant processing rate at each of N stages if all letter delays are concentrated at the first stage.

to find that t_1 can also be expressed in terms of the corner-point s_1, and the dispatch time T,

$$\Lambda(t_1)=k(T-s_1)+\Lambda(s_1). \tag{43}$$

Equation (43) states that flows that enter in a capacity constrained interval of time, t_1-s_1, must be processed in the longer interval $T-s_1$.

It is not possible for t_1 to lie inside S when the dispatch T is outside S. Geometrically, (Fig. 12), this statement is correct only if the slope of $\tau_1(t)$ is greater than the slope of the trailing edge of the saw-tooth. For capacity constrained sorting or processing rates at stage 1, the slope of $\tau_1(t)$ is

$$d\tau_1/dt=k^{-1}\lambda(t)-1>-1. \qquad (t\in S) \tag{44}$$

Hence, $\tau_1(t)$ always lies under $T-t$ if the dispatch time is later than the corner point s_2.

Average Delays

To study the effect of a storage stage on the flow of mail through serial processing stages, one must formulate average letter delay in terms of the dispatch time as well as the processing rates. The average delay of letter mail is found, as before, by multiplying the total delay of each letter by the rate of arrival of letters and integrating this expression over the interval (0, 1). All letters arriving before the time t_1 have a delay $T-t$ and all letters arriving after t_1 have a delay of $1-t$. The expression for average delay is therefore given by,

$$D=\int_0^1 \lambda(t)\tau(t)\,dt=1-\bar{t}-(1-T)\Lambda(t_1) \quad (T\in S) \tag{45a}$$

$$=1-\bar{t}-(1-T)\Lambda(T), \quad (T\in\bar{S}) \tag{45b}$$

where $\bar{t}=\int_0^1 t\lambda(t)\,dt$. The first terms in (45) depend on the shape of the input curve and the last terms are functions of the intermediate dispatch time T.

If we consider the case where T lies in S, the peak of the second saw-tooth in Fig. 12 is higher and located at an earlier time than it would be if there were no stage 1 delays. The fraction $\Lambda(t_1)$ rather than $\Lambda(T)$ makes the dispatch at time T. If there happen to be many (rather than one) succeeding dispatches in the interval (0, 1) the added delay to the mail fraction $\Lambda(T)-\Lambda(t_1)$ may not be serious; on the other hand, if there is only one intermediate dispatch the average delay of letter mail may increase sharply as a result of capacity restricted processing rates at stage 1. It is also interesting to note that as the maximum processing rate, k, increases, s_1 and s_2 get closer together until, finally, they coalesce. For larger values of k the average delay is always given by equation (45b).

We can also express
of the corner point s_1 by
The processing rate is co
the average delay is a qu

$$D=kT^2+[\Lambda(s_1)-$$
$$=1-\bar{t}-(1-T$$

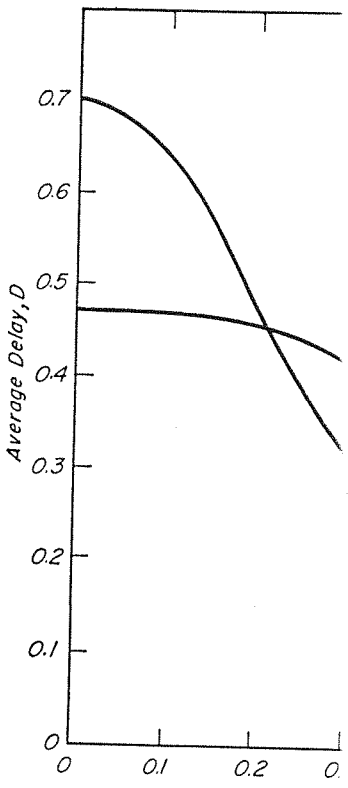


Fig. 13. Aver

The two expressions for a
value at the two corner poi
delay of letter mail throug
essing rates are not capac
are used for $\Lambda(t)$.

We have already show
S if they differ at all. He
small processing rates at st
S. This statement makes
delays to be an explicit fu
what fraction get delayed.

We can also express the average delay in the interval (s_1,s_2) in terms of the corner point s_1 by substituting equation (43) into equation (45a). The processing rate is constant in the capacity constrained region S, and the average delay is a quadratic function of the dispatch time, T,

$$D=kT^2+[\Lambda(s_1)-ks_1-k]T+[1-\bar{t}+ks_1-\Lambda(s_1)] \quad (T\in S) \quad (46a)$$

$$=1-\bar{t}-(1-T)\Lambda(T). \quad (T\in\bar{S}) \quad (46b)$$

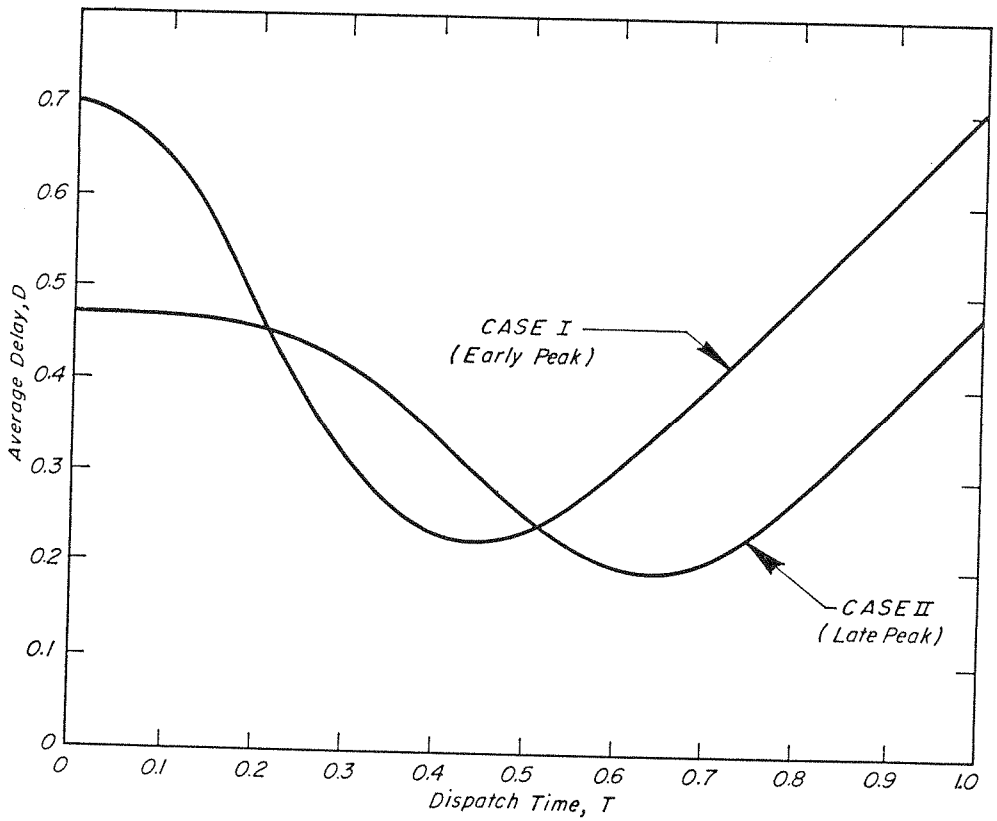


Fig. 13. Average letter delay versus dispatch time.

The two expressions for average delay are continuous in slope as well as value at the two corner points s_1 and s_2. Figure 13 is a plot of the average delay of letter mail through the two-stage system of Fig. 11 when processing rates are not capacity restricted and the cumulative flows of Fig. 3 are used for $\Lambda(t)$.

We have already shown that t_1 and T must simultaneously lie within S if they differ at all. Hence, average letter delay, D, is not affected by small processing rates at stage 1 so long as the dispatch time T is outside S. This statement makes physical sense since we do not expect average delays to be an explicit function of *where* letters are delayed but rather, *what fraction* get delayed.

Optimal Timing of the Dispatch

As one can see from Fig. 13 there is an intermediate dispatch that minimizes average delay. In the event that processing rates at stage 1 are not capacity constrained, equation (46b) is the expression for average letter delay in the interval (0, 1). Excluding the term $1-\bar{t}$ which is constant for a given input curve, it is easy to show that D has a single minimum. Both terms $(1-T)\Lambda(T)$ and $(1-T)\Lambda(t_1)$ in equation (45) start at zero, increase to a peak, and then decrease to zero as T ranges from zero to one. When inventories do not build up in front of stage 1, the optimal dispatch time, T^*, is found by setting the first derivative of equation (46b) with respect to T equal to zero. T^* is then the solution of

$$\lambda(T)(1-T)=\Lambda(T). \tag{47}$$

It seems reasonable that processing rate restrictions at stage 1 will affect the timing of the dispatch that minimizes average delay. We find that T^* either lies inside the interval S or is the solution of (47). On the other hand, the solution of (47) may still be the optimal dispatch time even though capacity processing rates are operative at stage 1.

If the minimum value of the average delay occurs in the interval S, T^* is the timing of the dispatch that minimizes equation (46a). In this case, T^* can be explicitly written in terms of the corner-point s_1,

$$T^*=\tfrac{1}{2}\,(1+s_1)-\Lambda(s_1)/2k. \tag{48}$$

We have found two possible solutions for the optimal dispatch time depending on whether or not the capacity processing rate, k, reduces the amount of mail processed by dispatch time. The obvious question that must be answered is, "when is the solution of equation (48) to be used in place of (47)?"

Since the average delay (as a function of T) has only one minimum in the interval (0, 1), the subinterval S can either lie to the left, to the right, or on either side of the optimal dispatch time. Since equations (46a) and (46b) are equal and tangent at both corner points and since (46a) is always less than or equal to (46b), we see that the solution of equation (48) can only be the dispatch that minimizes average letter delay if the solution of (47) also lies in S. If the solution of (47) lies outside S then this solution is the optimal dispatch time.

For the solution of (48) to be optimal, $s_1<T^*<s_2$ and the capacity processing rate, $k<\mathrm{Max}_t\,\lambda(t)$, must also satisfy the two inequalities

$$\Lambda(s_1)/(1-s_1)<k<\Lambda(s_1)/(1-2s_2+s_1). \tag{49}$$

intermediate dispatch that
processing rates at stage 1
s the expression for average
ng the term $1-\bar{t}$ which is
) show that D has a single
$-T)\Lambda(t_1)$ in equation (45)
crease to zero as T ranges
uild up in front of stage 1,
ting the first derivative of
T^* is then the solution of

(47)

tions at stage 1 will affect
ge delay. We find that T^*
on of (47). On the other
timal dispatch time even
at stage 1.
ecurs in the interval S, T^*
tion (46a). In this case,
er-point s_1,

k. (48)

ptimal dispatch time de-
sing rate, k, reduces the
he obvious question that
quation (48) to be used

has only one minimum
ier lie to the left, to the
. Since equations (46a)
oints and since (46a) is
solution of equation (48)
letter delay if the solu-
lies outside S then this

$T^*<s_2$ and the capacity
y the two inequalities

$+s_1)$. (49)

Both of these inequalities can be obtained from equation (48) or from the equivalent statement that the slope of D in equation (46) be negative at s_1 and positive at s_2.†

The optimal solution of T^* in S can also be obtained by substituting $\Lambda(s_1)+k(T^*-s_1)$, the total output flow of stage 1, for $\Lambda(T)$ and k, the constant processing rate for $\lambda(T)$ in equation (47).

Since the processing rates and dispatch times are independent decision variables, one can replace a large network of serial processing stages by a one-stage network. For an optimal schedule of the kind discussed in the early parts of the preceding section we found that all letter delays were concentrated at the first stage; hence if we replace k in equations (40)–(49) by the constant $k(\sum_i a_i)^{-1}$ of equation (30), we are able to find the optimal timing of the intermediate dispatch for an N-stage network.

When T^* lies in S, the minimum average delay is found by substituting

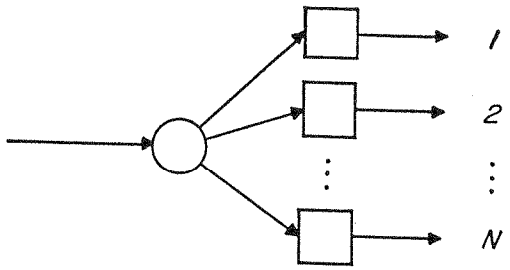


Fig. 14. Sorting into N storage stages.

equation (48) into (46a),

$$D^*=1-\bar{t}-\tfrac{1}{4}k+[ks_1-\Lambda(s_1)][\tfrac{1}{2}-\tfrac{1}{4}s_1+\Lambda(s_1)/4k]. \qquad (50)$$

Comparison of this expression for D^* with D in equation (46b) gives a quantitative estimate of the merits of increasing processing rates with present nonoptimal dispatch times versus using optimal dispatch times with present capacity restrictions on processing rates.

At a time when it was popular to believe that high-speed automatic processing equipment would cure the storage and delay problems brought about by infrequent dispatches, discussions were often focused upon these expressions for average delay.

Many Branches and Many Dispatches

The mathematical models of this section serve three purposes: first, they are natural extensions of those considered in the optimal timing of a

† Alternatively, if $k<\Lambda(s_1)/(1-s_1)$, T^* is the solution of (47) and lies to the left of s_1; if $k>\Lambda(s_1)/(1-2s_2+s_1)$, T^* lies to the right of s_2.

single dispatch. Secondly, they serve as an introduction to the final part of this section. Thirdly, and perhaps most important of all, they answer some of the dispatch problems that will arise when semi-automatic and fully automatic sorting machines are introduced into Secondary sorting areas.

Sorting machines under consideration have hundreds and thousands (rather than tens) of branches into which mail can be sorted and stored. If the total number of distinguishable addresses in a major post office remains in the thousands only a small and very special portion of letter mail may require Secondary sorting. Storage areas will be located at the end of each Primary branch, as in Fig. 14. Mail sorted at stage 1 will collect in each of the N storage stages and await dispatch.

The optimal timing of *one* intermediate dispatch from each of the N storage stages in Fig. 14 is no more difficult to find than the one-branch case considered in Fig. 11. Since the dispatch times of each branch are independent of one another, the average delay of letters is given by

$$D=1-\bar{\imath}-\sum_{i=1}^{i=N}\alpha_i(1-T_i)\Lambda(t_i), \qquad (51)$$

where we again interpret t_i as the time of arrival of the last letter making the intermediate dispatch T_i in the ith branch. The optimal dispatch time T_i^* of the ith branch equals T^* in equations (47) or (48), since the optimal dispatch times are independent of the constant fraction sorted at stage 1. Hence, the arguments following equation (48) apply; all dispatches are scheduled at the same time and the minimum average delay is still given by equation (50).

In the many dispatch (per branch) case we need only consider one branch at a time since the optimal timing of the dispatches will again be independent of the constant fraction sorted into any one branch. Supressing the index which refers to a branch but including one which refers to the jth dispatch, we now obtain a simple generalization of equations (45) and (46)

$$D=1-\bar{\imath}-\sum_{j=1}^{j=m}(T_{j+1}-T_j)\Lambda(t_j), \qquad (52)$$

where T_j is the time of the jth dispatch, $T_0=0$, $T_{m+1}=1$, and t_j is the arrival time of the last letter that makes the jth dispatch. The timing of the optimal dispatches, when sorting rates are not capacity restricted at stage 1, is also a simple generalization of equation (47):

$$\Lambda(T_j)-\Lambda(T_{j-1})=\lambda(T_j)(T_{j+1}-T_j). \quad (1\leqq j\leqq m) \quad (53)$$

In equation (53) there are m equations in m unknown dispatch times. When sorting rates are capacity restricted a subset, M, of the m dispatches will lie inside S while the remaining $m-M$ will lie in $\bar{S}$. The optimal dispatch times in S are the solutions of equation (53) when we substitute

duction to the final parts
rtant of all, they answer
hen semi-automatic and
l into Secondary sorting

hundreds and thousands
an be sorted and stored.
s in a major post office
special portion of letter
eas will be located at the
il sorted at stage 1 will
dispa h.
tch rom each of the N
ind than the one-branch
imes of each branch are
f letters is given by

$t_i)$, (51)

of the last letter making
The optimal dispatch
(47) or (48), since the
istant fraction sorted at
on (48) apply; all dis-
minimum average delay

need only consider one
dispatches will again be
y one branch. Supress-
ling which refers to
ation of equations (45)

$t_j)$, (52)

$_{m+1}=1$, and t_j is the ar-
spatch. The timing of
t capacity restricted at
(47):

$_j)$. $(1\leqq j\leqq m)$ (53)

nknown dispatch times.
M, of the m dispatches
lie in $\bar{S}$. The optimal
53) when we substitute

$\Lambda(t_j)$ for $\Lambda(T_j)$ and k for $\lambda(T_j)$ on the right-hand side. There will still be m equations in m unknowns; while $m-M$ can be written in the form of (53), M are of the form,

$$\Lambda(t_j)-\Lambda(t_{j-1})=k(T_{j+1}-T_j). \qquad (T_j\in S) \quad (54)$$

The solutions of the optimal dispatch times inside as well as outside S are affected by the location and duration of the capacity-constrained interval; however there is one useful shortcut that reduces the dimensionality of the problem.

Label those M dispatches falling in S with the index $i=I+1, I+2, \cdots, I+M$. Substituting t_i for t_1 and T_i for T in equation (43) we equate input flows and amounts sorted by the ith dispatch,

$$\Lambda(t_i)=k(T_i-s_1)+\Lambda(s_1). \qquad (55)$$

Substituting (55) into (54) gives the difference equation,

$$T_{i+1}-2T_i-T_{i-1}=0. \qquad (56)$$

We can write the solution of (56) in terms of the two dispatches T_I and T_{I+M+1} closest to the boundary of S,

$$T_i^*=(M+1)^{-1} \cdot\{(i-I)T_{I+M+1}-(M+1-i+I)[k^{-1}\{\Lambda(s_1)-\Lambda(T_I)\}-s_1]. \qquad (57)$$

Substituting (57) for T_{I+1} and T_{I+M} in equation (53) we obtain $m-M$ equations in $m-M$ unknown dispatch times in $\bar{S}$. It should be clear that because of equations (55) and (57) the optimal dispatch times in $\bar{S}$ are functions of s_1. The arguments which show that there is a one-to-one correspondence between the solutions of (53) and (54) in S is similar to the argument for the one dispatch case.

If the capacity-constrained interval occupies a large part of the time scale there may be no dispatches, other than those at $t=0$ and $t=1$, located outside S. In this case, equation (57) becomes,

$$T_i^*=(M+1)^{-1}\{i-(M+1-i)[\{\Lambda(s_1)/k\}-s_1]\}. \qquad (58)$$

When $M=1$ this reduces to the single dispatch case of equation (48).

The flow of mail out of a storage stage is zero except at the jth dispatch time when there is a pulse equal to $\Lambda(t_j)-\Lambda(t_{j-1})$. In the limit of large m, i.e., many dispatches from any one branch, we expect the output flows from the storage stage to resemble closely the output from the sorting or processing stage. Each pulse of mail will be small but if the dispatches are evenly spread out they have the effect of replacing one or two high-volume mail pulses by continuous flow. That several (three or four) dispatches did provide a good approximation to continuous flow in our

experiments was fortunate, since the simplicity of the graphical solutions of equations (47) and (53) does not extend to more than three dispatches.

For a large number of evenly spaced dispatches when sorting rates at stage 1 are not capacity restricted, the average delay of letter mail in equation (52) has the limiting expression

$$\text{Lim } D_{m\to\infty} = \text{Lim } D^*_{m\to\infty} = \text{Lim}_{m\to\infty}\{1-\bar{t}-\textstyle\sum_{j=1}^{j=m}(1/m)\Lambda(j/m)\} = 0.$$

The Primary and the Secondary

If a post office contained only one major sorting area the dispatch problems we have studied might go a long way towards understanding and improving letter delays. But there is at least one additional complication: while dispatches from Primary storage stages may be subject to control, dispatches from the Secondary storage stages may be fixed and not under the explicit control of the post office. Historically, the fixed train and plane dispatch schedules, and hence those of the Secondary, are more closely tied to passenger and freight than to mail service.

Although dispatch times in the Secondary may be fixed, decision rules must be sought for the release of mail inventories from the Primary storage areas. Before we do this we must attempt to answer a question that is often asked: Why not feed each Primary branch directly into its Secondary sorting stage? Since queues of mail can build up in front of each sorting stage, why should a normally smooth, direct, and fast flow process be artificially interrupted by storage?

Throughout this paper we have made reference to a number of *distinct* serial and parallel sorting, processing, and storage operations. In no case have we had any need to analyze the effects of parallel stages that only serve to increase the flow capacity. For our purposes it has been sufficient to replace a large group of identical, parallel, processing operations by a single stage and assign to that stage a number or function that reflects the capacity or actual flow rate assigned to the group.

In practice, of course, one increases processing or sorting rates by adding more parallel stages, i.e., a duplication of facilities. Consider the physical layout of a manual sorting stage in a large post office. Because of the limitation on a man's reach, it was seldom the case that more than 50 sorts were made per stage. To increase total flow rates through the Primary as many as 100 identical parallel sorting stages might be used. If there were to be a direct connection between *each* Primary sorting stage and each Secondary sorting stage, upwards of 5000 direct links would be needed. If there were a similar structure in the Secondary this number might again be multiplied by a large factor. Until recently, the sheer cost and space considerations of such a network have been prohibitive, and post offices have had to resort to storage facilities that could collect the mail until economical and timely transportation was provided.

Quite naturally, transportation facilities that carry mail between sorting areas within a post office differ quite radically from those that haul mail between post offices. Several distinct methods have been used. The oldest and perhaps the one still in greatest use is a manual one where at scheduled or random intervals the pigeonholes (storage stage) in each primary Case (Primary storage area) are 'swept clean' (inventory is released). More modern techniques, which have been introduced since 1957 and which will be described in greater detail in the next section are those that depend on a system of belts and conveyors to carry the mail from sorting to storage areas. There are more refined semi-automatic conveying and sorting devices in existence or in development; almost invariably, however, there is a question of optimal release rules if mail sorted in the Primary must also pass through a second or third major sorting area.

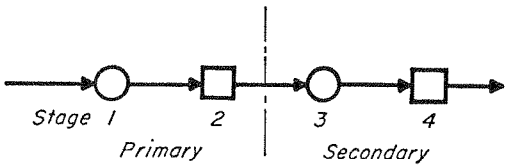


Fig. 15. A branch through the Primary and Secondary.

The problems that arise in the Secondary sorting areas are due to the large pulses of mail that flow into the Secondary as a result of the simultaneous release of the contents of many parallel Primary storage stages. If the sorting rate in a Secondary stage were infinite, the pulse of mail could be processed instantly and prepared for immediate dispatch to trains, planes, or busses. We use the word 'prepare' because mail in a Secondary branch must in fact be tied, bundled, labelled, and packaged in some form that preserves the distinct categories into which the mail has already been sorted.

The sorting and processing rates that are available in a Secondary are, however, not infinite. The obvious result is that one must calculate the amount that can be processed in a known interval of time and release the contents of storage (assuming very fast transportation between stages) at the Primary that much earlier than the final dispatch time in the Secondary.

In Fig. 15, stages 1 and 2 are in the Primary, stages 3 and 4 are in the Secondary. The input rate to stage 1 is $\lambda(t)$. The output rate of stage 1 looks like $\lambda(t)$ or is constant depending on whether or not sorting rates at stage 1 are capacity restricted. Mail inventories in stage 2 increase until the contents of storage are released at a dispatch time, T. This pulse of

mail is then processed in stage 3. Again, the output flows from stage 3 are stored in stage 4 until the inventory is released.

Fix a single dispatch time of stage 4 at $t=X$. If the capacity sorting rate at stage 3 is c, the largest inventory that can be released from stage 2 and completely processed in stage 3 by the stage 4 dispatch equals $\Lambda(T)$. T is the solution of

$$\Lambda(T)=c(X-T). \tag{59}$$

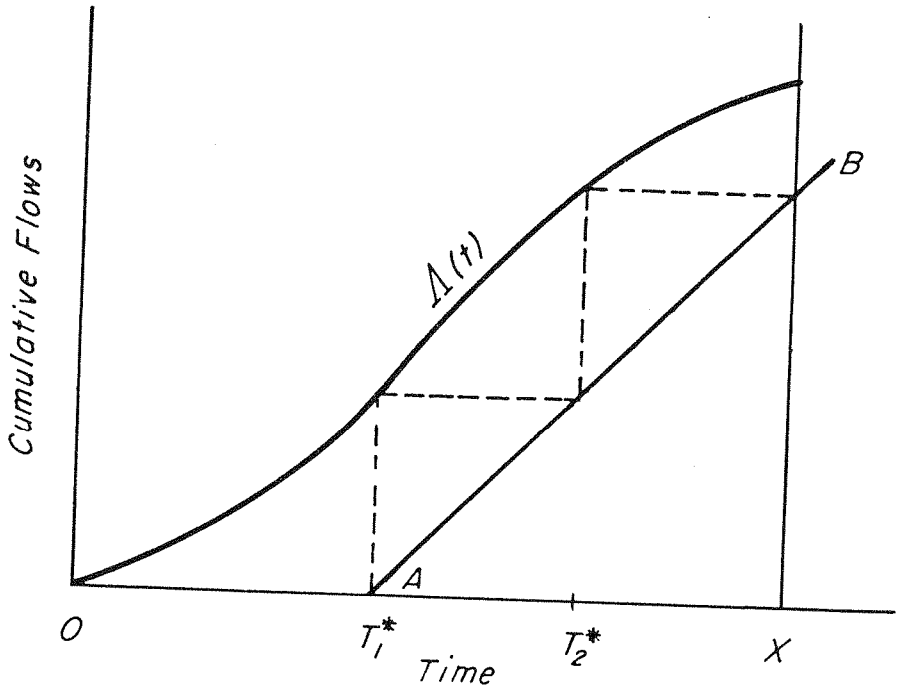


Fig. 16. A graphical solution of the two-dispatch case.

If stage 1 is capacity-restricted, its output function is substituted for $\Lambda(T)$. The argument follows that if T is earlier than the solution of equation (59) the amount processed by stage 1 and released by stage 2 is smaller than $\Lambda(T)$. Likewise if T is later than this solution, capacity processing rates in stage 3 cannot completely process the inventory released from stage 2.

In general, for m dispatches from stage 2 the dispatch times are given by

$$\Lambda(T_j)=c(T_{j+1}-T_1). \qquad (1\leqq j\leqq m) \tag{60}$$

Again, we interpret $T_0=0$, $T_{m+1}=X$.

The graphical solution of equation (60), for general $\lambda(t)$ and two dispatches, is shown in Fig. 16. T_1 and T_2 are the two dispatch times at

stage 2. X is the fixed c
line AB is the maximum
ventory sorted in the i
ventory sorted in the i
difference between the c
amount of mail that is
stage 4. In general, the
Primary dispatches is sim
until one line just touches
dashed lines in Fig. 16.

We have shown in (5
mail that makes an early

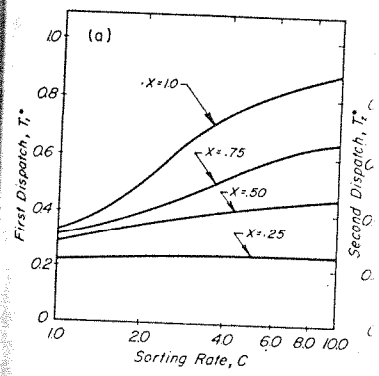


Fig. 17. Optim

would seem to be identical
letter delay. The proof
timing of the second Secon
mail not making the first dis
one; the average delay of
by the product of the fract
dispatch. There is no way
dispatch at X can be matc

Figure 17 is a plot of th
for a Secondary dispatch at
dispatch times are plotted
and are based on the cumul
that the optimal dispatch t
large values of c. In these c
carries most of the mail whil

Analysis of Fig. 17 also s
mary by X can be dispatch

stage 2. X is the fixed dispatch time of the final stage. The slope of the line AB is the maximum sorting rate, c, at stage 3. The amount of inventory sorted in the interval (T_1,T_2) is $\Lambda(T_1)$. The amount of inventory sorted in the interval (T_2,X) is $\Lambda(T_2)-\Lambda(T_1)$. Hence, the difference between the cumulative input curve, $\Lambda(X)$, and $\Lambda(T_2)$ is the amount of mail that is *not* processed by the fixed dispatch time, X, at stage 4. In general, the problem of locating the optimal timing of the Primary dispatches is simply one of drawing parallel lines (with slope c) until one line just touches the intersections of the horizontal and vertical dashed lines in Fig. 16.

We have shown in (59) that these solutions maximize the fraction of mail that makes an early (rather than a late) Secondary dispatch. This

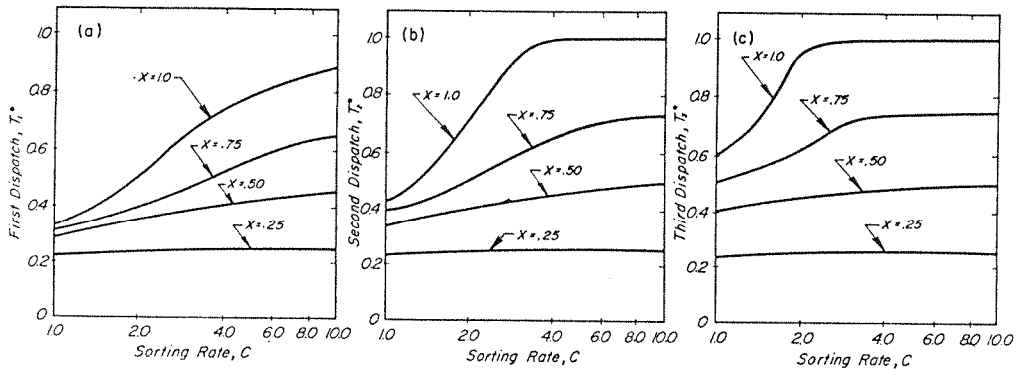


Fig. 17. Optimal timing of three primary dispatches.

would seem to be identical to the optimal policy that minimizes average letter delay. The proof of this statement is simple. Whatever the timing of the second Secondary dispatch the important point is that any mail not making the first dispatch will at least have to wait until the second one; the average delay of letters in that branch is essentially increased by the product of the fraction that has to wait and the time to the next dispatch. There is no way by which the failure of some mail to make the dispatch at X can be matched by delay reductions of other mail.

Figure 17 is a plot of the optimal timing of three Primary dispatches for a Secondary dispatch at $X = 0.25$, 0.50, 0.75 and 1.00. The optimal dispatch times are plotted as a function of the capacity starting rate, c, and are based on the cumulative input flows of Case II in Fig. 3. We see that the optimal dispatch times are least sensitive to changes in c for large values of c. In these cases, the first Primary dispatch comes late and carries most of the mail while succeeding dispatches become less essential.

Analysis of Fig. 17 also shows that almost all mail processed in the Primary by X can be dispatched from the Secondary if $c \geqq 2.0$, i.e., $c(X-$

$T_1^*) = \Lambda(X)$. In our experiments we found c greater than 2.0 in almost all manual Secondary stages. It therefore becomes doubtful whether high-speed sorting equipment in the Secondary will reduce letter delays.

In completing this section on the mathematical models of sorting and storage operations in the Primary and Secondary, we want to record a number of problems that have been studied and solved. We have not included their analysis because of space limitations and because we feel that much of it will be obviated with the introduction of modern processing and sorting machines. The research included a study of the effects of: (*i*) capacity restrictions on the sum of processing and sorting rates in the Primary and Secondary; (*ii*) sum restrictions on the total number of Primary dispatches, and (*iii*) different dispatch times for groups of Secondary branches.

EXPERIMENTS

Introduction

Part of our task in this study was the actual experimental evaluation of scheduling and dispatching policies at a large United States Post Office. Partly because of its location and partly because of the recent introduction of a modern system of conveyors and belts between processing stages, the office chosen was the Roosevelt Park Annex Post Office in Detroit, Michigan. This office handled flows of mail to and from branch offices on the one hand, and inter-city flows on the other. Its main function was the sorting and rerouting of all classes of mail through the City of Detroit.

While the Roosevelt Park Annex sorted and routed all classes of mail, the experiments that are described in this section were restricted to sorting and processing operations of first-class letter mail. Since the peak of the mail flows occurred in the late weekday afternoons, the experiments were confined to the Monday through Friday 4 p.m. to midnight shift—the so-called Tour 3. The abstract properties of the sorting operations and the flow processes are duplicated to one degree or another in every post office; the major differences are in the number of branches emanating from a sorting stage, the mode of transportation of mail from stage to stage, the total volumes of mail processed by the post office, and the fractional flows through each branch of a sorting stage.

As we mentioned earlier, a modern system of inter-stage conveyors had been installed in the Annex prior to the introduction of new scheduling policies. This system replaced a manual one in which bundles of letters were hand-carried between sorting stages. A schematic diagram of this trademarked "Mail-Flo" system is shown in Figs. 18 and 19.

At this point it seems desirable to describe the actual sorting and storage process more fully. The description is based on operations at the Roosevelt Park Annex.

Some mail (precancel
livery) branched from the
bulky items were remove
bundles marked 'local' and
mail was put into trays;
directly to the Incoming
were aligned or faced, ca
Primary; airmail and spec

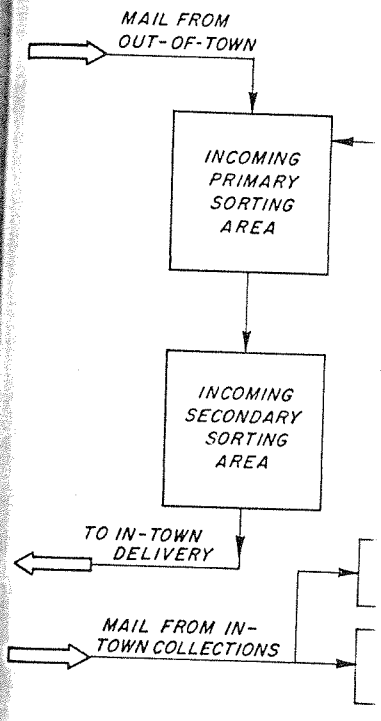


Fig. 18. M

sorted separately. The ma
containing about 580 letters
each containing 49 pigeonh
labeled with the names of t
Detroit, zoned; Detroit, un
(these were 'directs,' requi
Georgia; New England; For
ondary sorting.

The Secondary sorting sta
218). Pigeonholes in the 'C
such as Akron, Athens, Co
far exceeded the number
destinations (receiving ~ 5 pe

Some mail (precancelled local mail, bulky items, airmail, special delivery) branched from the main stream before the Primary. Of these, the bulky items were removed at the dumping stage where metered mail (in bundles marked 'local' and 'out-of-town') was also separated. The metered mail was put into trays; the local and out-of-town portions were sent directly to the Incoming and Outgoing Primaries. Uncancelled letters were aligned or faced, cancelled, and shipped directly to the Outgoing Primary; airmail and special delivery mail were removed at this stage and

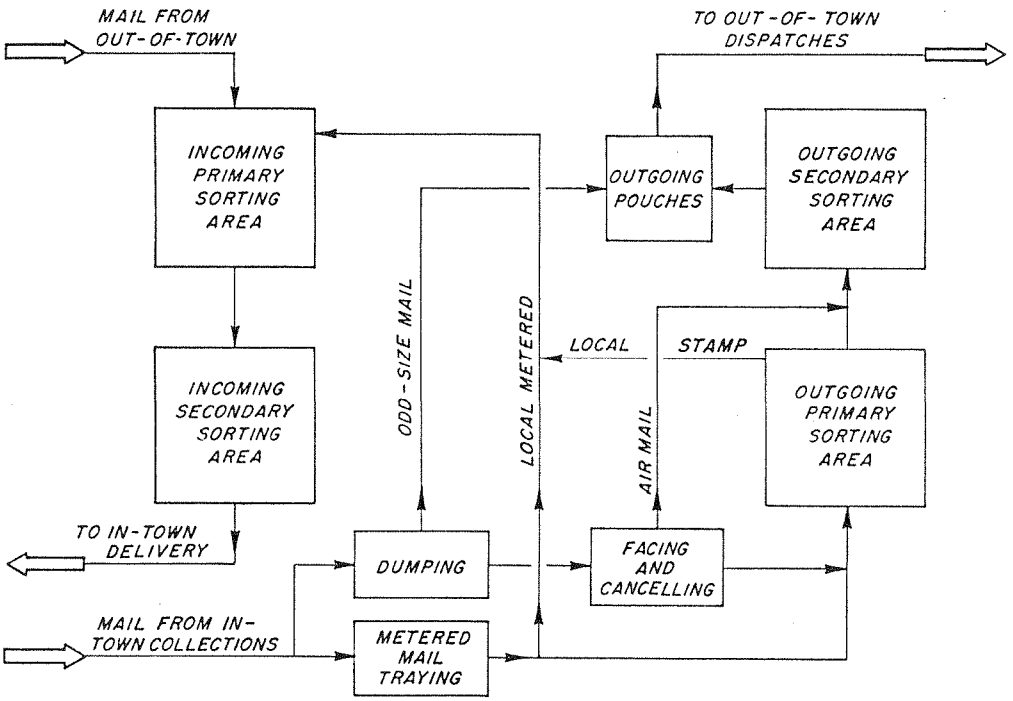


Fig. 18. Mail flows through a post office.

sorted separately. The main stream reached the Primary in trays, each containing about 580 letters. The Primary consisted of 218 sorting cases each containing 49 pigeonholes, i.e., storage stages. Most of these were labeled with the names of the Primary destinations or branches such as: Detroit, zoned; Detroit, unzoned; New York City; Lansing, Michigan (these were 'directs,' requiring no secondary sorting); Ohio; Florida-Georgia; New England; Foreign. Twenty-two destinations required Secondary sorting.

The Secondary sorting stages contained far fewer cases (3–14 rather than 218). Pigeonholes in the 'Ohio' secondary, for instance, had destinations such as Akron, Athens, Columbus. Since the number of destinations far exceeded the number of available pigeonholes, the less important destinations (receiving ~ 5 per cent of the mail) were classified as 'Residue.'

They were sorted in a Tertiary labelled according to their railroad branch lines.

Transport facilities for mail between the Primary and Secondary provided one of the major restrictions on mail flow. The Primary cases were arranged in aisles each containing 12–14 cases. Conveyor belts

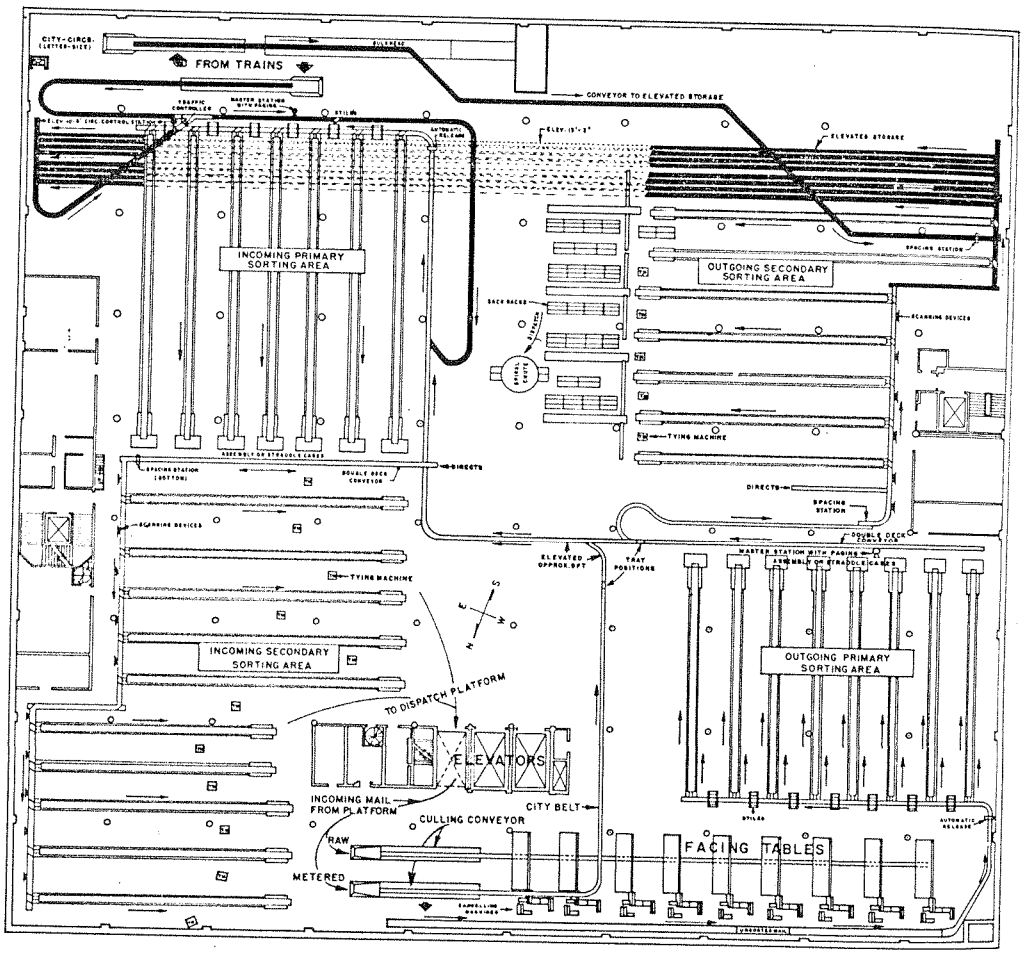


Fig. 19. A Mail-Flo system.

moved under the cases. A 'caller' gave a voice signal, e.g., 'Ohio,' and all sorters released the mail from the Ohio pigeonhole onto the belts. The mail was conveyed to the end of the aisle, packaged in trays, and then conveyed to the Ohio Secondary. The clearing time of the belt, i.e., the time required for mail introduced at one end to reach the other end, was one minute. Therefore, 'calls' were restricted to a maximum of 60 per hour and primary storage stages had to compete for places on the call schedule.

Before leaving the Po
pigeonholes in the Secon
tion, collected in pouches
Because these operations
of the mail volume, they a
times.

Incoming mail (includi
ment. The Primary sort
Because almost all the ma

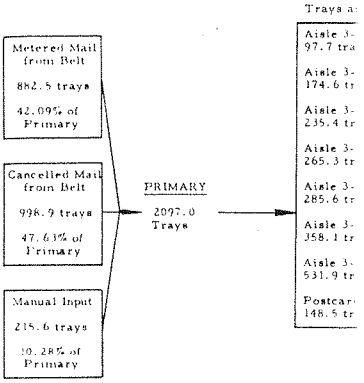


Fig. 20. Mail fl
on ave

delivered before morning,
incoming section. Hence
mail. It is possible that
reversed (e.g., an importa
our analysis could also b

A detailed breakdown
branches is shown in Fig.
from 1.0 to 1.5 million lett

As soon as the researc
Detroit (July 1958), it b
delays in the early proce
matical treatment that di
mail on the processing ra
and introduced during Se

Before leaving the Post Office, mail was collected ('swept') from the pigeonholes in the Secondary, tied into bundles with a common destination, collected in pouches and conveyed from the floor to a loading dock. Because these operations take little time, which is practically independent of the mail volume, they are of only minor interest in the discussion of delay times.

Incoming mail (including intra-city mail) had a similar sorting arrangement. The Primary sorted by zones, the Secondary by carrier routes. Because almost all the mail was received before midnight and could not be

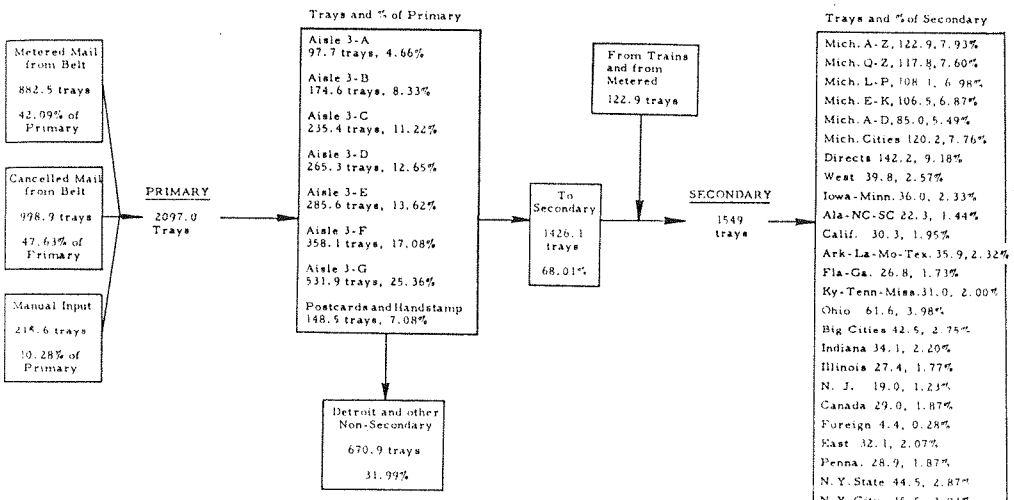


Fig. 20. Mail flows into Primary and Secondary (based on averages Sept. 30–Nov. 12, 1957).

delivered before morning, there were practically no avoidable delays in the incoming section. Hence, the experimental emphasis was on outgoing mail. It is possible that in another post office the situation might be reversed (e.g., an important train might arrive at 4 A.M.). In this case, our analysis could also be applied to the incoming section.

A detailed breakdown of the flow volumes observed through various branches is shown in Fig. 20. The total volume of mail in Tour 3 ranged from 1.0 to 1.5 million letters daily, excluding weekends.

As soon as the research project was initiated and a visit was made to Detroit (July 1958), it became evident that there were excessive letter delays in the early processing stages. After a relatively trivial mathematical treatment that did not take into account the effect of queues of mail on the processing rates at a stage, a manpower schedule was set up and introduced during September and October of 1958. This experiment

and certain aspects of the data-collection problem are described in the next section.

We then studied the scheduling and dispatch problems posed by the Secondary; these led to the mathematical models of the preceding section. It was then a simple matter to draw up a set of rules that foremen could use to reduce letter delays. These dispatch and scheduling rules are also reported; they were introduced in March and April of 1959.

In addition to the design and evaluation of over-all and local decision rules, this research study included an analysis of the predictability of total daily mail flows as well as development of a simple manual that described in detail the steps a foreman should follow in calculating and assigning processing and sorting rates. While we consider the latter a necessary and useful exercise, the details are not of interest here.

Measurement of Letter Delays

Before discussing experimental results it is necessary to consider how experimental data were obtained. The most direct way to measure delays would be to insert marked letters at selected points (e.g., the initial dumping stage) and retrieve them at later stages. This would give a direct measure of individual letter delays. But we could not use this method because of the large manpower requirements and the difficulties of retrieving test letters (many of the dummy letters were returned by the nominal addressees with notes expressing puzzlement or indignation!). Moreover, it was found that the sorting staff spotted the letters and gave them special priority. We therefore resorted to more indirect methods. Two methods seemed practical: (*i*) the delay of a letter could be computed by measuring the actual input rate and processing rate over a period of time; (*ii*) cumulative flows and inventories could be measured as a function of time, and average delay of letter mail through a particular stage could be obtained by calculating the area between these two curves. The second method was used because of the ease of actual measurements and because measurements of cumulative flows are at worst functions with changes in slope.

The success of this experimental method depends on the validity of several assumptions. The first is that the 'mix' or fraction of a branch to the main stream flow remains constant with time. The second is that so-called conversion factors do not vary. Inventories and cumulative mail volumes were not tabulated by individual letter count but rather by the unit of mail transported or processed at each stage. At the dumping stage, for instance, sacks of mail were weighed in bulk; cancelling machines recorded in terms of individual letter count; along conveyors, mail was counted as trays. We were furnished with the following conversion fac-

tors experimentally deter
580 letters = 1 tray.

The 'Before' Measurem

In anticipation of the r
historical information abo
Roosevelt Park Annex as e
we could obtain estimates
ing rates by reviewing an
States Post Offices. Afte
evident that this plan wo
kept by the Post Office did
in our analysis but report
power assignments then in
ment, namely the Mail-F
manpower assignments and
the Post Office. We felt t
structural changes that litt
to the new system. Hen
detailed survey of the syst
policies were introduced.

Initial observations an
made over a six-week peri
input and output flow ra
measured at 12-minute int
and Secondary sorting are
made at the dumping, met
ing stages. To give the
values obtained in these m
volumes of mail sorted in
ber-November period in 1
and Fig. 22 shows input an

After these initial exp
throughout the period Oct
flow pattern. The latter

1. Output volumes of m
2. Output volumes of fa
3. Inventories at the Prin
4. Sorting rates in the Pri
5. Output volumes of the

In addition, mail flows le
destinations (one for each

tors experimentally determined by Detroit personnel: 43 letters = 1 lb.; 580 letters = 1 tray.

The 'Before' Measurements

In anticipation of the research reported here, plans were made to obtain historical information about the processing and sorting operations at the Roosevelt Park Annex as early as September of 1957. The hope was that we could obtain estimates of letter delays, inventories of mail, and processing rates by reviewing and analyzing data officially recorded in United States Post Offices. After a brief survey of the historical data it was evident that this plan would not be successful for two reasons: (*i*) data kept by the Post Office did not contain specific delay measurements needed in our analysis but reported instead the number and cost of actual manpower assignments then in use; (*ii*) the recent introduction of new equipment, namely the Mail-Flo system, had resulted in natural changes in manpower assignments and local dispatch rules to transport the mail within the Post Office. We felt that the routing system had undergone so many structural changes that little if any of the earlier data would be applicable to the new system. Hence we decided to make an independent and detailed survey of the system before new scheduling, dispatch or routing policies were introduced.

Initial observations and data collection in the Mail-Flo system were made over a six-week period (Sept. 30 to Nov. 12, 1957). In this period, input and output flow rates, queues, and manpower assignments were measured at 12-minute intervals from 4 p.m. to midnight in all Primary and Secondary sorting areas. Measurements of mail volumes were also made at the dumping, metered mail traying, facing, cancelling, and pouching stages. To give the reader some idea of the scope and numerical values obtained in these measurements Fig. 20 lists the average daily flow volumes of mail sorted in the Primary and Secondary during the September-November period in 1957. Figure 21 shows cumulative output flows and Fig. 22 shows input and output flows of a typical Secondary.

After these initial experiments, measurements were made regularly throughout the period October 1958–May 1959 at selected points in the flow pattern. The latter set of data included measurements of:

1. Output volumes of metered mail traying operations (measured in trays).
2. Output volumes of facing-cancelling operations (measured in letters).
3. Inventories at the Primary Sorting Area (measured in trays).
4. Sorting rates in the Primary (measured in manpower).
5. Output volumes of the Primary Sorting Area (measured in trays and pounds).

In addition, mail flows leaving the Annex on each train to 22 selected destinations (one for each group of stages in a Secondary branch) were

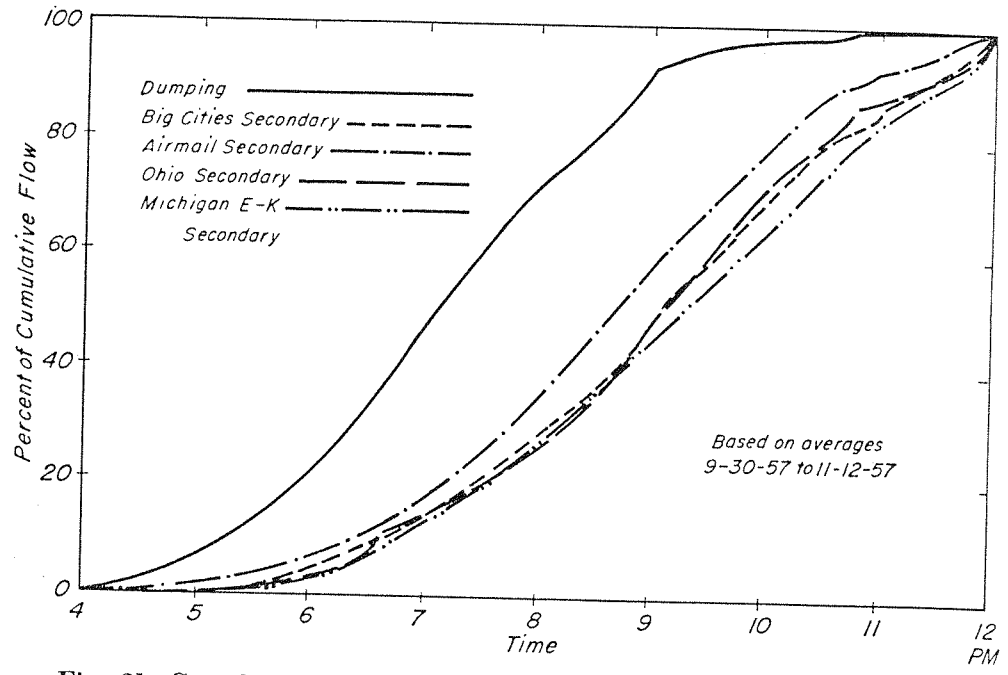


Fig. 21. Cumulative output of Dumping stage and selected Secondaries.

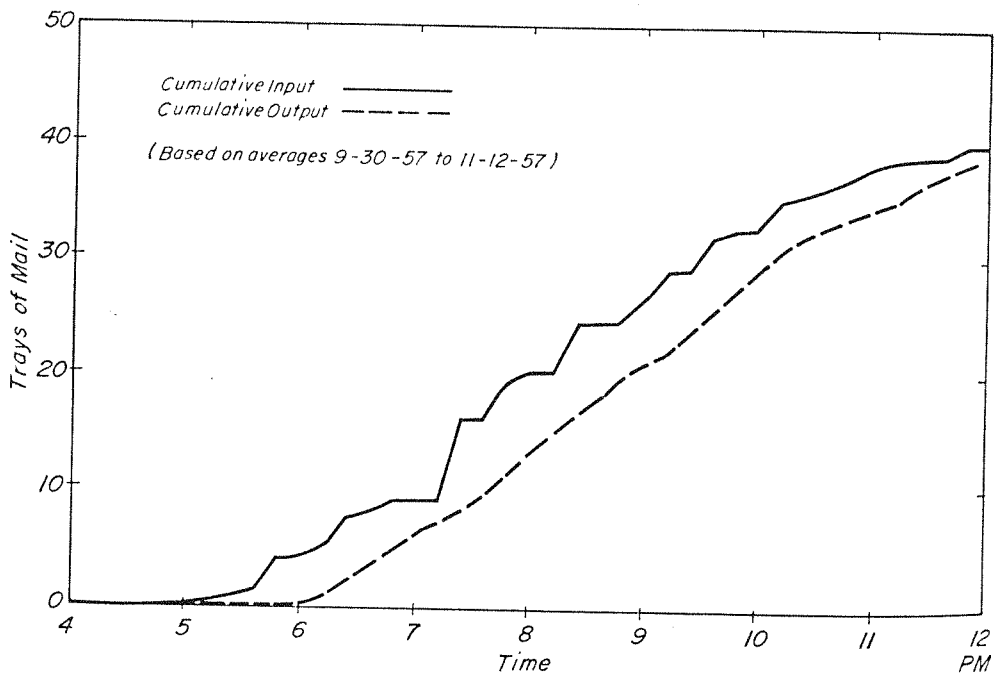


Fig. 22. Cumulative input and output flows, West Secondary.

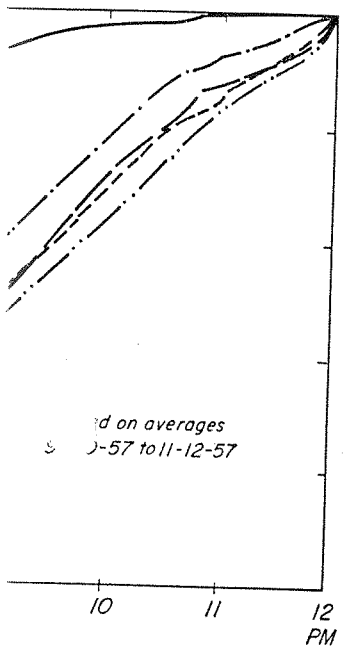
weighed over two two-w
experiments in the Second

Rescheduling Experime

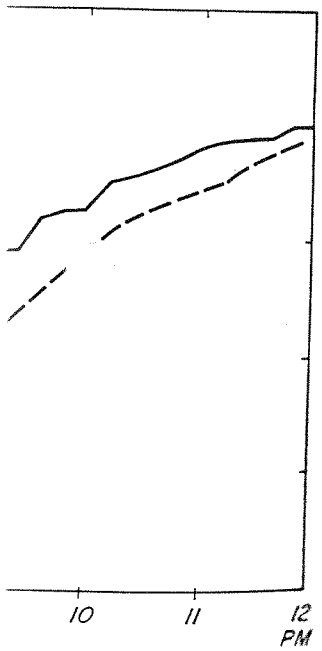
The data collected an
November 1957 revealed
delays at the Primary So

	Inventories at
4:00 P.M.	
4:15	
4:30	
4:45	
5:00	
5:15	
5:30	
5:45	
6:00	
6:15	
6:30	
6:45	
7:00	
7:15	
7:30	
7:45	
8:00	
8:15	
8:30	
8:45	
9:00	
9:15	
9:30	
9:45	
10:00	
10:15	

(a) Average of inventory
include mail waiting on the
increased by ~100 trays b
(b) Average of inventory



d selected Secondaries.



Vest Secondary.

weighed over two two-week periods, before and after the rescheduling experiments in the Secondary Sorting Area.

Rescheduling Experiments—Processing and Primary Sorting Stages

The data collected and analyzed during the months of October and November 1957 revealed that there were very large inventories and long delays at the Primary Sorting Area. These are shown in Table I. The

TABLE I

INVENTORIES AT PRIMARY BEFORE AND AFTER RESCHEDULING

	Before[a]	After[b]
4:00 P.M.	45	112
4:15	54	100
4:30	68	109
4:45	82	125
5:00	117	153
5:15	136	158
5:30	145	196
5:45	154	208
6:00	185	191
6:15	200	191
6:30	232	182
6:45	268	199
7:00	283	256
7:15	280	283
7:30	275	295
7:45	284	295
8:00	295	263
8:15	287	246
8:30	281	226
8:45	290	210
9:00	314	210
9:15	312	172
9:30	306	119
9:45	298	60
10:00	297	34
10:15	261	31

[a] Average of inventory (trays) Sept. 30–Oct. 3, 1957. These numbers do not include mail waiting on the belt between facing table and Primary and should be increased by ~100 trays between 7:00 and 9:30 P.M.

[b] Average of inventory (trays) Sept. 30–Oct. 3, 1958. Mail on belt included.

large delays and inventories of mail were attributed to overanxiety on the part of the foremen to avoid idle labor charges. Because of the constant checks on mail processing costs and the almost absolute lack of checks on mail processing delays, the schedules reflected manpower assignments that would ensure a large inventory of mail in front of processing stages.

The computation of optimal assignments in the early processing stages followed our mathematical models in the second section. Essentially, processing rate assignments maintain zero inventory at each stage until a point is reached where input flow rates exceed the maximum processing rates available.

An example of the actual manpower assignments in the early processing stages is given in Table II.† These assignments were for days of heavy mail flow (first and last of month). It will be seen that earlier assignment of capacity sorting rates in the Primary made it possible to reduce manpower assignments earlier, reflecting an earlier reduction of inventories.

The simple models of the second section were complicated by the existence of stochastic terms in the mail flows. The effect of small positive fluctuations in the input rate was to increase the inventory levels of unprocessed mail. The effect of small negative fluctuations was to increase the fraction of time that the assigned manpower was not processing mail, i.e., idle labor. On the first day of operations under the revised schedule, 137 hours of idle labor were recorded because of unpredictable negative fluctuations in the mail flow. It was later found that, by allowing a controlled queue of 0.75 trays per man to develop at the Primary, an acceptable level of 3 man-hours of idle labor (out of 700 man-hours per shift) could be maintained. Naturally, the build-up of a queue resulted in larger letter delays; in the Primary, the penalty was about 15 minutes for the average letter.

The results of the experiment are shown in Fig. 23 and Tables II and III. Figure 23 shows that the average letter delay up to the end of the Primary was reduced by about 0.8 hours. Table II shows that capacity sorting rates in the Primary were reduced earlier while Table III shows that input flows to the Secondary peaked earlier than before the experiment. Because the 'call' schedule had not yet been revised this measurement of Secondary input flows is evidence of a corresponding change in Primary output flows.

It is appropriate to mention one feature of the system that proved to be of considerable practical importance. While there was, at any time, a maximum amount of available manpower, no penalties were attached to less-than-capacity assignments. Because of the frequent arrival of trains

† No figures were available for manpower assignments at the facing tables prior to 1958.

MANPOWER ASSIG
(Sept.

Time	Facing tabl	
	(9/30/58)	(10/
4:00	7	
4:15	7	
4:30	24	
4:45	24	
5:00	24	
5:15	24	
5:30	40	
5:45	40	
6:00	50	
6:15	50	
6:30	70	
6:45	70	
7:00	70	
7:15	70	
7:30	70	
7:45	61	
8:00	58	
8:15	58	
8:30	78	
8:45	78	
9:00	22	
9:15		
9:30		
9:45		
10:00		
10:15		
10:30		
10:45		
11:00		
11:15		
11:30		
11:45		
12:00		

TABLE II

MANPOWER ASSIGNMENTS AT FACING TABLES AND PRIMARY
(Sept. 30 and Oct. 1, 1957 and 1958)

Time	Facing tables		Primary			
	(9/30/58)	(10/1/58)	(9/30/57)	(10/1/57)	(9/30/58)	(10/1/58)
4:00	7	16	16	15	25	30
4:15	7	16	17	13	25	30
4:30	24	24	36	25	50	50
4:45	24	24	51	31	50	50
5:00	24	24	49	37	72	70
5:15	24	24	53	51	75	70
5:30	40	40	63	69	103	110
5:45	40	40	62	58	107	147
6:00	50	50	82	83	157	178
6:15	50	50	84	91	189	184
6:30	70	70	93	83	184	175
6:45	70	70	112	161	189	189
7:00	70	70	139	175	169	183
7:15	70	70	155	194	189	188
7:30	70	70	163	176	186	188
7:45	61	70	185	158	186	188
8:00	58	70	183	178	181	188
8:15	58	70	163	178	189	188
8:30	78	70	163	176	191	189
8:45	78	70	166	175	191	189
9:00	22	40	162	179	167	158
9:15		10	161	179	119	136
9:30			158	192	160	147
9:45			84	107	160	93
10:00			101	102	40	40
10:15			145	154	14	
10:30			175	191	14	
10:45			133	126		
11:00			114	106		
11:15			143	120		
11:30			151	111		
11:45			43	45		
12:00			61	49		

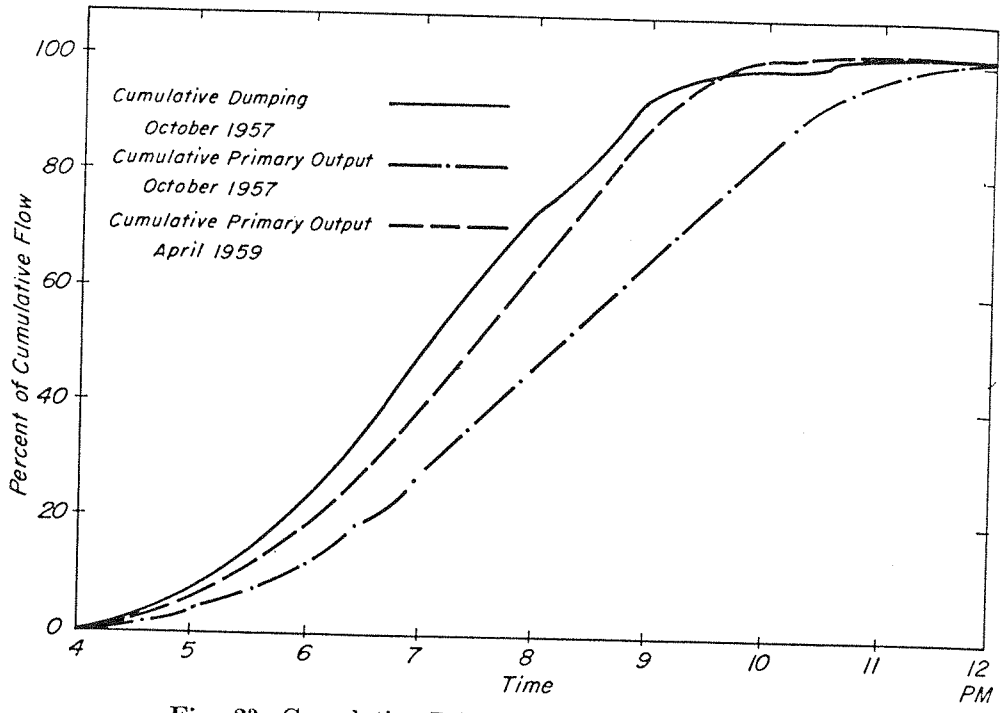


Fig. 23. Cumulative Primary output in 1957 and 1959, cumulative output of Dumping stage in 1957.

during the evening hours, there was always a large backlog of unprocessed mail in the Incoming Primary. Since this mail had to be processed before early deliveries of the following morning, workers who were not needed in the Outgoing Primary were shifted to the Incoming Primary. As a

TABLE III

MAIL FLOWS (POUNDS) INTO SECONDARY BEFORE AND AFTER RESCHEDULING OF PRIMARY

Time	Before (Sept. 17, 18, 19, 22, 23, 1958)	After (Sept. 24, 25, 26, 29, 30, 1958)
4–5 P.M.	2825	2440
5–6	6634	7875
6–7	11009	14476
7–8	16159	20442
8–9	18299	17110
9–10	14570	12042
10–11	4930	5254
11–12	495	696

result there was little idle la
Outgoing Primary.

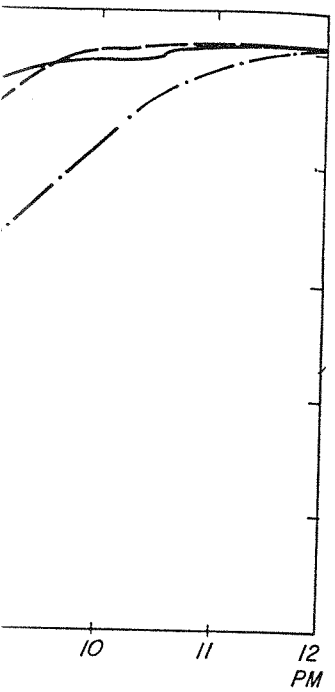
Rescheduling of the Sec

Rescheduling of the Sec
Primary storage rules were
matical models had been d
ficient time and experience
large scale. Initially, each
separately; as conflicting a
for mail inventories in the
handled by the mechanical
the processing rate assignm
sorting assignments and the
least the following steps:

1. Determination of outpu
2. Determination of indiv
stage.
3. Tabulation of fixed Se
tie, and transport mail bundle
4. Calculation of the opt
schedule. This was done grap
5. Assignment of the num
mized average letter delay or
each Secondary dispatch.
6. Compilation of an over
storage rules.

Trains often served se
for mail inventories in the
Whenever the 'calling rate
the Mail-Flo system, adjust
the Primary 'call' schedules
A simple numerical ex
Primary branch labelled
corresponding to this branc
and equipment for a maxim
breakdowns of that mail cat
on New York Central tra
fixed Secondary dispatch
times within the Post Office
A.M. On the same scale, w

† Airmail is processed in a



957 and 1959,
in 1957.

backlog of unprocessed
d to be processed before
s who were not needed
coming Primary. As a

RE AND AFTER

After
(Sept. 24, 25, 26, 29, 30, 1958)
2440
7875
14476
20442
17110
12042
5254
696

result there was little idle labor from less-than-capacity assignments in the Outgoing Primary.

Rescheduling of the Secondary

Rescheduling of the Secondary sorting assignments and revision of the Primary storage rules were accomplished during March 1959. The mathematical models had been developed by the beginning of the year and sufficient time and experience had been gained to use the new decisions on a large scale. Initially, each Secondary and each Primary was scheduled separately; as conflicting assignments arose and as the number of 'calls' for mail inventories in the Primary exceeded the number that could be handled by the mechanical conveying devices, many of the storage rules and the processing rate assignments were revised. Determination of the final sorting assignments and the 'calls' for Primary mail inventories involved at least the following steps:

1. Determination of output flows from the Primary sorting stages.
2. Determination of individual processing rates for each Secondary sorting stage.
3. Tabulation of fixed Secondary dispatches, lead times needed to package, tie, and transport mail bundles to a loading dock.
4. Calculation of the optimal 'call' times for a given Secondary dispatch schedule. This was done graphically (*see* Fig. 16).
5. Assignment of the number of 'calls' to each Primary branch, which minimized average letter delay or which guaranteed a large fraction of mail making each Secondary dispatch.
6. Compilation of an over-all schedule of manpower assignments and Primary storage rules.

Trains often served several Secondary sorting areas; hence, 'calls' for mail inventories in the Primary tended to bunch closely together. Whenever the 'calling rate' exceeded that which could be tolerated by the Mail-Flo system, adjustments of both the Secondary sorting rates and the Primary 'call' schedules were made.

A simple numerical example may be of interest. We consider the Primary branch labelled 'California.' In the Secondary sorting area corresponding to this branch destination there were four cases, i.e., space and equipment for a maximum of four trained sorters who made still further breakdowns of that mail category. Surface mail† to California left Detroit on New York Central trains 357, 369, and 39 and the corresponding fixed Secondary dispatch times (allowing for packaging and transport times within the Post Office building) were 4:20 P.M., 10:05 P.M., and 2:15 A.M. On the same scale, which is normalized to the eight-hour period of

† Airmail is processed in an Airmail Secondary.

Tour 3 beginning at 4:00 P.M. and ending at midnight, these dispatch times become $X_1=0.04$, $X_2=0.76$, and $X_3=1.28$. The total volume of California mail during that period was about 18,000 letters and sorting rate of skilled labor was found to be about 1700 letters per man-hour. The normalized capacity sorting rate, c, in the California Secondary is just the ratio of the total mail volume that can be processed to the actual mail volume; in this case, we calculate $c=3.00$. In other words, if the Secondary sorting stage operated at capacity in Tour 3 it could process about three times the normal volume of California mail.

Using the Case II curve of Fig. 3 as the output flows from the Primary sorting stages, we see that little mail can make the first dispatch at 4:20 P.M. For the second train at 10:05 P.M. a single call for Primary inventories should be made at $T^*=0.56$(8:35 P.M.).† In this one-call case, approximately 60 per cent of the evening's mail would make the fixed Secondary dispatch. If two calls are made for inventories in the Primary they should be located at $T_1^*=0.50$(8:00 P.M.) and $T_2^*=0.65$(9:12 P.M.); if three calls, the optimum timing is $T_1^*=0.47$(7:40 P.M.), $T_2^*=0.59$(8:40 P.M.) and $T_3^*=0.69$(9:28 P.M.). In the three-call case we get 86.7 per cent of the the evening's mail onto the Secondary dispatch. Had we been able to make an infinite number of calls, i.e., continuous output of the mail sorted in the Primary, 95 per cent of the evening's mail could have made that same Secondary dispatch. However, as we have already mentioned, the total calling-rate was severely restricted by the Mail-Flo system.

TABLE IV

SCHEDULE FOR CALIFORNIA SECONDARY

Calls	Manpower assignment (time period)	Trains
4:12 P.M.	4 men (4:17–4:20 P.M.)	NYC 357; 4:40 P.M.
7:39, 9:04	4 men (7:44–10:05 P.M.)	NYC 369; 10:25 P.M.
Midnight	1 man (12:05–1:11 A.M.)	NYC 39; 2:35 A.M.

In preparing the new scheduling instructions for postal personnel, the timing of the Primary 'calls,' manpower assignments, and times when Secondary sorters came on and off duty had to be calculated. This information was printed on a sheet and given to foremen in charge of each

† T^* is just the solution of $\Lambda(T)=3.00\ (0.76-T)$; in general we refer to equation (60) for T_j^*, the optimal timing of the jth dispatch.

Secondary sorting area. A t
California Secondary. In thi
based on the actual output fl
and February of 1959 (rather

A booklet, prepared befo
and supervisors of 'call' times
each 'call,' manpower sched
Secondary at certain critica
modifying the manpower an
amounts in each branch. I
manpower pool was schedul
tions in mail inventories.

It is gratifying that the
mail inventories arrived in t
predicted times, that invent
that mail inventories were
were to be rescheduled, and

The 'After' Measurement

With the introduction of
storage rules, mail volumes
could be measured and ave
For the purposes of measu
observed mail flows to each
destinations were individu
Secondary mail categories
ence between the latter an
example, an entire Primary

Post office personnel we
periods in February and Ap
fore and after the reschedu
ments it was possible to ob
train for each city and to
in an analysis of the 1957
were then obtained by in
output curves and the cur
stages. As an example, t
Fig. 24. The reduction in
between the 1957 and 195

The accuracy of these
tions of 'constant mix' a
but also on the assumptio

Secondary sorting area. A typical sheet is shown in Table IV for the California Secondary. In this case, the calculations of the 'call' times were based on the actual output flows of the Primary sorting stages in January and February of 1959 (rather than Case II of Fig. 3).

A booklet, prepared before the experiment started, informed foremen and supervisors of 'call' times, number of trays of mail to be expected with each 'call,' manpower schedules, inventory levels in the Primary and Secondary at certain critical times during Tour 3, and local rules for modifying the manpower and 'call' schedule if inventories exceeded stated amounts in each branch. In the early days of the experiment a flexible manpower pool was scheduled to process any unpredictably large fluctuations in mail inventories.

It is gratifying that the experiment was successful in the sense that mail inventories arrived in the Secondary within several minutes of their predicted times, that inventory pile-ups were small (1–5 trays of mail), that mail inventories were processed by the time manpower assignments were to be rescheduled, and that measurements of idle labor were small.

The 'After' Measurements

With the introduction of the new sorting rate assignments and Primary storage rules, mail volumes handled in time for each Secondary dispatch could be measured and average delays of letter mail could be calculated. For the purposes of measuring the reductions in average letter delay we observed mail flows to each one of twenty Secondary destinations. These destinations were individual cities (listed in Table V) rather than the Secondary mail categories of Fig. 20 because of the imprecise correspondence between the latter and areas served by trains, busses, or planes; for example, an entire Primary branch is not usually serviced by a single train.

Post office personnel weighed mail shipped nightly over two two-week periods in February and April of 1959. The measurements were made before and after the rescheduling of the Secondary. From these measurements it was possible to obtain the fractional mail volume leaving on each train for each city and to compare the figures with ones obtained earlier in an analysis of the 1957 'before' measurements. Average delay times were then obtained by integration of the area between the cumulative output curves and the curves of cumulative input to the initial dumping stages. As an example, the Canton, Ohio Secondary branch is shown in Fig. 24. The reduction in average letter delay is the area of the rectangles between the 1957 and 1959 output curves.

The accuracy of these calculations depended not only on the assumptions of 'constant mix' and constant conversion factors discussed earlier but also on the assumptions that mailing habits were unchanged and that

train and plane dispatch schedules had not undergone major revisions. Any improvements in collecting mail or any trend on the part of the public towards earlier mailing habits would have affected our calculations of average letter delay. In such a case our estimates of delay reductions could have been too high (hint of such an effect can be seen in Fig. 23). While we knew that there had been some changes in plane and train schedules and hence in the Secondary dispatch and tie-out times, we were not able to obtain detailed dispatch schedules for 1957. But this fact made it apparent that dispatch schedules had not been used in determining

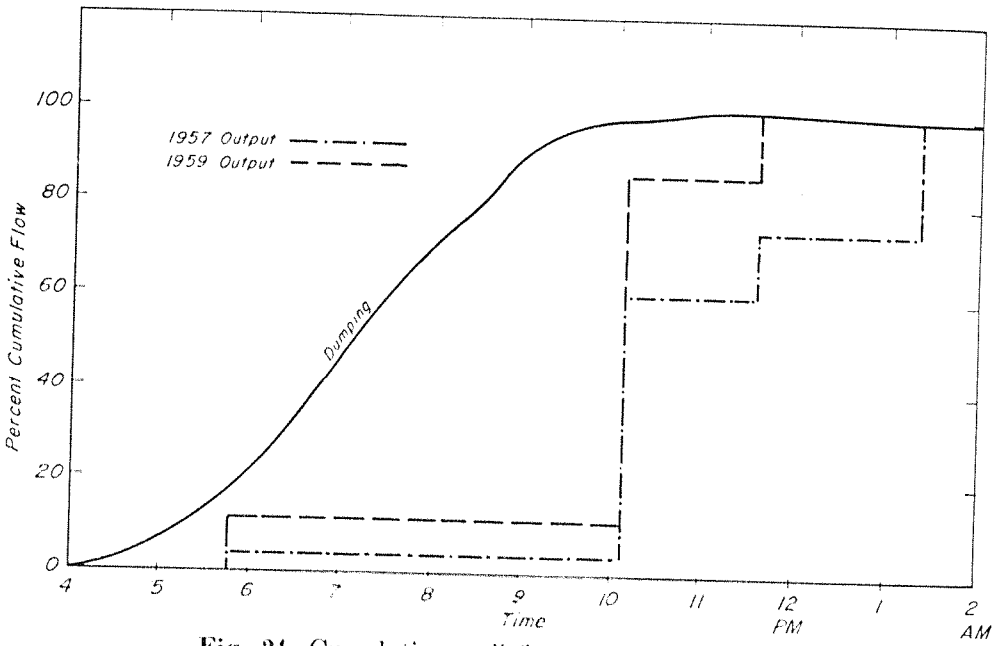


Fig. 24. Cumulative mail flows to Canton, Ohio.

optimal release times of Primary storage areas or manpower assignments in Secondary sorting areas. Since we also knew that flows out of the Secondary areas were relatively constant during Tour 3, we felt that the comparison of 'before' and 'after' measurements was justified.

Naturally, the delay reduction varies with the address or Secondary branch being considered. This statement follows from the nature of the Secondary dispatches. For example, if there were only one dispatch from a Secondary branch at 4 A.M., it is doubtful if the new schedules could have brought any reduction in average letter delay. Since this Secondary dispatch would come long after the peak mail flows and sorting operations in the post office, the entire mail volume destined for that dispatch would have been sorted under the old as well as the new schedules. On the other hand, large improvements were seen in those branches where one of several Secondary dispatches came close to the time of peak mail flows. The

general effect was that l
dispatches and smaller an

Table V shows the r
routed to the twenty city
for all destinations; num

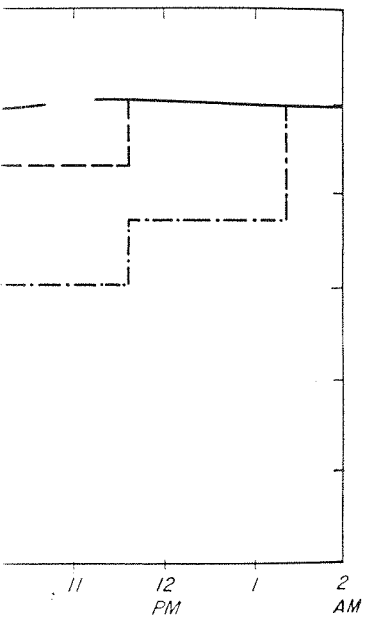
AVERAGE 1957 DELAY TIMES
AND IN

City (Secondary B
Albany, N. Y.
Baltimore, Md.
Benton Harbor, Mich.
Canton, Ohio
Charleston, W. Va.
Charlotte, N. C.
Chattanooga, Tenn.
East Lansing, Mich.
Erie, Pa.
Fort Worth, Texas
Grand Rapids, Mich.
Hamilton, Ont.
Jacksonville, Fla.
Jersey City, N. J.
Madison, Wis.
Salt Lake City, Utah
San Diego, Calif.
South Bend, Ind.
Traverse City, Mich.
Worcester, Mass.

average delay can be qu
median reduction in Tab

It was difficult, if not
delay reductions in the I
mailed in Detroit, i.e., th
addressee. Mail volum
times might be ready for
cities; hence, any increa
lead to delay reductions

undergone major revisions. nd on the part of the public ffected our calculations of imates of delay reductions ct can be seen in Fig. 23). hanges in plane and train and tie-out times, we were s for 1957. But this fact ot been used in determining



Canton, Ohio.

s or i apower assignments new that flows out of the ng Tour 3, we felt that the ts was justified.

the address or Secondary ows from the nature of the vere only one dispatch from ie new schedules could have Since this Secondary dis- vs and sorting operations in d for that dispatch would w schedules. On the other anches where one of several of peak mail flows. The general effect was that larger amounts of mail were processed by early dispatches and smaller amounts by the last dispatch.

Table V shows the results of average delay measurements of letters routed to the twenty city destinations. It is seen that delays are reduced for all destinations; numbers range from 0.36 to 2.69 hours. While no average delay can be quoted for the Detroit Post Office as a whole, the median reduction in Table V is approximately 25 per cent.

TABLE V

AVERAGE 1957 DELAY TIMES IN HOURS; APRIL 1959 DELAY TIME REDUCTION IN HOURS AND IN PER CENT OF 1957 DELAY TIMES

City (Secondary Branch)	1957 delay, hr	1959 delay reduction, hr	1959 delay reduction, %
Albany, N. Y.	3.87	1.06	27.4
Baltimore, Md.	3.34	1.48	44.4
Benton Harbor, Mich.	4.80	1.87	38.9
Canton, Ohio	3.96	1.17	29.5
Charleston, W. Va.	4.51	1.80	40.0
Charlotte, N. C.	4.15	1.83	44.0
Chattanooga, Tenn.	4.57	2.69	58.9
East Lansing, Mich.	4.05	0.65	16.0
Erie, Pa.	3.74	0.36	9.6
Fort Worth, Texas	3.25	1.12	34.6
Grand Rapids, Mich.	4.02	1.16	28.8
Hamilton, Ont.	3.99	0.45	11.2
Jacksonville, Fla.	4.02	1.12	27.7
Jersey City, N. J.	4.47	2.03	45.4
Madison, Wis.	4.80	0.58	12.0
Salt Lake City, Utah	4.05	1.00	24.7
San Diego, Calif.	4.46	0.98	22.0
South Bend, Ind.	4.39	0.92	21.0
Traverse City, Mich.	3.98	0.79	19.8
Worcester, Mass.	6.14	0.98	15.9

It was difficult, if not impossible, to calculate the effect of the average delay reductions in the Detroit Post Office upon the total delay of a letter mailed in Detroit, i.e., the total interval between mailing and receipt by the addressee. Mail volumes processed in Detroit for Secondary dispatch times might be ready for early morning delivery in nearby as well as distant cities; hence, any increase in volumes of mail for these dispatches could lead to delay reductions of one day for some letters. On the other hand,

increases in mail volumes for ill-timed Secondary dispatches might not affect the total letter delay.

We are aware of many imperfections in our data-gathering techniques and in our experiments; some of these were due to budgetary restrictions, some to the requirement that experiments not interfere with the flow of mail and, finally, some to our own lack of experience in postal systems. The experimental results should be considered only as approximations of the actual situations. Nevertheless, as a result of distinct theoretical and experimental checks of mail inventories, manpower counts, and letter delays, we believe that the results are meaningful.

CONCLUSIONS

AT A TIME when postal systems are under pressure from the mailing public, from the competition of additional modes of communication and transportation, and from increasingly stringent government fiscal policies, there is a need for accurate evaluation of operational and design problems. As mail volumes increase, it is undoubtedly true that automatic sorting devices will replace manual ones. As the need for faster communication arises, novel methods of processing and routing mails will be brought into use. As the population spreads from city to rural areas, new techniques of collecting, storing, and dispatching mails will be sought. Fast transportation and the need for quick mail service will make the operations of one postoffice more dependent on the operations of a distant one. Hence, as postoffices of the future are redesigned and relocated, it may become imperative to develop and control more centralized processing and storage operations. Contrary to the focus of an earlier day, less emphasis should be given to the design of isolated sorting devices and transport systems; more emphasis must be given to the complete system and the rules that organize its over-all behavior.

To obtain new operating rules and design criteria, postal management will be faced with the selection of courses of action from many possible alternatives. Along these lines, this study has reported the effects of certain feasible, suitable, and optimal operational decisions that reduce average letter delay. The processing, sorting, and storage decisions were obtained from several mathematical models and experiments. We feel that there are several important aspects of mail flows and delays that are contained in the body of the report; we also feel that there are at least five major conclusions that can be drawn.

The first of these relates to the timing and the frequency of dispatches of mail inventories. From the results in the third section we found that one or two well-timed dispatches would often result in smaller delays for letter mail than could many ill-timed dispatches. This, in principle, applies as much to collection times of mail boxes, postman delivery schedules,

metropolitan and commercia
it does to the dispatch probl
and Secondary sorting areas.
will automatically prepare la
livery and may create an ever
we have yet witnessed.

The second conclusion foll
saying that increases in sortin
the delays of mail waiting to
section also points out that
changed for an identical incre
dispatch. In other words, tl
letters that previously waite
now wait in storage in a diffe
role of processing and sorting
to understand their effects o

The third conclusion con
conveying device or transpo
It has not been uncommon
available transportation facil
riers that are only infreque
delays are more sensitive to t
of the carrier, the net result
One extreme example is the
based primarily on elapsed f
frequency and timeliness of d

The fourth conclusion is a
facilities. A more complete
as a partial substitute for t
hardware designs. While ne
day sorting and storage tech
understanding of existing flo
letter delays will provide imp
fact that rescheduling oper
achieved in a time period th
research, development, testi
ment.

Finally, any sorting, stora
the encouragement of new n
has been noticeable in all
might be incentives for early
increased use of pre-post-offi
new methods of coding add

ndary dispatches might not

ur data-gathering techniques
ue to budgetary restrictions,
ot interfere with the flow of
xperience in postal systems.
d only as approximations of
lt of distinct theoretical and
anpower counts, and letter
gful.

sur om the mailing public,
nmunication and transporta-
ent fiscal policies, there is a
d design problems. As mail
t automatic sorting devices
aster communication arises,
ls will be brought into use.
l areas, new techniques of
e sought. Fast transporta-
make the operations of one
f a distant one. Hence, as
relocated, it may become
lized processing and storage
r day, less emphasis should
ices and transport systems;
system and the rules that

riteria, postal management
action from many possible
as reported the effects of
ional decisions that reduce
and storage decisions were
nd experiments. We feel
il flows and delays that are
eel that there are at least
n.

the frequency of dispatches
hird section we found that
result in smaller delays for
. This, in principle, applies
ostman delivery schedules,

metropolitan and commercial inter-post-office transportation systems as it does to the dispatch problems within a post office, i.e., in the Primary and Secondary sorting areas. The introduction of fast sorting machines will automatically prepare large volumes of sorted mail for next day delivery and may create an even more critical need for timely dispatches than we have yet witnessed.

The second conclusion follows easily from the first one. It goes without saying that increases in sorting rates will reduce the long queues and hence the delays of mail waiting to be sorted. Embarrassingly enough, the third section also points out that savings in sorting times may be merely exchanged for an identical increase in the time that the mail waits for a fixed dispatch. In other words, the delay of letters may be the same as before; letters that previously waited for sorting and processing operations may now wait in storage in a different part of the postal network. Hence, the role of processing and sorting should be compared with that of mail storage to understand their effects on delays.

The third conclusion centers around the problems of the speed of a conveying device or transport system and the frequency of its departures. It has not been uncommon to find examples where slow but frequently available transportation facilities have been replaced by high-speed carriers that are only infrequently available. In those cases where letter delays are more sensitive to the availability of a dispatch than to the speed of the carrier, the net result may be an increase in average letter delay. One extreme example is the argument for 'mail-by-missile'; this has been based primarily on elapsed flight times without full consideration of the frequency and timeliness of departures.

The fourth conclusion is a plea for the best use of today's equipment and facilities. A more complete understanding of existing processes will serve as a partial substitute for the rush into new and sometimes ill-founded hardware designs. While new designs should not be restricted by present-day sorting and storage techniques, we also know that a more thorough understanding of existing flow patterns, storage and sorting policies, and letter delays will provide important benefits. More important still is the fact that rescheduling operations in the present system can often be achieved in a time period that is short compared to that required for the research, development, testing, and production of manufactured equipment.

Finally, any sorting, storage, and delay problems could be reduced with the encouragement of new mailing habits. The effect of a peaked input has been noticeable in all our calculations. Means of encouragement might be incentives for early (or late) mailings, new types of mail service, increased use of pre-post-office sorting techniques (such as metered mail), new methods of coding addresses, or even the relocation of mail boxes.

It will, however, be difficult to appraise the over-all effects of such measures until mail delays and the costs of providing various types of service are better understood.

ACKNOWLEDGMENTS

IN A research effort of this type there must be many acknowledgments. The entire project depended on the collaboration of United States Post Office personnel, both in Detroit and in the Office of Research and Engineering in Washington, D. C. We mean to slight no others if we mention by name especially MESSRS. H. H. KUSISTO, A. J. MICHAELS, and F. LEWANDOWSKI in Detroit, and MESSRS. FEIMSTER and J. N. LEWIS in Washington, D. C.

At Broadview Research Corporation we owe thanks to N. R. WALLACE and C. HANSON who were responsible for most of the data-processing problems and programming of the Alwac III-E computer; to W. S. JEWELL for helpful discussions, to B. RAGENT for numerical computations, to J. V. ZACCOR and J. H. BOYES for collation of experimental results. Finally, we owe many thanks to W. ALDEN for many interesting discussions and for the original research contract.

PERT AS AN ANAL
PLANNING—ITS

J

Booz·Allen Appli

(Rece

THE PAST decade has witn
analytical aids to planning.
solely to the injection of opera
it is clear that operations rese
development. In this area th
and PERT-like systems in pro
widely discussed.

The term 'PERT-like' is u
economy. In a recent survey s
have, at least at one time or an
the network approach and modi
of terminology has certainly bee
those on the technique-develop
pap er the term PERT as a gen

PERT—A TO

MANAGEMENT has long been s
control more effective, particular
tions, and relations is involved.
PERT strives, although noble, is
it may not be the ultimate answ
not—three years' practice indicat
and being so, warrants our contin

Since we are speaking here of
place ourselves on management's
attempt to develop a broad over
management is thus far. This is
we must certainly remind oursel
technical minded, work with the

* Presented at the Tenth Annive
SOCIETY OF AMERICA, Washington, D.