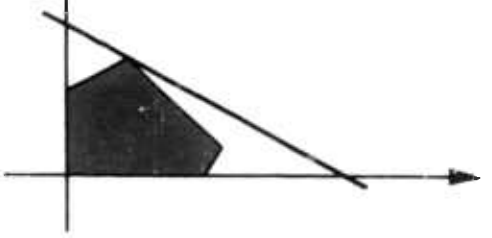
AD 261 850

RESEARCH REPORT 11
14 July 1961
I.E.R. 172-12

REDUCING LETTER DELAYS IN POST OFFICES

by

R. M. Oliver and Aryeh H. Samuel



OPERATIONS RESEARCH CENTER

INSTITUTE OF ENGINEERING RESEARCH

UNIVERSITY OF CALIFORNIA - BERKELEY

REDUCING LETTER DELAYS IN POST OFFICES

by

Robert M. Oliver* and Aryeh H. Samuel**

14 July 1961 Research Report 11

This research has been partially supported by the Operations Research Center and the Institute of Transportation and Traffic Engineering of the University of California at Berkeley. Most of the work reported here is the direct result of a sub-contract awarded to Broadview Research Corporation, Burlingame, California by the prime contractor, Automation Management, Incorporated, Westboro, Massachusetts through the intermediacy of Bruce Payne and Associates, Boston, Massachusetts. The original contract was awarded by the Office of Research and Engineering, United States Post Office, Washington, D. C.

* Now with the University of California, Berkeley.

** Now with Stanford Research Institute, Menlo Park, California.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

This paper reports a group of mathematical models and experiments which 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 which 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.

This paper includes a theoretical and numerical analysis of letter delays as well as a description and evaluation of a series of full-scale experiments performed at one of the larger United States Post Offices.

1. INTRODUCTION

Summary and Objectives

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 which were obtained in the course of experiments at one of the larger United States Post Offices.

This paper includes a discussion of flow problems which arise within the confines of mail sorting and classifying operations, i.e., intra-post office 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 had been less systematized, and operating rules can be found which 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 which 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 important variables: the arrival time, the inhomogeneous nature of the mail stream entering a post office, 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 leaves 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 which transport mail from place to place within a post office.

A letter enters the post office, is dumped in loose or packaged form and is transported to and through one or more serial dumping, culling, priority sorting, facing, stamp-cancelling and counting stages.

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 which 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 other sorting systems:

[†] See Figure 1.

successive post offices or perhaps the ultimate addressee of the mail. The time which 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 which 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 width of the peaks spread out and 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% 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 which, 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 which arise in the serial production process, our analysis of the operational problems centered around deterministic queueing 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 which arises is that of scheduling manpower and machines to sort the peak flows as they propagate through the many processing stages. Intuitively, both of us felt, as did many postal supervisors, that delays would be reduced when sorting and processing rates matched the mail flows at each stage. That is to say, a plot of processing rates should closely parallel mail flow rates at each stage. It is interesting to note that this intuitive solution is correct only so long as strict inequality restrictions on sorting and processing rates are operative.

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 finally sorted in say 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 which carry the mail to another post office. In 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 elegant 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 expects to find intervals where the decision variables, i.e., sorting rates, lie on the edge of the constraint region followed by intervals where the classical Euler equations are satisfied within the interior of the region. Many of the postal scheduling problems represent an interesting sub-class of derivative-constrained variational processes in that the classical Euler equations may be satisfied only at an isolated number of points; on the other hand, one or more inequality constraints are always operative in any interval of time.

Contents of Paper and Notation

In addition to this introductory section, the paper is divided into three parts. Section 2 studies the scheduling problems in a network which 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 which minimize average letter delay are obtained and illustrated.

In Section 3 attention is given to the effects of storage and some of the new scheduling and dispatching problems which naturally arise.

In Section 4 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.

In Section 5 a summary of the major objectives achieved by this research is presented.

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 $0, 1, \ldots, j$ referring to the j^{th} 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 i^{th} 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.

2. MODELS OF SERIAL PROCESSING AND SORTING STAGES

Although mail arrives at a post office loose or packaged in bundles of various sizes and shapes, the bulk of it must go through a predetermined number of serial processing stages. In the early stages letters are prepared for their journey through manual or automatic sorting schemes.

These early processing stages consist of one or more dumping and weighing operations where mail flows into the main stream, culling or sieving operations where non-first-class mail and other bulky items are rejected from the main stream, facing operations where letters are oriented in preparation for stamp cancellations, some simple sorting operations which separate priority from non-priority mails, and finally the more complicated sorting operations where addresses are subdivided into a large number of categories. Although there are many variants to this routing pattern, a large fraction of mail flows can be analyzed in terms of serially connected sorting stages where branching also occurs. Figure 1 is a schematic diagram of a typical flow pattern. Figures 21 and 22 in Section 4 are diagrams of an actual flow pattern at one of the larger United States Post Offices.

If there is little stochastic variation in the input flows to a number of serially connected sorting stages and if processing rates at each stage are known functions of time, mail inventories and letter delays can be calculated at each stage. One's intuitive picture of the relation between letter delays and sorting rates is certainly a simple one. Mail inventories can be reduced to essentially zero by assigning to each stage a processing rate that is equal to the input rate. Inventories and delays of mail arise when cumulative input flows exceed cumulative amounts processed as a result of restrictions on processing rates.

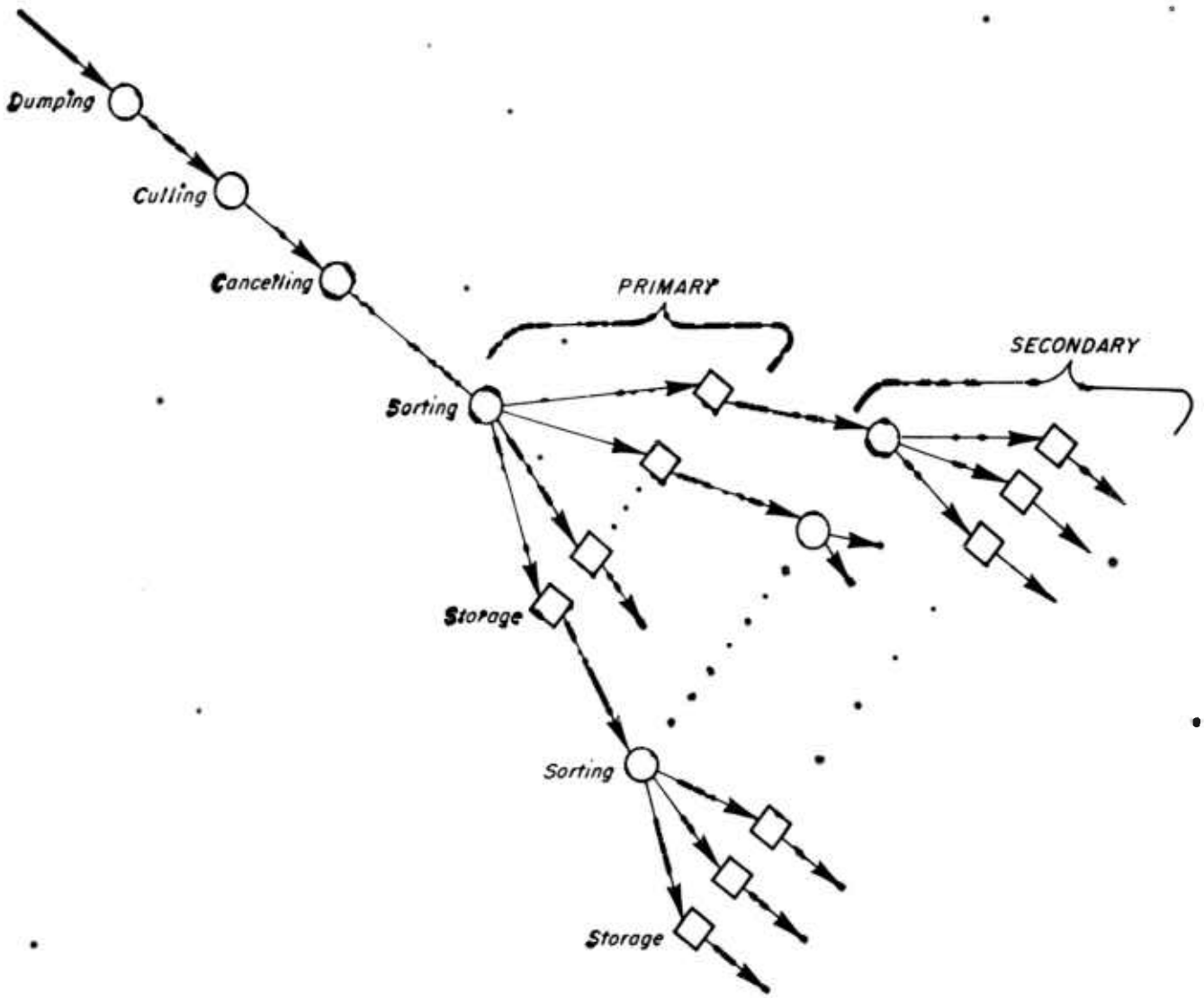


Fig. 1 - Flow Diagram.

Individual stage restrictions are due to hardware or plant limitations, such as the maximum flow capacity of a conveyor or the maximum seating capacity around tables where mail is faced, culled or sorted. Restrictions on the total number of men or machines which can sort or process mail but which can be assigned to any one of a number of stages give rise, mathematically, to inequality restrictions on linear sums of processing rates.

The Delay of a Letter

Consider first of all, some of the flow characteristics of two serial processing stages in Figure 2. All of the mail which 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 which enter either stage 1 or 2 at time t.

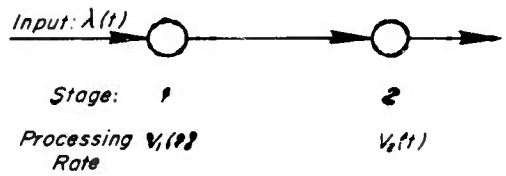


Fig. 2-Two serial processing stages.

A letter entering stage 1 at time t enters the second stage queue at time $t + \tau_1(t)$. A letter which 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,

$$\tau(t) = \tau_1(t) + \tau_2(t + \tau_1(t)) \qquad (2.1)$$

It is as important to note that the argument of the delay in the second term is $t + \tau_1(t)$ as it is to remind oneself that these delays are functions of the processing rates, $v_1(t)$ and $v_2(t)$, and the shape of the input rate, $\lambda(t)$.

One surprising result which will be obtained shortly is that the total system delay, $\tau(t)$, is only an implicit function of the cumulative input to the first stage and the cumulative output of the last stage.

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.

As we have said, a letter which enters the first stage at time t leaves at $t + \tau_1(t)$; all letters preceding the letter which 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.

$$\int_0^t \lambda(s)\,ds = \int_0^{t+\tau_1(t)} v_1(s)\,ds \qquad (2.2)$$

and for the delay $\tau_2(t)$, similarly:

$$\int_0^t v_1(s)\,ds = \int_0^{t+\tau_2(t)} v_2(s)\,ds$$

If we let $\Lambda(t)$, $V_1(t)$, $V_2(t)$ be the cumulative input and processing flows we can write Equation (2.2) in terms of cumulative flows,

$$\Lambda(t) = V_1(t + \tau_1(t)) \qquad (2.3a)$$

$$V_1(t) = V_2(t + \tau_2(t)) \qquad (2.3b)$$

The solution for $\tau(t)$ is found by substituting Equation (2.1) and (2.3a) into (2.3b).

$$\Lambda(t) = V_2(t + \tau(t)) \tag{2.4}$$

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:

$$\tau(t) = k^{-1}(\Lambda(t) - \Lambda(s_1)) - (t - s_1) \qquad t \in S \tag{2.5}$$

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 Figure 3 and the indicated values of the constant processing rate, k, are used. Case (I) corresponds to an early peak, Case (II) to a late peak in the input rate. We 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 Figure 3.

As another example we note that a constant delay, τ_0, could be obtained in stage 1 if the processing rate, $v_1(t)$, lagged the input rate by τ_0, i.e., if $\lambda(t) = v_1(t + \tau_0)$. On the other hand, a scheduling policy which maintains a constant inventory, I, in front of this stage gives rise to a delay $\tau(t)$ which is the solution of $V_1(t + \tau(t)) - V_1(t) = I$.

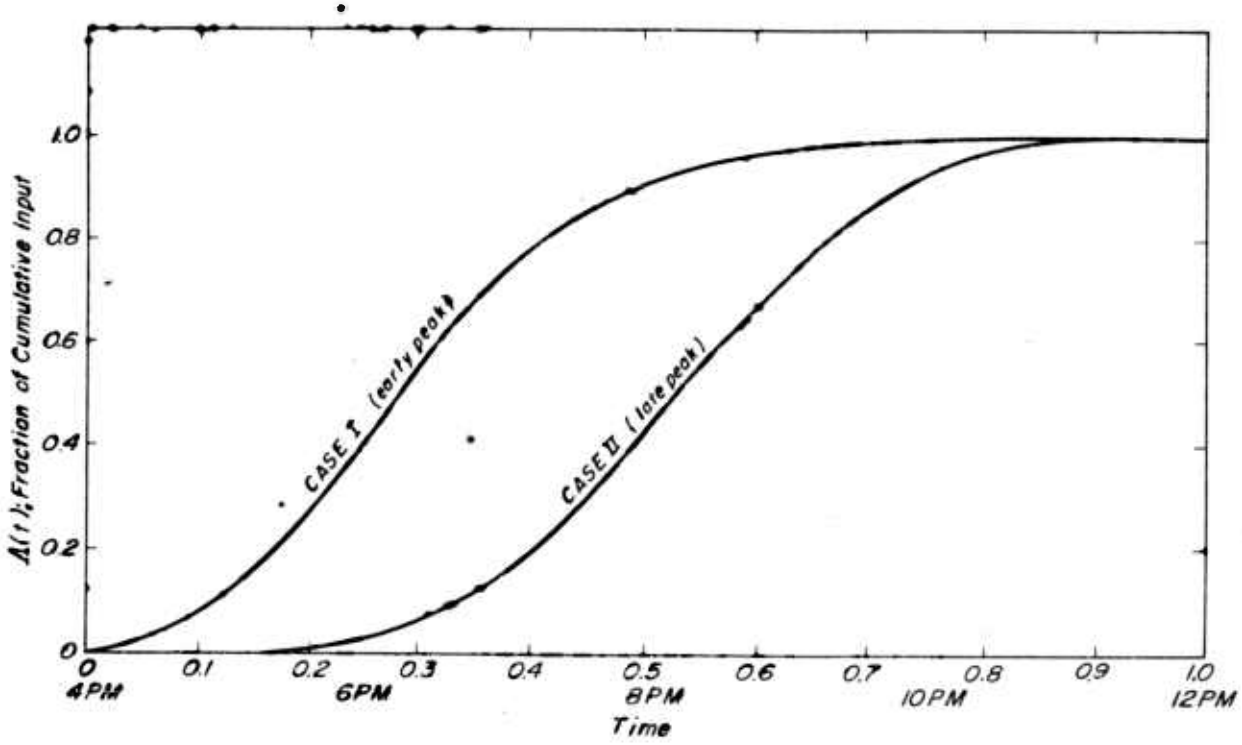


Fig. 3- Cumulative input as a function of time.

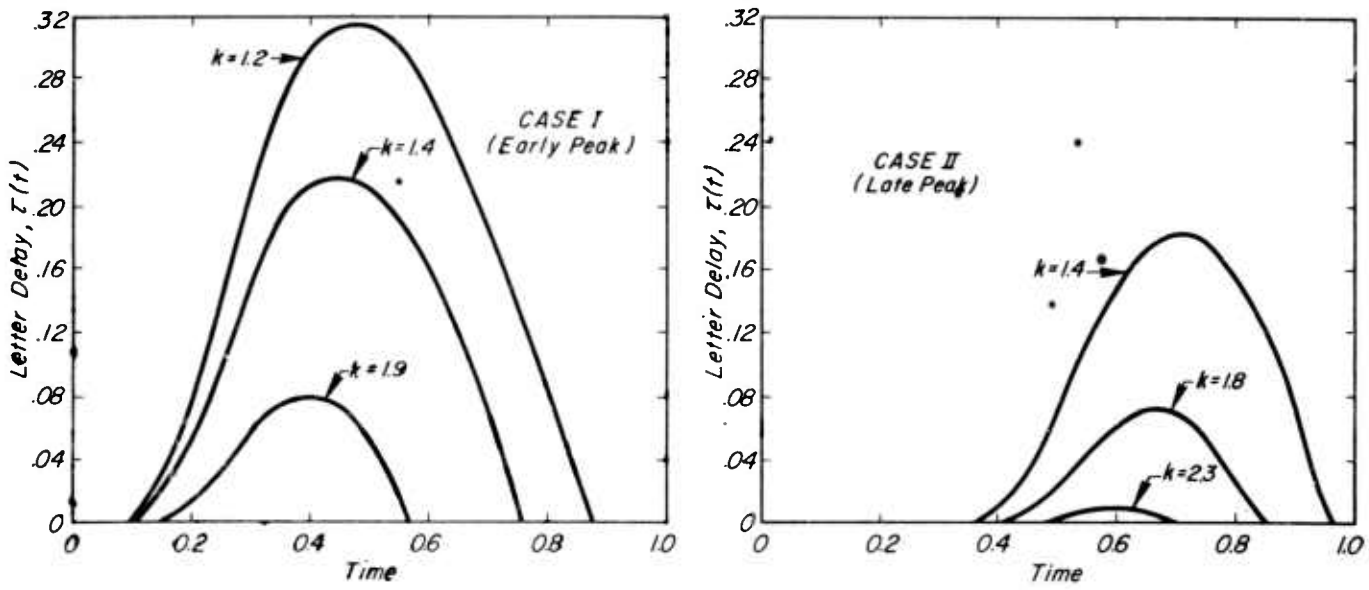


Fig. 4–Letter delay as a function of time (constant processing rate, k).

Average Letter Delay

As we mentioned in the introduction, we will analyze the effect of various processing and sorting rates upon 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 $t+dt$ is $\tau(t)$ we can write the average delay as,

$$D = \int_0^1 \lambda(t)\, \tau(t)\, dt \tag{2.6}$$

D is always greater than or equal to zero when we impose non-negative processing rates and inventory restrictions at each stage; i.e., if

$$v_1(t),\ v_2(t) \geq 0 \tag{2.7a}$$

$$\Lambda(t) \geq V_1(t) \geq V_2(t) \tag{2.7b}$$

the solutions of Equations (2.3) and (2.4) lead to non-negative letter delays.

If processing rates are unrestricted, we can easily satisfy the equality restrictions in Equation (2.7b). The solutions for $\tau_1(t)$, $\tau_2(t)$ and $\tau(t)$ in Equations (2.2), (2.3), and (2.4) can then be made zero; hence the average

---

[†] With these normalizations the dimensionless values of D (average delay), $\tau(t)$ (the delay of a letter), and other time-dependent variables yet to be defined need only be multiplied by the actual length of the interval to obtain the real value of D, $\tau(t)$.

delay of letters traveling through the two-stage system of Figure 2 can also be made zero. However, restrictions on the sum of processing rates,

$$v_1(t) + v_2(t) \leq k \tag{2.8}$$

and, more generally, of the form

$$v_1(t) + av_2(t) \leq k(t) \tag{2.9}$$

play an important part when the total sorting rate at any time t is capacity restricted. Manpower pools, though interchangeable between processing stages, may be restricted in total size by the availability of skilled personnel. Budgets may restrict the hiring of unskilled labor. Even semi-automatic and fully automatic machines may be limited in their sorting and processing capacities.

Before we obtain solutions which minimize average letter delay under these, and other, more severe restrictions, we want to briefly sketch the intuitive procedures which 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 Figure 2 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 (2.6) is a simple variational problem: find that (feasible) processing rate assignment $v_1(t)$, $v_2(t)$ which makes D a minimum while satisfying the inequality constraints of (2.7) and (2.8). 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 (2.6). Secondly, the inequality constraints (2.7) and (2.8) will lead to optimal processing rate solutions which, 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. Conceptually, the switch-over process may be simple. Algebraically, the exact point of transition may be difficult to compute.

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 (2.7) and (2.8) becomes a strict equality. It is fortunate, as we will soon see, that in the majority of our scheduling problems one need only divide the interval (0, 1) into two types of sub-intervals 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; $\Lambda(t) > V_1(t) = V_2(t)$.

We use the classical notation $\delta x(t)$ to denote a variation in the function $x(t)$ while holding t constant. The variation in average delay of Equation (2.6) is then given by the linear functional,

$$\delta D = \int_0^1 \lambda(t)\, \delta\, \tau(t)\, dt \qquad (2.10)$$

If the delay of a letter $\tau(t)$ is related to a sorting rate $v(t)$ by an implicit functional equation of the form

$$g[\tau(t),\ v(t),\ t] = 0 \tag{2.11}$$

then the variation in $\tau(t)$ due to small variations in $v(t)$ is simply

$$\delta\tau(t) = \frac{\partial g}{\partial v}\left(\frac{\partial g}{\partial \tau}\right)^{-1} \delta v(t) \tag{2.12}$$

By straightforward differentiation of Equation (2.2), we find that $v_2(t)$ can be expressed in terms of the processing rates.

$$\delta\tau(t) = \frac{-\int_0^{t+\tau(t)} \delta v_2(s)\, ds}{v_2(t+\tau(t))} = \frac{-\delta V_2(t+\tau(t))}{v_2(t+\tau(t))} \tag{2.13}$$

whenever the demoninator does not equal zero.

It is interesting to note that any change in the sorting rate in the interval $(0,\ t+\tau(t))$ affects $\tau(t)$, the delay of the letter entering at time t. In other words, variations in processing rates before or after the arrival of a letter in the system may have equal effects on the delay of that letter.

The Region $\bar{S}$

Assume that $\lambda(t)$ is always greater than zero in the interval (0, 1). In Equation (2.10) 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

$$\tau(t) = 0; \quad \lambda(t) = v_1^*(t) = v_2^*(t) \qquad t \in \bar{S} \qquad (2.14)$$

where the star (*) denotes the optimal processing rates.[†] It is also true that if Equation (2.14) holds in $\bar{S}$ then

$$\lambda(t) = v_1^*(t) = v_2^*(t) \qquad t \in \bar{S} \qquad (2.15)$$

Since $v_1^*(t) + v_2^*(t) \leq k$ from Equation (2.8), we find that

$$\lambda(t) \leq k/2 \qquad t \in \bar{S} \qquad (2.16)$$

is a necessary condition for the optimal solutions to be given by Equation (2.15). That is to say, the input rate must be less than or equal to one-half the processing rate capacity.

[†] The asterisk (*) will always be used in this paper as a superscript to denote optimum decision variables or the minima or maxima of functions.

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 want to 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.

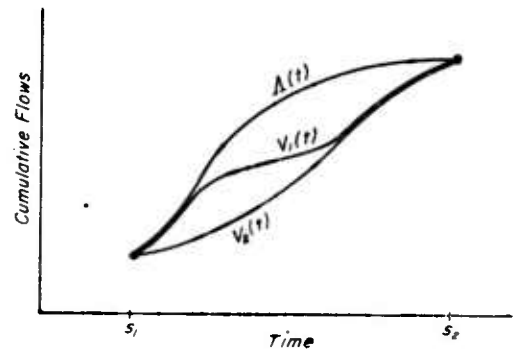


Fig.5 - Feasible cumulative processing rates.

If $\tau(t)$ is greater than zero in the optimal program, we expect that 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 Figure 5 we see that any one of the three following solutions for $V_1(t)$, $V_2(t)$ is feasible when $\Lambda(t) > V_2(t)$:

$$\Lambda(t) > V_1(t) = V_2(t) \tag{2.17a}$$

$$\Lambda(t) > V_1(t) > V_2(t) \qquad t \in S \tag{2.17b}$$

$$\Lambda(t) = V_1(t) > V_2(t) \tag{2.17c}$$

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,

$$v_1(t) + v_2(t) = k \qquad t \in S \tag{2.18}$$

Since letter delays in $\bar{S}$ are zero, variations in average letter delay of Equation (2.10) can be rewritten

$$\delta D = \int_S \lambda(t)\, \delta\tau(t)\, dt \tag{2.19}$$

If we substitute Equation (2.13) for $\delta\tau(t)$ and make the change of variable $r = t + \tau(t)$ we get[†]

$$\delta D = - \int_S \frac{\lambda(t)\, \delta V_2(t + \tau(t))}{v_2(t + \tau(t))}\, dt \tag{2.20a}$$

$$= - \int_S \delta V_2(r)\, dr \tag{2.20b}$$

Positive variations in $V_2(t)$ decrease D and from Equation (2.18) we also know that positive variations in $v_2(t)$ equal negative variations in $v_1(t)$. Hence

$$\delta V_2(t) = - \delta V_1(t) \qquad t \in S \tag{2.21}$$

To obtain a minimum average delay program we increase $V_2(t)$ and decrease $V_1(t)$ 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,

$$v_1^*(t) = v_2^*(t) = \frac{k}{2} \qquad t \in S \tag{2.22}$$

In the optimal program, only the first line of Equation (2.17) holds; inventories and delays are concentrated at the first stage and there are no delays and no inventories at the second stage. When k is replaced by k(t), we

[†] r is just the exit time of a letter arriving at t.

impose a capacity rate restriction which varies with time. Again, the optimal assignment of processing rates is one which splits the time-varying capacity rates between both stages so that,

$$v_1^*(t) = v_2^*(t) = \frac{k(t)}{2} \qquad t \in S \qquad (2.23)$$

Proof of this schedule follows directly from the variational arguments.

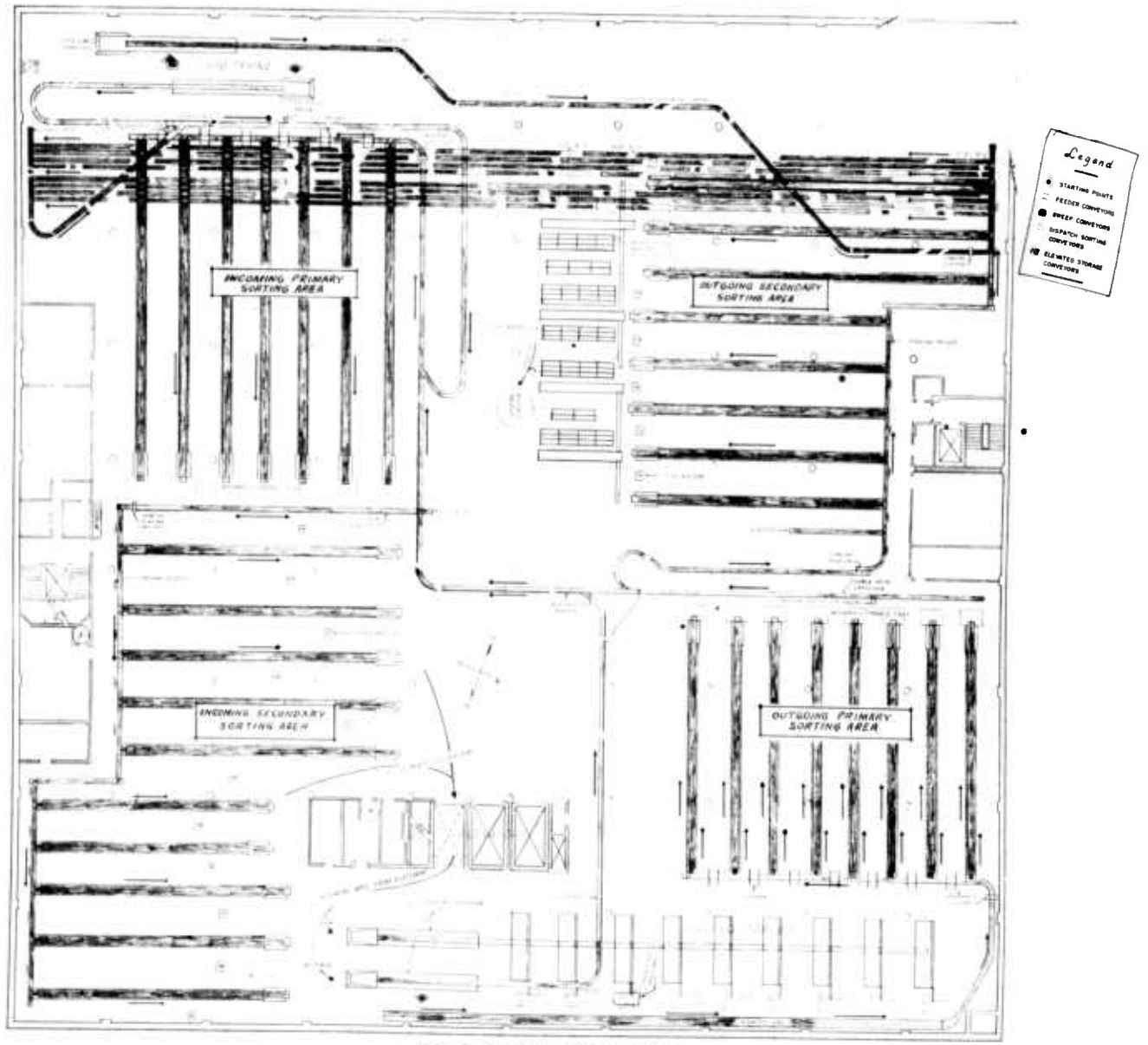


Fig. 22 - A MAIL-FLO SYSTEM.

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 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.[1] 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 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 account for little time and are 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 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.

---

[1] Throughout Section 4 a "call" is equivalent to a release or dispatch time of Primary storage. Other post office terms such as "cases", "sweeps", "tie-outs", "mix" are explained when they are introduced in the text.

A detailed breakdown of the flow volumes observed through various branches is shown in Figure 23; at this point it is important to point out that the total volume of mail in Tour 3 ranged from 1.0 to 1.5 million letters daily excluding weekends. This volume indicates that the Detroit office numbers in the top ten United States Post Offices.

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 which 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 Section 3. It was then a simple matter to draw up a set of rules which 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 and development of a manual which would describe 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.

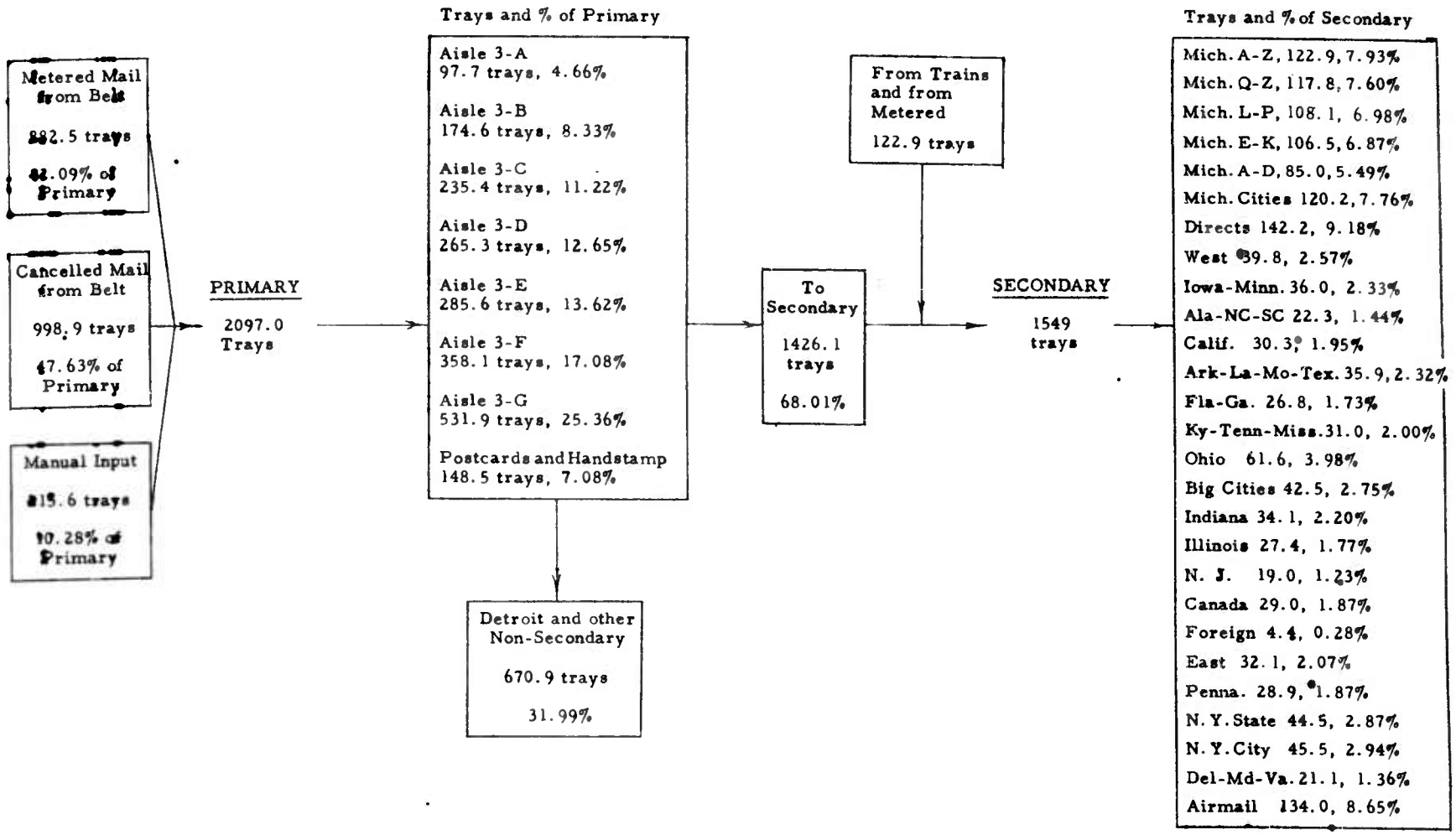


Figure 23: Mail Flows Into Primary and Secondary (Based on Averages 9/30/57 - 11/12/57)

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. See Equation (2.32). 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 a certain fraction of individual letter count; along conveyors, mail was sent in trays. We were furnished with the following conversion factors experi-

mentally determined by Detroit personnel: 43 letters = 1 lb., 580 letters = 1 tray. A more careful analysis showed that individual trays contained 8.5 to 16 lbs. and 350 to 1200 pieces, with an approximately Gaussian distribution about the above averages.

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 Figure 23 lists the average daily flow volumes of mail sorted in the Primary and Secondary during the September-November period in 1957. Figure 24 shows cumulative output flows and Figure 25 shows input and output flows of a typical Secondary.

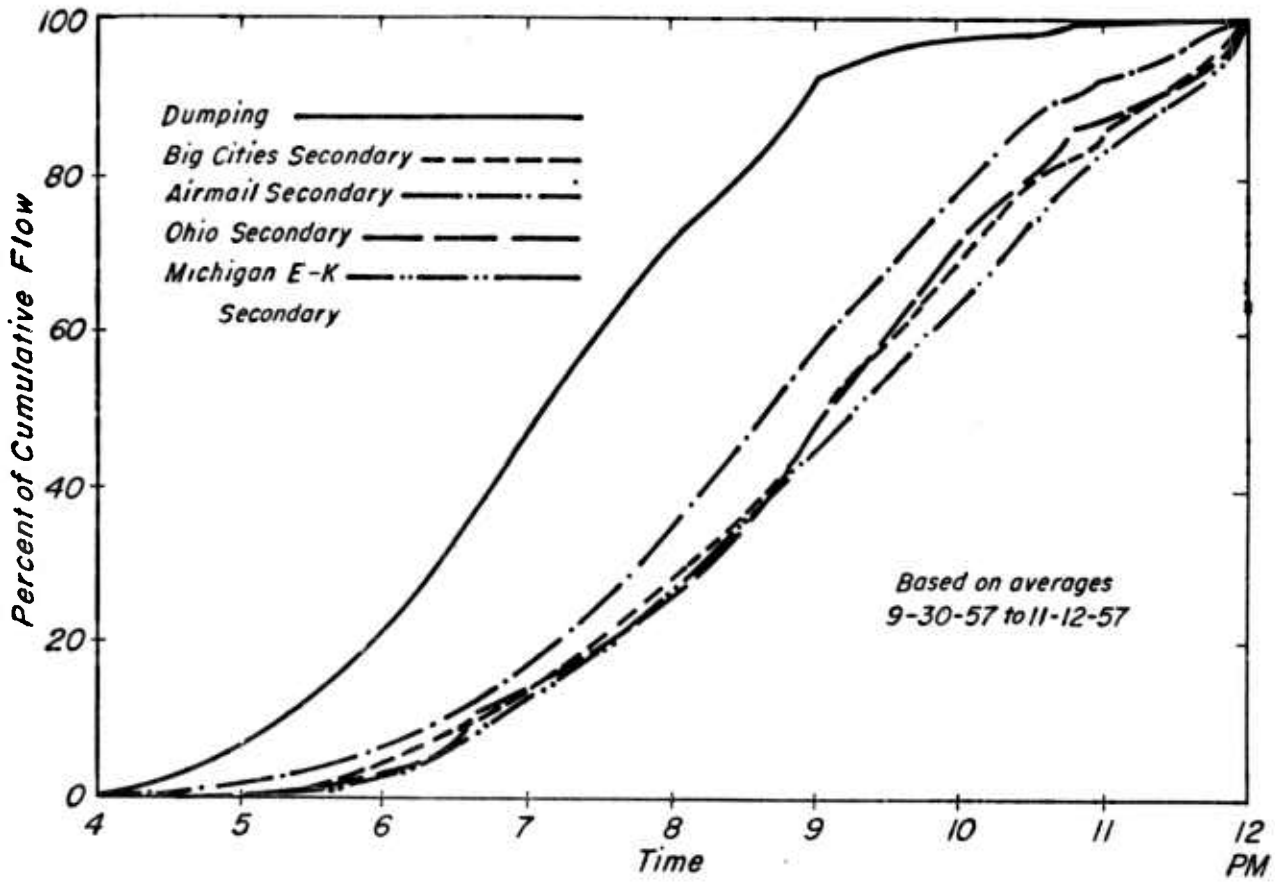


Fig. 24 - Cumulative output of dumping stage and selected secondaries.

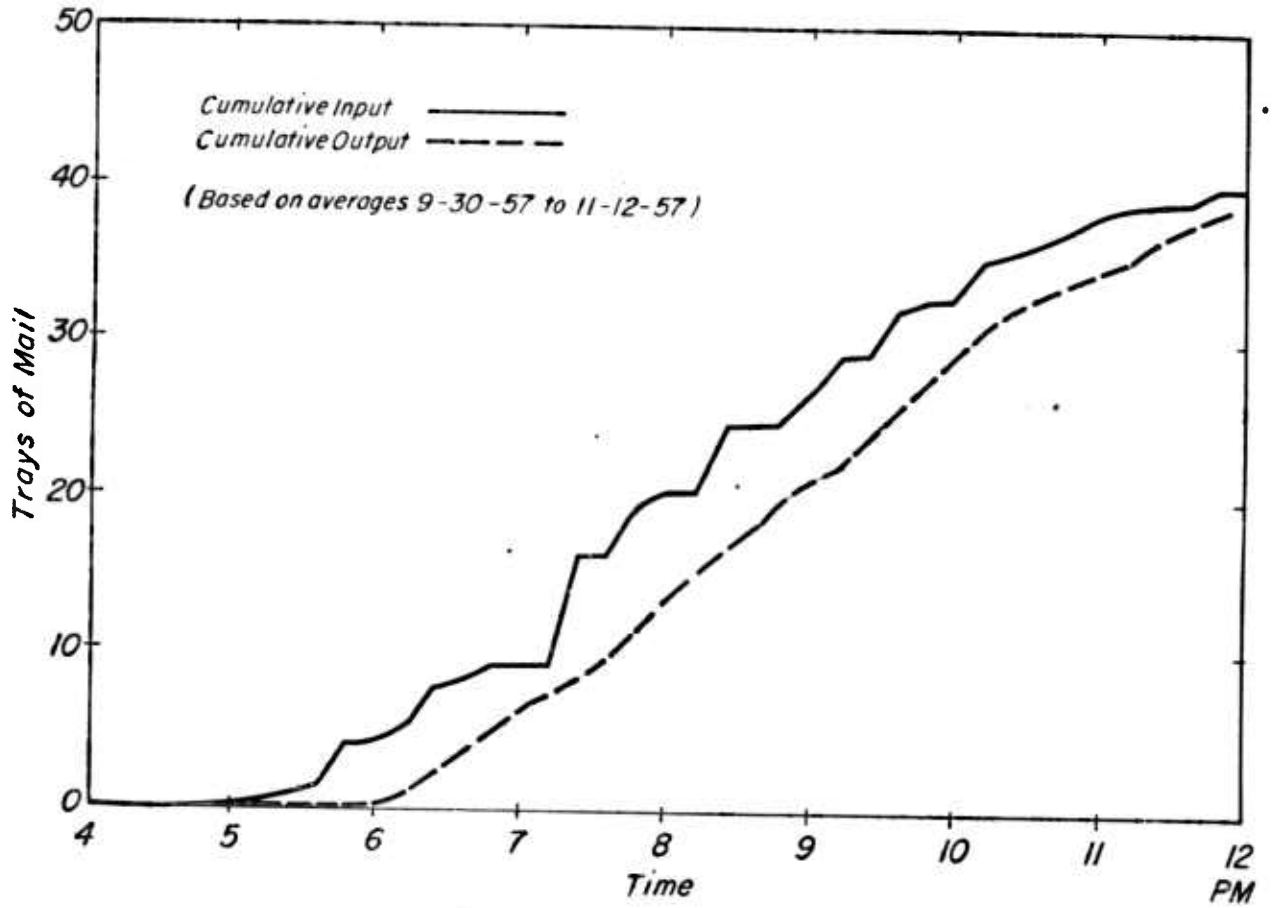


Fig.25 – Cumulative input and output flows, West Secondary.

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 weighed over 2 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 2. The large delays and inventories of mail were attributed to overanxiety on the part of the foremen to avoid idle labor charges. Due to the constant checks on mail processing costs and the almost absolute lack of checks on mail processing delays, the schedules reflected manpower assignments which 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 Section 2. 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 3.[7] 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 mail inventories.

The simple models of Section 2 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

[7] No figures were available for manpower assignments at the facing tables prior to 1958.

Table 2. 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) 9/30-10/3/1957. 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) 9/30-10/3/1958. Mail on belt included.

Table 3. Manpower Assignments at Facing Tables and Primary, (Sept. 30 and Oct. 1, 1957 and 1958)

	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		

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 1000 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 Figure 24 and Tables 3 and 4. Figure 26 shows that the average letter delay up to the end of the Primary was reduced by about 0.8 hours. Table 3 shows that capacity sorting rates in the Primary were reduced earlier while Table 4 shows that input flows to the Secondary peaked earlier than before the experiment. Because the "call" schedule had not yet been rescheduled this measurement of Secondary input flows is evidence of a corresponding change in Primary output flows.

	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

Table 4. Mail Flows (pounds), into Secondary Before and After Rescheduling of Primary.

[†] Refer to the paragraph following Equation (2.5).

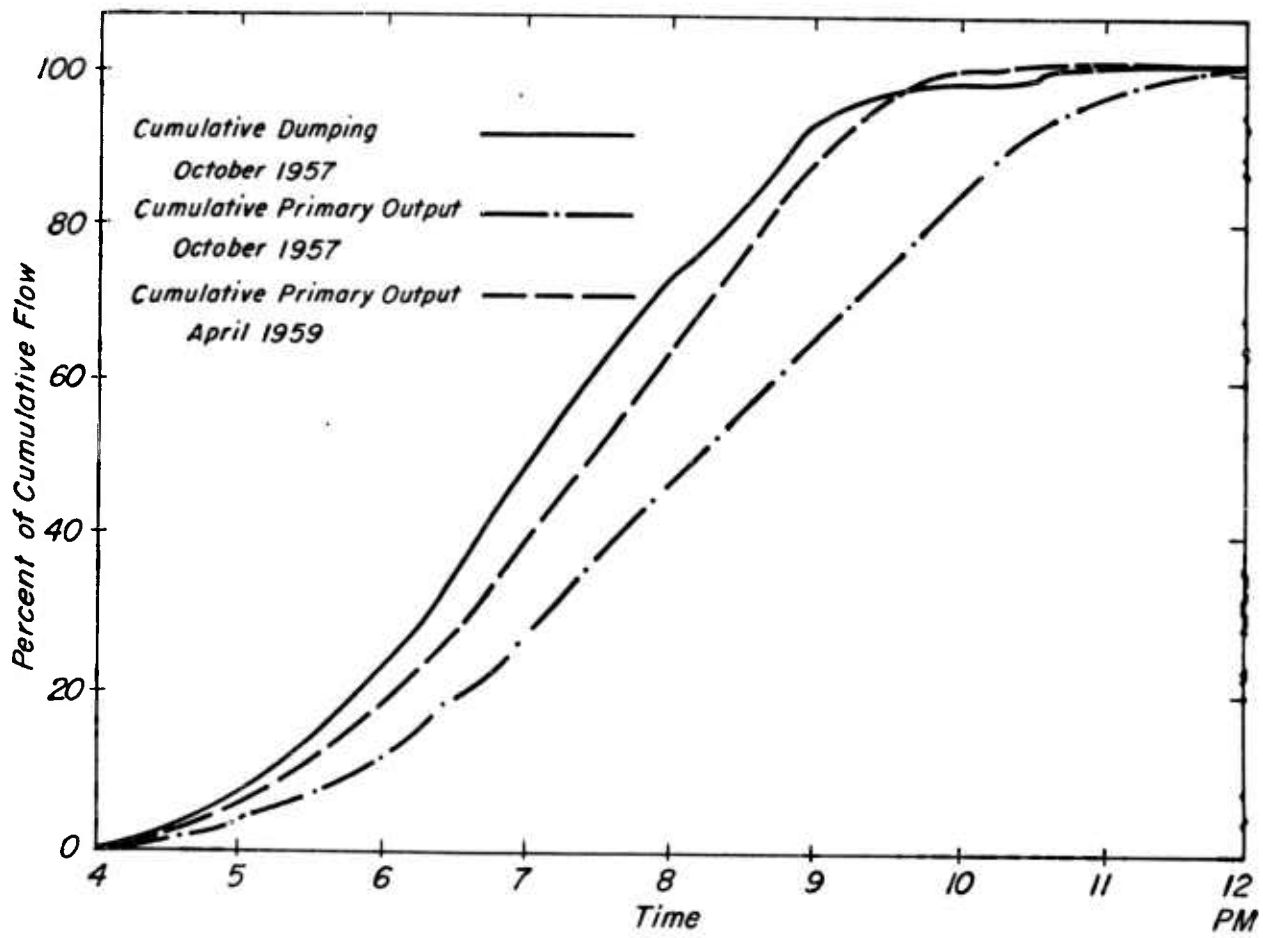


Fig.26 -Cumulative Primary output in 1957 and 1959, cumulative output of Dumping stage in 1957.

It is appropriate to mention one feature of the system which 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 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 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 which 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.
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.

It is important to mention again that 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 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 which 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 Figure 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% of the evening's mail would make the fixed Secondary

† Airmail is processed in an Airmail Secondary.

†† T^* is just the solution of $\Lambda(T) = 3.00\ (0.76 - T)$; in general we refer to Equation (3.18) for T_j^*, the optimal timing of the j^{th} dispatch.

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% of 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% 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.

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 Secondary sorting area. A typical sheet is shown in Table 5 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 Figure 3).

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.

Table 5. Schedule for California Secondary

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 6) rather than the Secondary mail categories of Figure 23 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 Figure 27. 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.

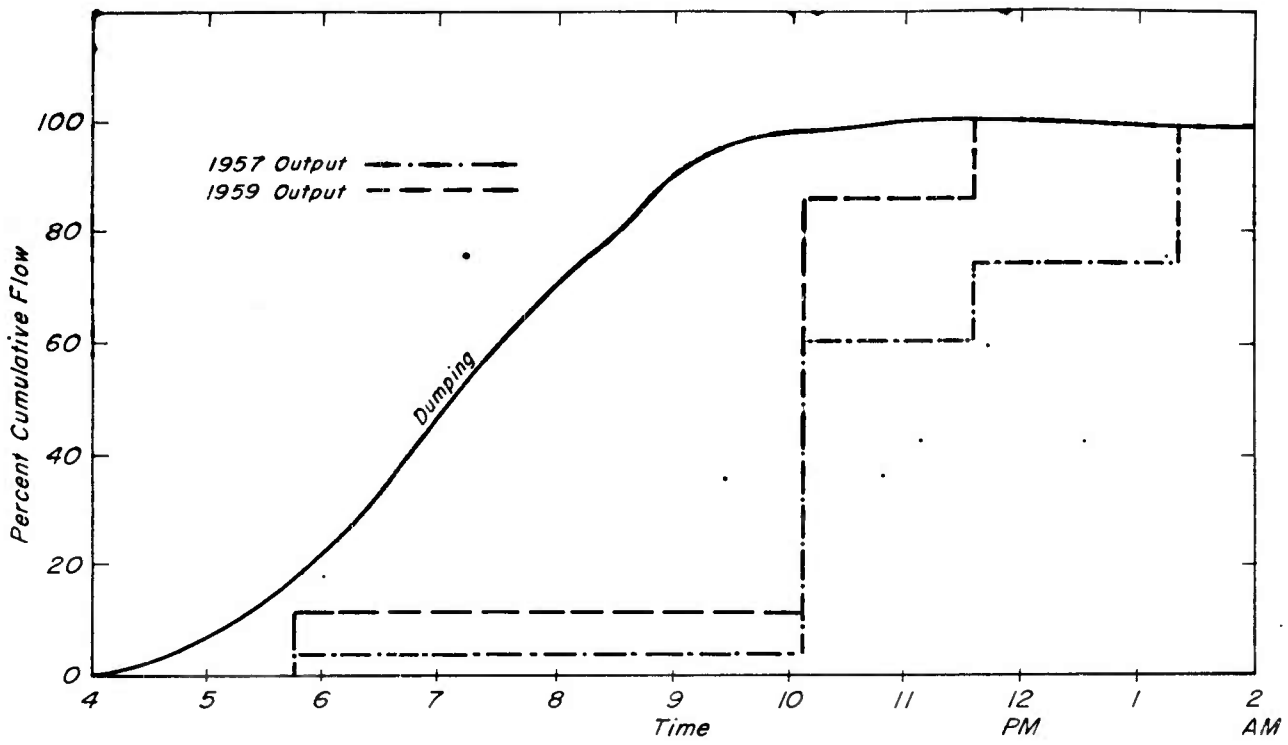


Fig.27 - Cumulative mail flows to Canton, Ohio

In such a case our estimates of delay reductions could have been too high (hint of such an effect can be seen in Figure 26). 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 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. in the morning, it is doubtful if the new schedules could have brought any reduction in average letter delay. Since this Secondary dispatch came 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 larger amounts of mail were processed by early dispatches and smaller amounts by the last dispatch.

CITY (Secondary Branch)	1957 Delay (Hours)	1959 Delay Reduction (hours)	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

Table 6. Average 1957 Delay Times in Hours; April 1959 Delay Time Reduction in Hours and in Per Cent of 1957 Delay Times.

Table 6 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 6 is approximately 25%.

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. 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.

5. CONCLUSIONS AND ACKNOWLEDGMENTS

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 post office more dependent on the operations of a distant one. Hence, as post offices 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 will be given to the design of isolated sorting devices and transport systems; more emphasis will be given to the complete system and the rules which 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 same lines this operations research study has reported the effects of certain feasible, suitable, and optimal operational decisions which 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 which are contained in the body of the report; we also feel that there are at least five major conclusions which can be drawn.

The first of these relates to the timing and the frequency of dispatches of mail inventories. From the results in Section 3 we found that one or two well-timed dispatches would often result in smaller delays for letter mail than could many ill-timed dispatches. The theory applies as much to collection times of mail boxes, postman 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, Section 3 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 which 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 which 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 noteable example is the argument for "mail-by-missle"; this has been based primarily on elapsed flight times without full consideration of the frequency and timeliness of departures.

The fourth conclusion is written as a plea for the best use of today's equipment and facilities. It can be argued that 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, the authors feel that a more thorough understanding of existing flow patterns, storage and sorting policies, and letter delays will provide many benefits. More important still is the fact that rescheduling operations in the present system can often be achieved in a time period which is short compared to that required for the research, development, testing and production of manufactured equipment.

The fifth and final conclusion is reached by only one of us.† Many sorting, storage and delay problems could be reduced with the encouragement of new mailing habits. 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.

† This paragraph does not reflect the views of A. H. Samuel.

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 own many thanks to W. Alden for many interesting discussions and for the original research contract.

Corner Conditions and Average Minimum Delay

Up to this point we have discussed only the structure of the optimal processing rates in the capacity constrained regions S and in the unconstrained regions $\bar{S}$. The next questions which must be answered are (i) where are the beginning and end points of S and $\bar{S}$ and (ii) what conditions must the processing rates satisfy at these points? A simple physical argument seems particularly appropriate at this time.

Let us assume that $\lambda(t)$ starts at zero, increases, attains a single maximum and falls off to zero before or at 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 sub-intervals, $\bar{S} = (0, s_1)$; $S = (s_1, s_2)$ and $\bar{S} = (s_2, 1)$. $S = (s_1, s_2)$ is the capacity constrained region where maximum processing rates are equally divided between the two serial stages, and s_1 and s_2 denote the beginning and end points of this interval.

As $\lambda(t)$ increases from zero, t will reach the point s_1 where for the first time the capacity restriction of Equation (2.8) becomes a strict equality; mathematically, s_1 is the smaller root of

$$\lambda(s_1) - \frac{k}{2} = 0 \qquad (2.24)$$

For the two-stage case this corresponds to the simple graphical solution where the tangent of the cumulative input flows is first equal to $\frac{k}{2}$. 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 maximum at a point in time which is the larger root of Equation (2.24). The inventory will always peak after the input flow rate has reached its maximum value and will become zero when the flows which have accumulated in the capacity

constrained interval (s_1, s_2) are completely processed by the maximum processing rates, i.e., when

$$\Lambda(s_2) - \Lambda(s_1) = \frac{k}{2}(s_2 - s_1) \qquad (2.25)$$

This expression and the solution for s_2 can also be obtained by setting the delay of a letter in the first stage, (Equation (2.5) and Figure 4)

$$\tau_1(t) = \frac{2}{k}(\Lambda(t) - \Lambda(s_1)) - (t - s_1) \qquad t \in S \qquad (2.26)$$

equal to zero at the end-point s_2. Again it is interesting to point out that the graphical solution of s_2 in Figure 6 is simply the intersection of the solid tangent line $V_1(t)$ with the cumulative input $\Lambda(t)$.† Whereas the optimal processing rates are continuous at the corner-point s_1 they are discontinuous at s_2,

$$v^*(s_2^-) = \frac{k}{2} \; ; \qquad v^*(s_2^+) = \lambda(t) \; .$$

With the optimal assignments of Equations (2.15) and (2.22) 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 (2.26) for $t \in S$ and zero for $t \in \bar{S}$. Using this optimal assignment of processing rates the minimum delay of all letters becomes

$$D^* = \int_S \lambda(t)\tau(t)dt = \int_{s_1}^{s_2} \lambda(t)\,\tau_1(t)\,dt \qquad (2.27)$$

D^* can be expressed in terms of the corner-points s_1 and s_2 when one substitutes Equations (2.24), (2.25) and (2.26) into (2.27).

† In the calculus of variations, Equations (2.24), (2.25) and (2.26) are known as the transversality or variable end point conditions.

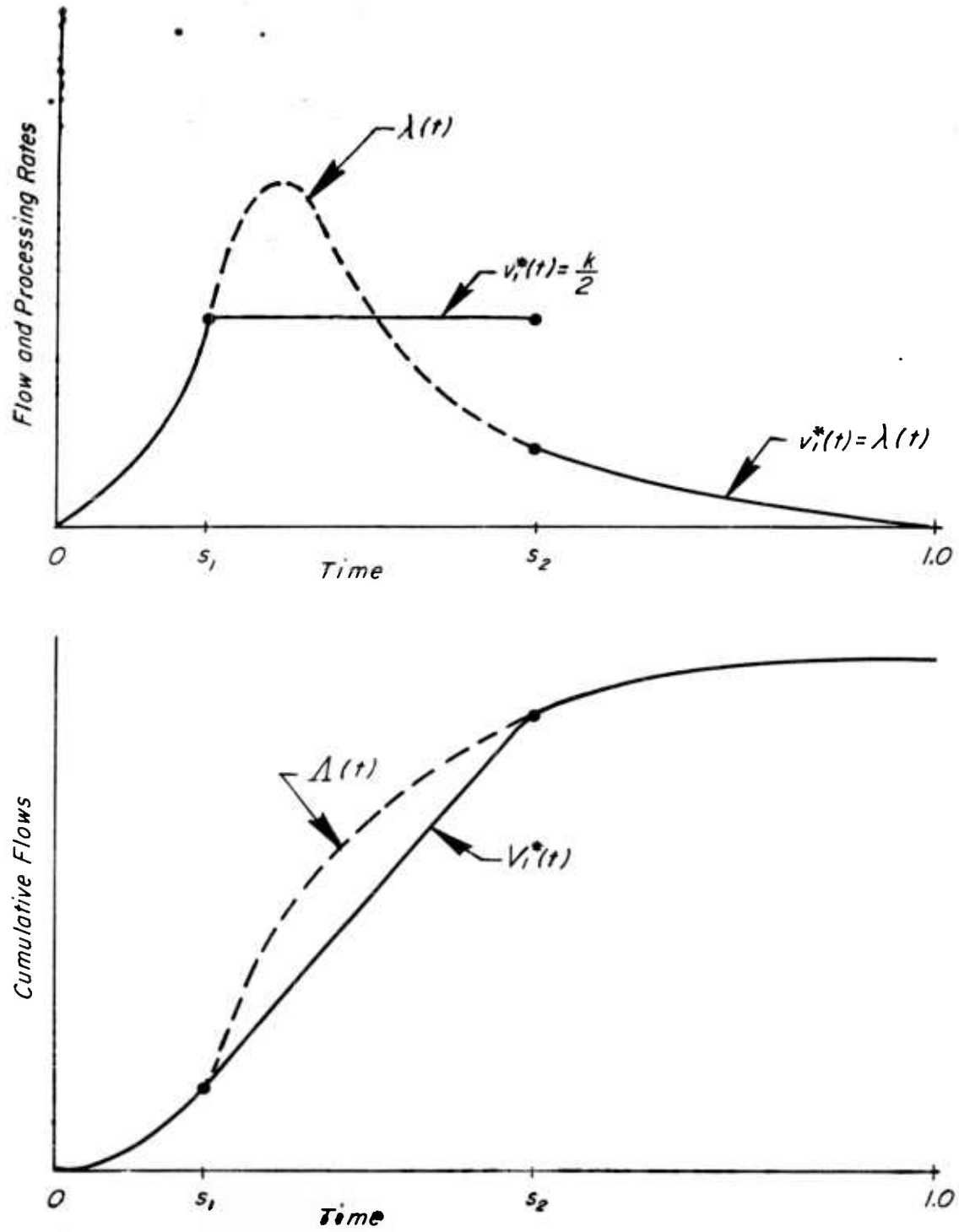


Fig. 6-Optimal processing rates as a function of time.

$$D^* = \frac{k}{2}(s_2 - s_1)^2 - \bar{\lambda}(s_2)(s_2 - s_1) + \int_{s_1}^{s_2} \lambda(t)\,dt$$

$$= \frac{k}{2}(s_2^2 - s_1^2) - \int_{s_1}^{s_2} t\lambda(t)\,dt \qquad (2.28)$$

We have already calculated the effect of varying the processing rates from their optimal values. Clearly the only feasible variation, which increases the average delay of letter mail, is one which increases the processing rate at stage 1, forces the processing rate at stage 2 to be less than the output rate of stage 1 and hence allows inventories to build up at stage 2 as well as at stage 1. While this variation no longer preserves optimality, it is an important situation to study since, in practice, it is hardly ever possible to exactly match processing rates with one another or with the input rates of letter mail.

It is also possible to calculate the effect of variations in the capacity processing rate k, available to the two-stage system. The effects of varying k by small amounts are twofold: first of all, the optimal processing rates within the region S are changed; secondly, the length of the capacity-constrained interval, i.e., the location of the beginning and end-points s_1 and s_2 are also changed.

The effect of these two types of variations are important, since the former give an estimate of increases in average delay at constant cost due to non-optimal assignments, while the latter give estimates of changes in average letter delay (using optimal assignments) due to small changes in cost.

N Serial Stages with Different Processing Efficiencies

Let us now consider delays through N serial stages where the inequality restriction of Equation (2.8) is generalized to include the interchange of resources between N stages.

$$\sum_{i=1}^{N} a_i v_i(t) \leq k \qquad a_i \geq 0 \qquad (2.29)$$

In many instances the interchange of one man or machine between stages does not result in the same processing rate. This effect is especially pronounced when an interchange of skilled and non-skilled labor is made between processing stages.

It is fortunate that one can write the expression for average delay in terms of the inventory level at time t. By arguments which are similar to those posed in Equation (2.2) and (2.4) the delay of a letter entering the queue of the first stage at time t and leaving the N^{th} stage at time $t + \tau(t)$ is the solution of the integral equation,

$$\lambda(t) = V_N(t + \tau(t)) \; . \qquad (2.30)$$

The non-negative inventory restrictions of Equation (2.7b) now become

$$V_i(t) \geq V_{i+1}(t) \qquad (2.31)$$

where $V_i(t)$ is the cumulative flow processed by the i^{th} stage at time t. The average delay can be written in terms of the unprocessed inventories by a change of variables from t to $\lambda(t)$,

$$D = \int_S \lambda(t)\tau(t)\,dt = \int_S \tau(t)\,d\lambda(t) = \int_S [\lambda(t) - V_N(t)]\,dt \qquad (2.32)$$

The variation in D is

$$\delta D = - \int_S \delta V_N(t)\, dt = - \int_S \int_0^t v_N(s)\, ds\, dt \qquad (2.33)$$

and it is clear that D decreases with increases in $V_N(t)$. To obtain an optimal schedule we do precisely this until either the non-negative inventory restrictions of Equation (2.31) or the capacity sorting rate restrictions of Equation (2.29) are violated. If all of the former inequalities are violated first then we are in a region $\bar{S}$ where $\tau(t) = 0$; otherwise, in a region S where $\tau(t) > 0$. In S, Equation (2.29) becomes a strict equality. In this case, the variation in the cumulative flow processed by stage N is expressible in terms of the cumulative flows processed by all other stages.

$$- \delta V_N(t) = \sum_{i=1}^{N-1} a_N^{-1}\, a_i \delta V_i(t) \qquad (2.34)$$

Since all $a_i \geq 0$, positive variations in $V_N(t)$ can be achieved by negative variations in one or more of the cumulative flows through the remaining processing stages until finally

$$\lambda(t) > V_i^*(t) = V_j^*(t) \qquad t \in S \qquad (2.35)$$

Again, the argument follows that since the cumulative amounts processed are equal for all t in S, the derivatives are also equal,

$$\begin{aligned} v_i^*(t) &= v_j^*(t) \qquad 1 \leq i,\ j \leq N \\ &= k\left[\sum_i a_i\right]^{-1} \qquad t \in S \end{aligned} \qquad (2.36)$$

Hence in an optimal program the delays to letters are concentrated at the first stage.

The optimal processing rates at each stage are simple to compute. In a region S they are given by Equation (2.36) whereas in a region $\bar{S}$, $\tau(t) = 0$ and the capacity sorting rate of Equation (2.29) is not restrictive:

$$v_i^*(t) = \lambda(t) \qquad t \in \bar{S}$$

$$\lambda(t) < k\left[\sum_i a_i\right]^{-1} \tag{2.37}$$

The only problem which remains is that of finding where the intervals S and $\bar{S}$ begin and terminate. When $\lambda(t)$ is a function with several peaks there may be several switches from S to $\bar{S}$. But, in the case where there is only one S, the beginning point, s_1, of the capacity constrained interval, is the smallest root of

$$\lambda(s_1) - k\left\{\sum_i a_i\right\}^{-1} = 0 \tag{2.38a}$$

and the end-point, s_2, is the solution of

$$\Lambda(s_2) - \Lambda(s_1) = k\left\{\sum a_i\right\}^{-1}(s_2 - s_1) \tag{2.38b}$$

When $\lambda(t)$ consists of multiple peaks, there is no difficulty in finding the solutions of the optimal processing rates. Intervals of S and $\bar{S}$ alternate. In the case where inventories have accumulated as a result of say the first peak in $\lambda(t)$, it may happen that the first capacity constrained interval S extends beyond the second, third, etc. peak in the input flow rates. However if processing rates are large enough, the inventories arising from one peak in the input rate are reduced to zero before the arrival of a second peak. The solutions of the corner points and the average minimum delay are in all cases simple generalizations of our earlier equations.

Initial Inventories

So far we have been concerned with the simple physical situation where cumulative flows are zero at time zero. Since initial inventories may be non-zero, one asks the question, "will these initial inventories affect the scheduling policies obtained in the previous section?"

Let us return, momentarily, to the two-stage system of Figure 2. If I_1 and I_2 are the initial inventories at stages 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 (2.3) and (2.4). Again, the variation of the delay of a letter depends only on variations in sorting rates at stage 2 and although D is obviously increased by the presence of initial inventories at either stage, the variation of a letter entering at t remains explicitly independent of I_1 and I_2.

But the important new feature is that one due to the inequality restrictions on non-negative inventories of Equation (2.7b). The effect of I_1 and I_2 is to replace (2.7b) by

$$I_1 + \Lambda(t) \geq V_1(t) \geq V_2(t) - I_2 \tag{2.39}$$

In an interval $\bar{S}$ where $\tau(t) = 0$, and the strict equalities of Equation (2.39) hold, we again find that $v_1^*(t) = v_2^*(t) = \lambda(t)$ and negative variations in processing rates at stage 2 can only increase average delay.

In an interval S where $\tau(t)$ is greater than zero in the optimal schedule, it is no longer true that one can arbitrarily force cumulative amounts processed at stage 1 to equal cumulative amounts processed at stage 2. Nevertheless, Equations (2.10) and (2.13) tell us that D always decreases by increasing the cumulative amounts processed by stage 2. It now appears that the capacity constrained interval S can consist of an initial sub-interval where

processing rates are not split between stage 1 and 2. This interval will correspond to a period in time when inventories of mail in stage 2 are positive. Once inventories in stage 2 are reduced to zero, the optimal solutions in earlier sections of this paper apply. To illustrate the optimal schedules in more detail, let us examine two extreme cases, drawn in Figure 7a, b and labelled as Case I and Case II. In Case I the single peak in the input rate, $\lambda(t)$, comes early and inventories at stage 1 and 2 are large. In Case II, inventories are relatively small and the peak in $\lambda(t)$ comes late in the period.

In the former case, the solutions for the optimal processing rates at both stages will differ only slightly from the solutions obtained in Equation (2.22). Since D always decreases with positive variations in $V_2(t)$ whenever $V_2(t) < V_1(t) + I_2$ we start at time zero by assigning the maximum processing rate k at stage 2. We continue with this assignment until the inventory in stage 2 is completely processed and then split the capacity processing rates between stage 1 and 2 until inventories in stage 1 are also completely processed. The optimal processing rates are

$$v_1^*(t) = 0; \; v_2^*(t) = k \qquad 0 \leq t \leq s_1 \qquad (2.40a)$$

$$= \frac{k}{2}; \quad = \frac{k}{2} \qquad s_1 < t \leq s_2 \qquad (2.40b)$$

$$= \lambda(t); \quad = \lambda(t) \qquad s_2 < t \leq 1 \qquad (2.40c)$$

where $s_1 = I_2/k$ and the solution of s_2 is obtained by substituting $I_1 + \Lambda(s_2)$ for the left-hand side of Equation (2.25).

In Case II where the peak in $\lambda(t)$ is late in the period, one is able to reduce inventories in both stages to zero before capacity processing rates would normally be assigned to handle the peak flow rates. Earlier arguments

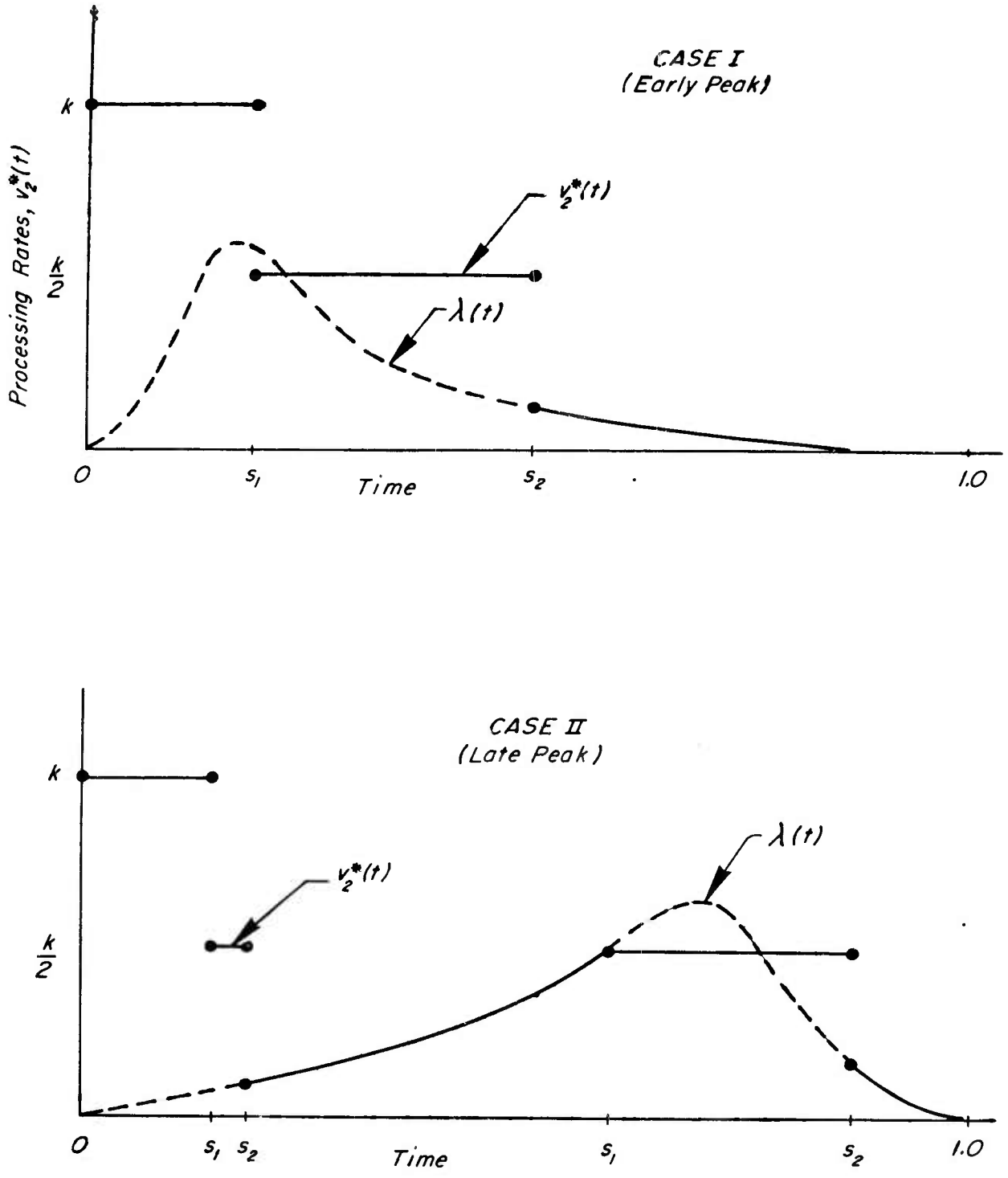


Fig. 7 – Processing initial inventories.

again apply. The assignment in $(0, s_1)$ is the capacity processing rate at stage 2 and in (s_1, s_2) is one-half the capacity processing rate at both stage 1 and stage 2. At s_2 inventories have been completely processed in both stages and processing rates are reduced to the input rate, $\lambda(t)$ until flow rates become so large that capacity processing rates are again assigned. We will denote this second capacity-constrained interval by $S' = (s'_1, s'_2)$; the solutions of the end-points are obtained by substituting s'_1 and s'_2 for s_1 and s_2 in Equations (2.24) and (2.25). We come to the immediate conclusion that once inventories in stage 2 are ever reduced to zero they do not appear again in an optimal schedule. The optimal assignments are

$$v_1^*(t) = 0 \qquad v_2^*(t) = k \qquad 0 \leq t \leq s_1 \tag{2.41a}$$

$$= \frac{k}{2} \qquad = \frac{k}{2} \qquad s_1 < t \leq s_2 \tag{2.41b}$$

$$= \lambda(t) \qquad = \lambda(t) \qquad s_2 < t \leq s'_1 \tag{2.41c}$$

$$= \frac{k}{2} \qquad = \frac{k}{2} \qquad s'_1 < t \leq s'_2 \tag{2.41d}$$

$$= \lambda(t) \qquad = \lambda(t) \qquad s'_1 < t \leq 1 \tag{2.41e}$$

The solutions for the corner points s_1 and s_2 are always obtained from the expressions which equate cumulative input flows to cumulative output flows,

$$s_1 = I_2/k$$

$$\frac{k}{2} s_2 - \lambda(s_2) = I_1 + I_2/2 \tag{2.42}$$

The solutions for s'_1 and s'_2 are obtained by substituting primed for unprimed variables in Equations (2.24) and (2.25). While solutions will always be

obtained for s_1' and s_2' the decision to actually allocate capacity processing rates in S' depends on whether or not

$$s_2 = \text{Min}(s_1', s_2) \quad .$$

If this statement holds, initial inventories in both stages will be reduced to zero before capacity processing rates are normally assigned. Otherwise, the inventories will be reduced to zero after the peak mail flows have occurred.

As one would expect, the N stage case is, in principle, no different. The optimal assignments can be stated as follows:

> "Start at the last, N^{th}, stage and assign capacity processing rates to that stage until all inventory is processed. When the inventory in the N^{th} stage becomes zero split the capacity processing rates such that flow out of the (N-1)st stage equals flow out of the N^{th} stage; i.e., maintain zero inventory at the N^{th} stage. Continue in this fashion until inventory at stage N-1 is completely processed; split the capacity processing rates between the N^{th}, (N-1)st and (N-2)nd stage etc..... until inventories at all stages are reduced to zero. Once inventories at stages 2, 3, ... (N) have been reduced to zero the optimal assignments are,

$$v_1^* = \lambda(t) \qquad t \in \bar{S} \qquad (2.43)$$

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

That is to say, inventories are only allowed to appear at the first stage.

Sorting and Branching

In its journey through the post office mail is processed through a number of serial stages. It is generally characteristic of a post office that the large bulk of sorting operations are made towards the end of the journey while only a few simple sorts are made in the early stages. Physically and conceptually a sorting or classification operation is one which differs from the usual processing operations of dumping, weighing, cancelling, etc. in that (i) the letter or package is observed piece by piece, (ii) a routing decision is made, which (iii) results in the choice of one out of many flow routes from a large network of successive sorting and processing operations.

Depending on whether the person making the decision at a sorting stage is or is not aided by automatic equipment, the number of distinct flow routes which branch from a single sorting stage may vary from two to several thousand. The sorting stages towards the end of the flow process also include certain special features of storage and inter-stage transportation; consequently their mathematical analysis is deferred to Section 3.

On the other hand the simpler (say binary and ternary) sorting stages are properly included in this section of the report since the predominant delay of a letter passing through these stages is a delay in queue, these queues again arising from temporary peaks in mail flow rates and capacity restrictions on sorting rates. The binary sorting operations are important since they either handle large volumes of mail flow or separate high-priority mail such as airmail and special delivery from the main flow stream. Letters which do not carry sufficient postage and oversized envelopes which require special handling are also good examples of flows which branch from the main stream. Again we proceed by considering the simplest sorting problem in the two-branch case of Figure 8. We assume that the input rate is $\lambda(t)$; in actual fact

this input rate may be the output flows from earlier stages whose operation is independent of the stages we are now studying. The sorting rate is $v_o(t)$ 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.

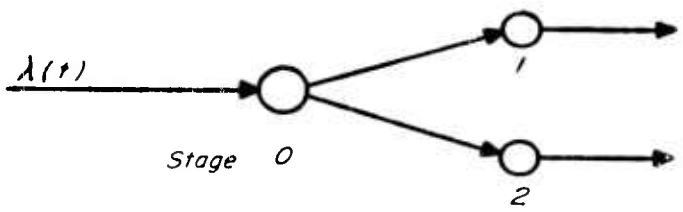


Fig. 8 – A sorting stage with branching.

In order to minimize the average delay of all letters which flow through the three-stage system of Figure 8 we must separately calculate the delay of a letter which goes through stage 0 and stage 1 as well as the delay of a letter which goes through the lower branch. In this section we will denote the former by $\tau_\alpha(t)$ and the latter by $\tau_{1-\alpha}(t)$. The delay of any one letter through, say, the top branch is made up of the sum of the delay in queue in stage 0 plus the delay time of a letter in the queue of stage 1 when the arrival time in this last queue is the exit time from stage 0.

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

$$D = \int_0^1 \Big(\alpha\tau_\alpha(t) + (1-\alpha)\,\tau_{1-\alpha}(t)\Big)\,\lambda(t)\,dt \qquad (2.44)$$

The non-negative inventory restrictions are, as before, statements that cumulative amounts processed at early stages must always be greater than or equal to cumulative amounts processed at a later stage. Since a

fraction α and $1 - \alpha$ flow into the two branches, these non-negative restrictions become,

$$\Lambda(t) \geq V_0(t) \tag{2.45a}$$

$$V_0(t) \geq V_1(t) \tag{2.45b}$$

$$(1-\alpha)V_0(t) \geq V_2(t) \quad . \tag{2.45c}$$

Basically two additional types of restrictions are encountered in postal operations. Either the sum of processing rates at stages 1 and 2 are restricted and $v_0(t)$ is not restricted,

$$v_1(t) + av_2(t) \leq k \tag{2.46a}$$

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

$$v_0(t) + a_1 v_1(t) + a_2 v_2(t) \leq k \quad . \tag{2.46b}$$

We have already pointed out that the constants a, a_1 and a_2 may not be equal to unity if labor or machines are not equally efficient at all of the three stages. For example, stage 0 represents a sorting stage where a certain amount of skill is called for in making the classification decision. The sorting rate of a man is usually reduced since he must consider each letter individually. In stages 1 and 2 which might be weighing stages, the mail is handled in bulk, the processing rate per man per unit time may be orders of magnitude larger than in the case of stage 0 and may call for relatively unskilled labor. Yet there are times when it will be advisable to shift labor between any one of the three stages mentioned above. The proof of the

optimal assignments of processing and sorting rates proceeds as before, but in the remaining parts of this paper we emphasize the solutions and certain characteristics of letter delays which result from these optimal assignments.

Restrictions on Parallel Processing Stages

We now turn our attention to policies which minimize Equation (2.44) subject to the non-negative inventory restrictions of Equation (2.45) and the restriction of Equation (2.46a) on processing rates. Again, we consider two types of intervals S and $\bar{S}$, where letter delays are non-zero and zero respectively. In $\bar{S}$, the sorting rate at stage 0 is just equal to the arrival rate, $\lambda(t)$, and the processing rates at stage 1 and 2 are $\alpha\lambda(t)$ and $(1-\alpha)\lambda(t)$ respectively. The region $\bar{S}$ terminates for the first time when the sum of processing rates equals the capacity[+]

$$\lambda(t) - k(\alpha - a\alpha + a)^{-1} = 0 \quad . \tag{2.47}$$

Since letter delays are zero in the interval $\bar{S}$, the average letter delay through the three stage network of Figure 8 and Equation (2.44) becomes

$$D = \int_S \Big(\alpha\lambda(t) - V_1(t) \Big) dt + \int_S \Big((1-\alpha)\lambda(t) - V_2(t) \Big) dt \quad . \tag{2.48}$$

Substituting Equation (2.46a) we find that the variation in average delay within the interval S for small variations in the processing rate of stage 1 is

$$\delta D = \int_S \left(\frac{1}{a} - 1 \right) \delta V_1(t)\, dt \quad . \tag{2.49}$$

Since the first variation in the variable end points of S in Equation (2.48) is simply the requirement that inventories (or delays) be zero at these points, we can now replace the variable end point problem of Equation (2.48) by a fixed end point problem. To see this clearly we momentarily step aside from the optimization problem.

[+] Again we assume that $\lambda(t)$ starts from zero, has a single peak and returns to zero before the end of the interval (0, 1). Of course we also assume that $\lambda(1) \leq k$, i.e., that all flows can eventually be processed in the interval.

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 processed must satisfy Equation (2.46a) and its integral

$$\Lambda(s_2) - \Lambda(s_1) = k(s_2 - s_1) + (1 - \tfrac{1}{a})\, V_2(s_2) - V_2(s_1) \quad . \qquad (2.50a)$$

As we have just mentioned, the first variations at the variable end-points require that inventories be equal to zero. Since inventories at each stage must be non-negative this statement is equivalent to the statement that inventories at stage 1 and 2 be zero independently of one another; hence,

$$V_2(s_1) = (1 - \alpha)\, \Lambda(s_1); \quad V_2(s_2) = (1 - \alpha)\Lambda(s_2) \qquad (2.50b,c)$$

and by substitution of this equation into (2.50a) we get a solution for s_2

$$\Lambda(s_2) - \Lambda(s_1) = k(s_2 - s_1)\,(a - a\alpha + \alpha)^{-1} \qquad (2.50d)$$

which is independent of the actual assignment of processing rates at stage 1 or 2 and is only a function of s_1, α, a, k and the shape of the input flows $\Lambda(t)$. But s_1 is just the smaller root of Equation (2.47) and we therefore find that s_2 is only a function of a, α, k and $\Lambda(t)$.

From the large set of schedules which are feasible in the sense of satisfying Equation (2.45) one need only consider a smaller set as candidates for an optimal scheduling policy. These are the schedules which make $s_2 - s_1$ as small as possible by assigning capacity processing rates and reducing inventories to zero as quickly as possible. However 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.

If a is greater than 1 in Equation (2.49), $\frac{1}{a} - 1 < 0$ and 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 (i) Equation (2.45b) becomes a strict equality or (ii) 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 (2.45b) will be the inequality which 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 $v_2^*(t) = \frac{k - \alpha\lambda(t)}{a}$. The optimal processing rate $v_2^*(t)$, is plotted in Figure 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 rather curious behavior in that the interval S will contain a sub-interval R during which time the processing rate at stage 2 is zero and the delays in the top branch are positive. See Figure 10 a, b. Figure 9 corresponds to Case A where R is empty. In analogy in the end-point solutions of s_1 and s_2 we now want to find solutions for the end-points of R, namely r_1 and r_2. For the case where R is not empty, Case B, r_1 is the first time when the flow rate into stage 1 equals the capacity processing rate, i.e., the smaller real root of

$$\alpha\lambda(t) - k = 0 \qquad (2.51)$$

† $a > 1$ corresponds to the case where stage 1 is more "efficient" than stage 2, i.e., the same allocation of men or machines results in a higher processing rate at stage 1.

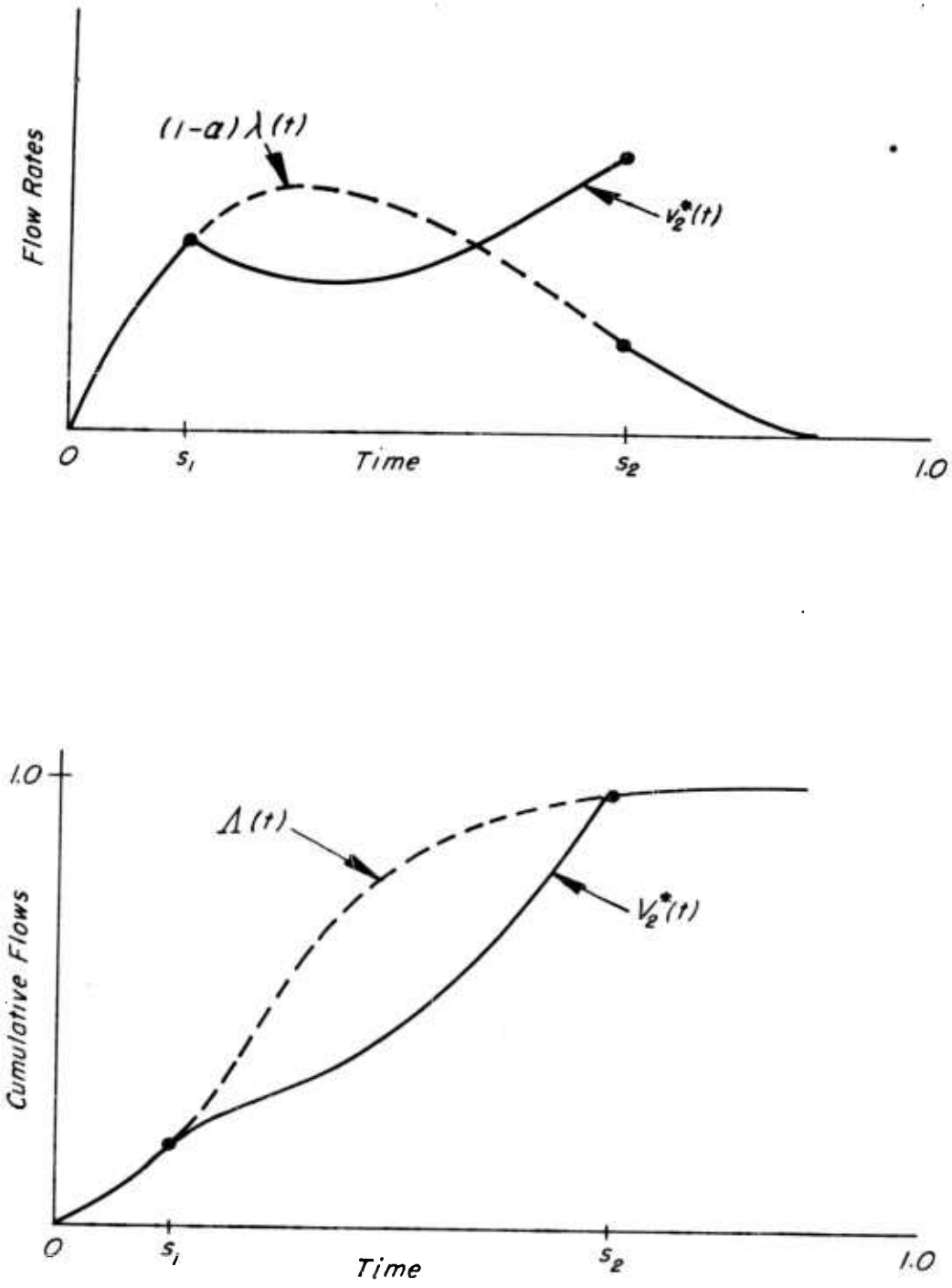
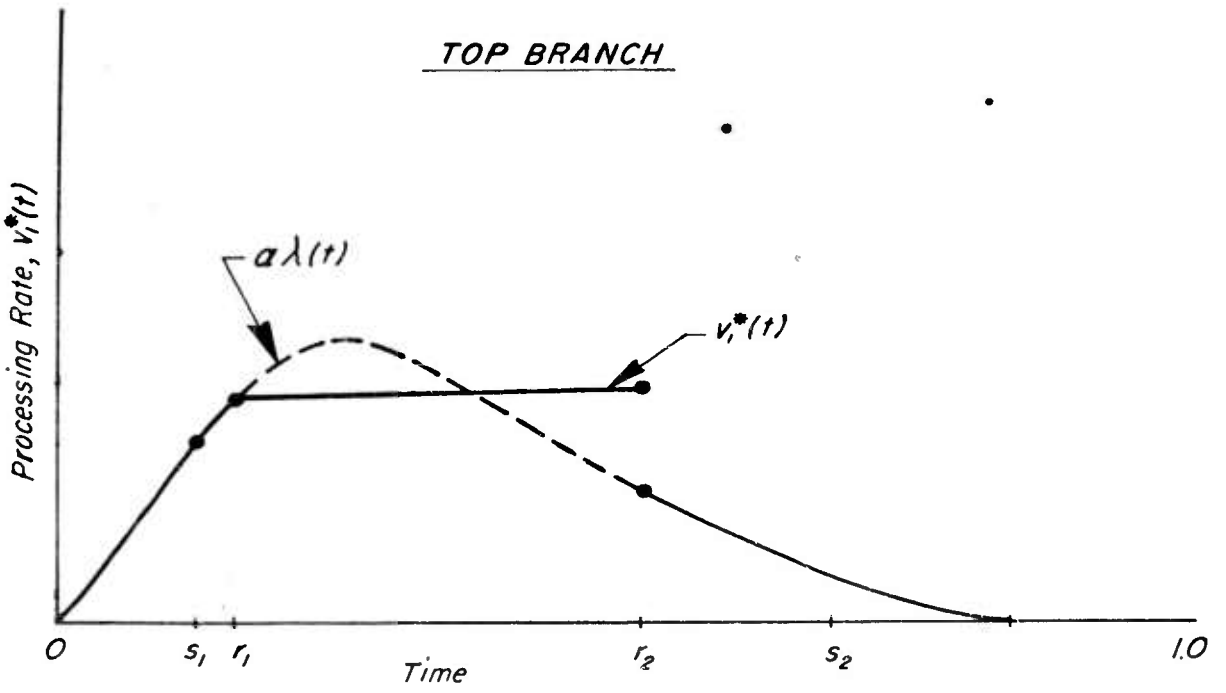


Fig. 9 – Optimal processing rates, case A.



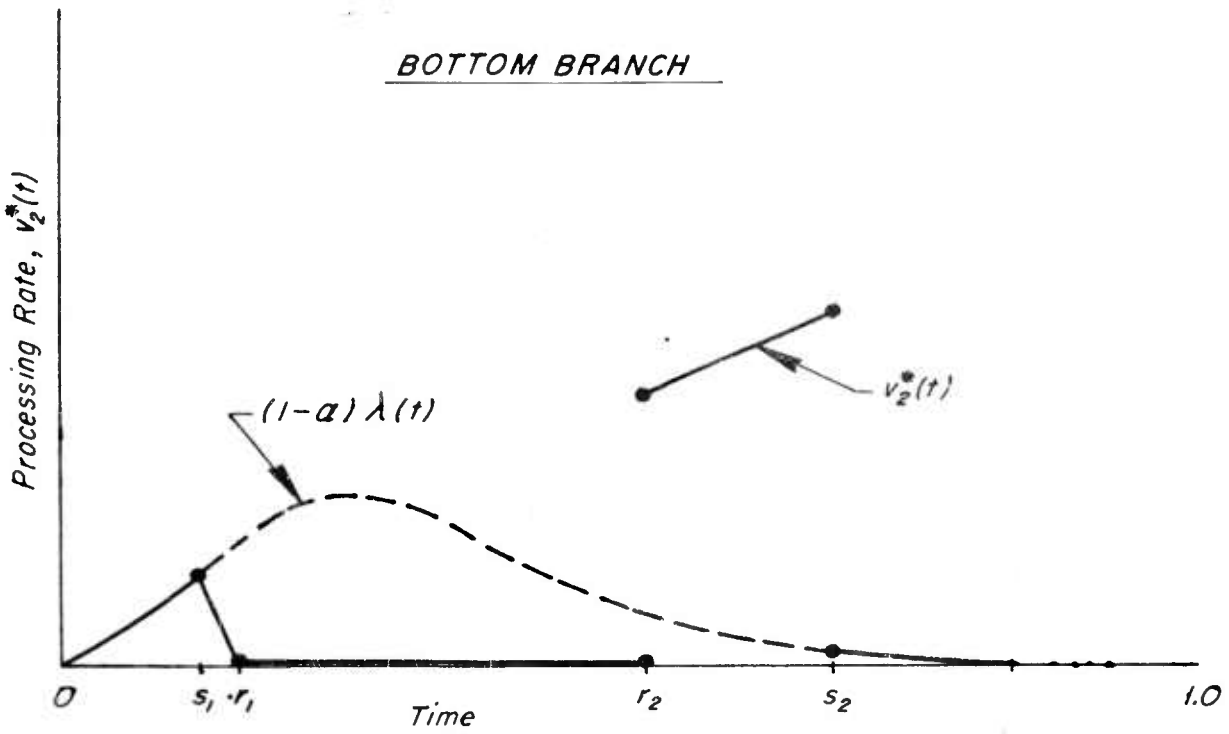


Fig. 10 - Optimal processing rates, case B.

The solution for r_2 is then obtained by substituting $\frac{k}{\alpha}$ for $\frac{k}{2}$ in Equation (2.25) and r_1 and r_2 for s_1 and s_2 respectively.

The optimal processing rates are summarized in Table 1 below for $a > 1$. For $a < 1$ the role of the top and bottom branch are reversed and for $a = 1$, the average delay is indifferent to the assignment of feasible processing rates so long as capacity rates are used.

TABLE 1

Optimal Processing Rate	Time Interval	Restrictions on capacity processing rates
Stage 0 $v_0^*(t) = \lambda(t)$	$0 \leq t \leq 1$	
Stage 1 $v_1^*(t) = \alpha\lambda(t)$	$0 \leq t \leq 1$	$k \geq \alpha \operatorname{Max}_t \lambda(t)$
Stage 2 $v_2^*(t) = (1-\alpha)\lambda(t)$	$s_2 \leq t < s_1$	$k < (a+\alpha-a\alpha)\operatorname{Max}_t \lambda(t)$
$= \frac{k - \alpha\lambda(t)}{a}$	$s_1 \leq t < s_2$	(CASE A)
Stage 1 $v_1^*(t) = \alpha\lambda(t)$	$r_2 \leq t < r_1$	$k < \alpha \operatorname{Max}_t \lambda(t)$
$= k$	$r_1 \leq t < r_2$	$k > a + \alpha - a\alpha$
Stage 2 $v_2^*(t) = (1-\alpha)\lambda(t)$	$s_2 \leq t < s_1$	(CASE B)
$= \frac{k - \alpha\lambda(t)}{a}$	$s_1 \leq t < r_1;\ r_2 \leq t < s_1$	
$= 0$	$r_1 \leq t < r_2$	

Expressions for average letter delay are easy to obtain. When no letters are delayed in the top branch, i.e., R is empty, the minimum average delay is,

$$D^* = \left(\frac{a - a\alpha + \alpha}{a}\right)\left\{\int_{s_1}^{s_2} \Lambda(t)dt - (s_2 - s_1)\,\Lambda(s_1)\right\} - \frac{k}{2a}(s_2 - s_1)^2. \quad (2.52a)$$

Equation (2.52) approaches (2.28) as a and α approach 1 and a processing rate of k/2 is used. When delays appear in both branches we calculate the average letter delay in each branch and add results. The average letter delay in the top branch is found by substituting k for $\frac{k}{2}$ in Equation (2.28), $R = (r_1, r_2)$ for $S = (s_1, s_2)$ and $\alpha\Lambda(t)$ for $\alpha\Lambda(t)$. The average letter delay in the bottom branch is the sum of two parts, one being the average delay when $v_2(t) = 0$, the other being outside the interval R but inside S. In the latter case we get terms which are similar to Equation (2.52a) except that the end points (s_1, s_2) are replaced by (s_1, r_1) and (r_2, s_2). Adding the minimum average letter delays in both branches we obtain

$$D^* = \gamma\int_{s_1}^{s_2} \Lambda(t)dt - \frac{\alpha}{a}\int_{r_1}^{r_2} \Lambda(t)dt - \frac{k}{a}\left((r_1 - s_1)(s_2 - r_1) + \frac{1}{2}(s_1 + s_2 - r_1 - r_2)^2\right) \quad (2.52b)^{\dagger}$$

$$- \gamma\,\Lambda(s_1)(r_2 - s_1) + \Lambda(r_1)\left((a^{-1}\alpha)(r_2 - r_1) - \gamma(s_2 - r_2)\right) .$$

for Case B.

The extension of these results to the many-branch case is simple when flow rates from the sorting stage exceed the processing capacity of N parallel stages. For the case of different efficiencies at each stage, rank the stages by their efficiency. If possible, keep inventories at the most efficient stage equal to zero by matching input rates with processing rates. The remainder of the processing capacity, if any, should be assigned to the next most efficient stage and so on to the least efficient stage. In general there will be zero letter delay in $m (1 \leq m \leq n)$ of n total branches.

† In (2.52b), $\gamma = (a + \alpha - a\alpha)a^{-1}$.

Cleaning Up Initial Inventories

A simple but important case arises when the input rate to each branch of Figure 8 is zero but initial inventories at stage 1 and 2 are I_1 and I_2 at time zero. In this case $\lambda(t)$ can be interpreted as a delta function with area $I_1 + I_2$ at the origin and $\alpha = I_1/(I_1 + I_2)$. With methods of the calculus we have no difficulty in finding the optimal assignment which minimizes average letter delay when processing rates are restricted by Equation (2.46a). Naturally, we assign capacity processing rates over an interval $S = (0, s)$. Both inventories will not be reduced to zero at the same time, but one inventory, say at stage 1, will be reduced to zero at a time $r < s$; capacity processing rates can then be reassigned to the remaining inventory at stage 2 in the interval (r, s).

Assign a constant processing rate v_1 at stage 1. At a time $r = \frac{I_1}{v_1}$ the inventory in front of this stage will be reduced to zero. During this time an amount $rv_2 = \frac{r}{a}(k-v_1)$ will be processed in stage 2 leaving an unprocessed inventory of size

$$I_2(r) = I_2 + \frac{I_1}{a}\left(1 - \frac{k}{v_1}\right)$$

to be processed by the new processing rate $v_2 = \frac{k}{a}$ in the time period $(s-r)$. Equating total amounts processed in the remainder of the interval with total inventory unprocessed, we find that the time of completion of the process is inversely proportional to k, namely

$$s = \frac{I_1 + aI_2}{k} \qquad (2.53)$$

We note that $s > r$ when $v_1 > \frac{kI_1}{I_1 + aI_2}$.

The average delay through the top branch of Figure 8 is quadratic in the initial inventory

$$D_\alpha = \frac{1}{2} r I_1 = \frac{I_1^2}{2v_1} \qquad (2.54a)$$

while the average delay through the bottom branch,

$$D_{1-\alpha} = \frac{1}{2}\left((r + s)I_2 - rsv_2\right) \qquad s > r \qquad (2.54b)$$

contains linear and quadratic terms in I_1. On substituting $\frac{k-v_1}{a}$ for v_2 in this equation and adding to D_α we get

$$D = \frac{I_1 + aI_2}{2ak} - \frac{1 - a}{a}\left(\frac{I_1^2}{2v_1}\right) \qquad (2.54c)$$

which is a constant plus a term inversely proportional to v_1. If we consider the case where $a > 1$, i.e., where stage 1 is the "efficient" processing stage, the coefficient of v_1 is negative and the average system delay can always be reduced by increasing the processing rates at stage 1. The optimal processing rates are therefore,

$$\begin{aligned} v_1^* &= k && 0 \leq t \leq r \qquad (2.55)\\ &= 0 && r < t \\ v_2^* &= 0 && 0 \leq t \leq r \\ &= \frac{k}{a} && 0 < t \end{aligned}$$

The optimal policy is identical to one which reduces inventories to zero in minimum time.

If we consider N rather than 2 parallel stages and linear restrictions on the sum of processing rates, we obtain the simple result that processing capacity should first be assigned to the most "efficient" stage. When inventories are completely processed at that stage, processing rates should be reassigned to the next most efficient stage and so on until the least efficient stage is processed last. This assignment is not optimal when one considers time-varying input rates in addition to initial inventories simply because the non-negative inventory restrictions of Equation (2.45) also play an important problem in the decision part.

More complicated flow patterns than the ones we have discussed arise in postal sorting operations. For example, a sorting operation may divide flow into N branches, and fractions of the flow through one branch may feed back into an earlier processing stage. Another important flow pattern arises when one branch completely bypasses a group of processing stages. While the former pattern is often used to re-route mis-sent mail, the latter is often used to expedite high-priority mails, say special delivery, around the normally high-volume and long-delay processing stages.

While the arrangement of processing stages or the restrictions on capacity processing rates may vary from one post office to another, the mathematical models and the optimal decision rules are not sufficiently different from the ones we have studied to warrant further discussion at this time.

3. 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 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 district 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 Secondary 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 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 which, 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 which 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 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 which 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 which 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 which accurately describe the effect of storage on letter delays.

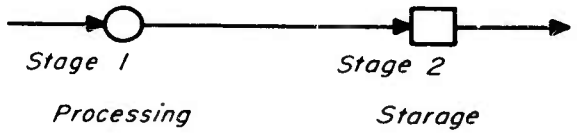


Fig. II-A serial processing and storage stage.

In Figure 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. One of many important questions 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 which leaves the system at dispatch time is just equal to the cumulative flow which 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 insures eventual dispatch of all mail. The total delay of a letter is again made up of two parts, the delays in stage 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 which enters the storage stage at time $t + \tau_1(t)$. Although the total delay is still given by the expression,

$$\tau(t) = \tau_1(t) + \tau_2\left(t + \tau_1(t)\right) \tag{3.1}$$

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.

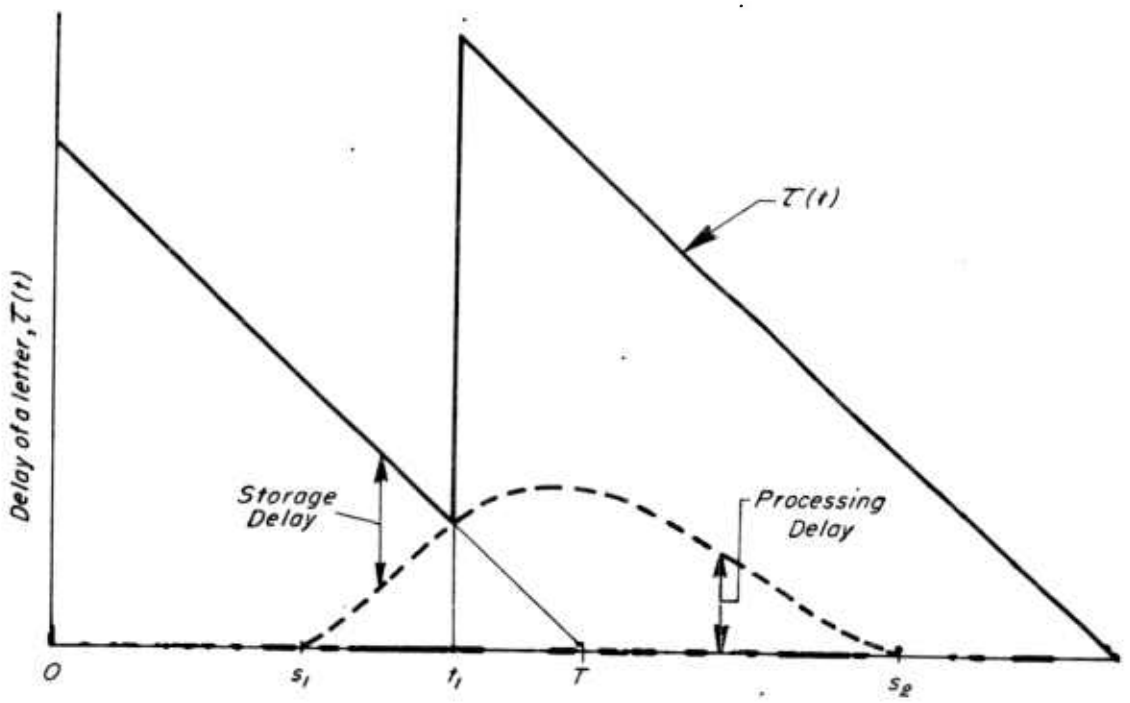


Fig. 12 - Letter delay with an intermediate dispatch at T

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 which 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 Figure 12 as the difference between the saw-tooth and the time axis. As one examines delays of those letters which arrive later in time, one arrives at the region 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 wait in queue at stage 1; any letter which arrives later does not make the intermediate dispatch at time T. Hence, t_1 is simply the solution of Equation (3.1) which adds the arrival time of the letter to its delay at stage 1 and equates this sum to the intermediate dispatch time,

$$\tau(t_1) = t_1 + \tau_1(t_1) = T \qquad (3.2)$$

With capacity restrictions on processing rates at stage 1, we have already obtained explicit solutions for $\tau_1(t)$, namely,

$$\tau_1(t) = \frac{\Lambda(t) - \Lambda(s_1)}{k} - (t - s_1) \qquad t \in S \qquad (3.3)$$

where k is the constant value for $v_1(t)$ in S.[4] It is a simple matter to substitute Equation (3.3) into the expression for $\tau_1(t_1)$ in Equation (3.2) to find that t_1 can also be expressed in terms of the corner-point s_1, and the dispatch time T,

$$\Lambda(t_1) = k(T - s_1) + \Lambda(s_1) \qquad (3.4)$$

Equation (3.4) states that flows which 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, (Figure 12), this statement is accurate so long as the slope of $\tau_1(t)$ is always greater than the slope of the trailing edge of the saw-tooth, namely -1. For capacity constrained sorting or processing rates at stage 1, the slope of $\tau_1(t)$ is found from Equation (3.3),

$$\tau_1'(t) = k^{-1}\lambda(t) - 1 > -1 \qquad t \in S \qquad (3.5)$$

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

---

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

Average Delays

To understand the over-all effects of a storage stage upon the flow of mail through a number of serially connected 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,

$$D = \int_0^1 \lambda(t)\,\tau(t)\,dt = 1 - \bar{t} - (1-T)\,\Lambda(t_1) \qquad T \in S \qquad (3.6a)$$

$$= 1 - \bar{t} - (1-T)\,\Lambda(T) \qquad T \in \bar{S} \qquad (3.6b)$$

where $\bar{t} = \int_0^1 t\lambda(t)\,dt$. The first terms in (3.6b) depend on the shape of the input curve and the last term is a function of the intermediate dispatch time T.

If we consider the case where T lies in S, the peak of the second sawtooth in Figure 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 (3.6b).

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 (3.4) into Equation (3.6a). The processing rate is constant in the capacity constrained region S, and the average delay is a quadratic function of the dispatch time, T,

$$D = kT^2 + \left(\Lambda(s_1) - ks_1 - k\right)T + \left(1 - \bar{t} + ks_1 - \Lambda(s_1)\right) \qquad T \in S \qquad (3.7a)$$

$$= 1 - \bar{t} - (1 - T)\,\Lambda(T) \qquad T \in \bar{S} \qquad (3.7b)$$

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 Figure 11 when processing rates are not capacity restricted and the cumulative flows of Figure 3 are used for $\Lambda(t)$.

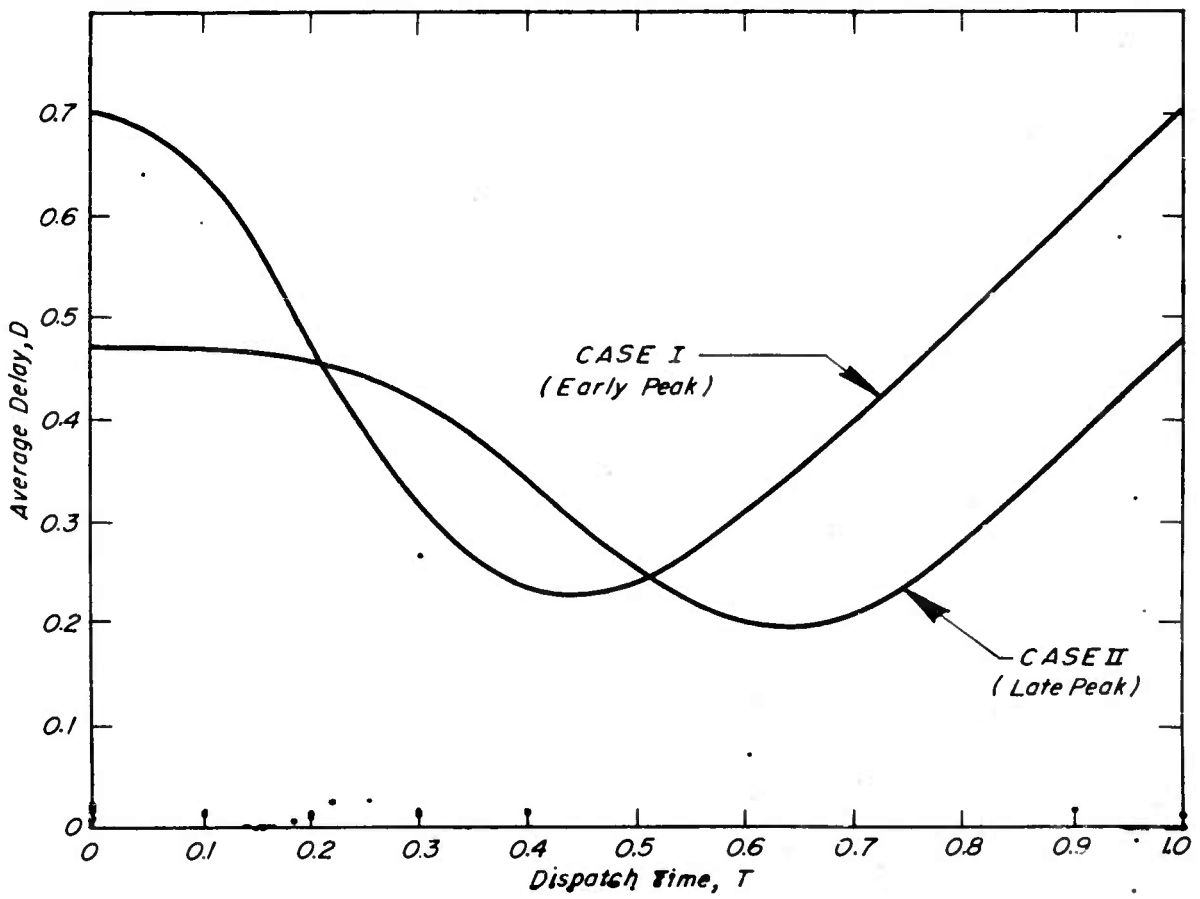


Fig.13–Average letter delay versus dispatch time.

We have already shown that t_1 and T must simultaneously lie within S or simultaneously lie outside S. 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 <u>where</u> letters are delayed but rather, <u>what fraction</u> get delayed.

Optimal Timing of The Dispatch

As one can see from Figure 13 there is an intermediate dispatch which minimizes average delay. In the event that processing rates at stage 1 are not capacity constrained, Equation (3.7b) is the expression for average letter delay in the interval (0, 1). Excluding the term $1 - \bar{t}$ which depends only on the shape of the 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 (3.6) 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 (3.7b) with respect to T equal to zero. T^* is then the solution of

$$\lambda(T)(1-T) = \Lambda(T) \tag{3.8}$$

It seems reasonable that processing rate restrictions at stage 1 will affect the timing of the dispatch which minimizes average delay. We find that T^* either lies inside the interval S or it will be the solution of (3.8). On the other hand, the solution of (3.8) 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 which minimizes Equation (3.7a). In this case, T^* can be explicitly written in terms of the corner-point s_1.

$$T^* = \frac{1 + s_1}{2} - \frac{\Lambda(s_1)}{2k} \tag{3.9}$$

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 which must be answered is, "when is the solution of Equation (3.9) to be used in place of (3.8)?"

Since the average delay (as a function of T) has only one minimum in the interval (0, 1), the sub-interval S can either lie (i) to the left (Figure 14a), (ii) to the right (Figure 14c), or (iii) on either side (Figure 14b) of the optimal dispatch time. Since Equation (3.7a) and (3.7b) are equal and tangent at both corner points and since (3.7a) is always less than (3.7b), we see that the solution of Equation (3.9) can only be the dispatch which minimizes average letter delay if the solution of (3.8) also lies in S, i.e., Figure 14b. If the solution of (3.8) lies outside S then this solution is the optimal dispatch time.

For the solution of (3.9) to be optimal, $s_1 < T^* < s_2$ and the capacity processing rate, $k < \underset{t}{\text{Max}}\, \lambda(t)$, must also satisfy the two inequalities

$$\frac{\lambda(s_1)}{1 - s_1} < k < \frac{\lambda(s_1)}{1 - 2s_2 + s_1} \qquad (3.10)$$

Both of these inequalities can be obtained from Equation (3.9) or from the equivalent statement that the slope of D in Equation (3.7) 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 (3.8).

It should now become clear to the reader that since the decision variables of processing rates and dispatch times are independent of one another,

† Alternatively, if $k < \frac{\lambda(s_1)}{1 - s_1}$, T^* is the solution of (3.8) and lies to the left of s_1; if $k > \frac{\lambda(s_1)}{1 - 2s_2 + s_1}$, T^* lies to the right of s_2.

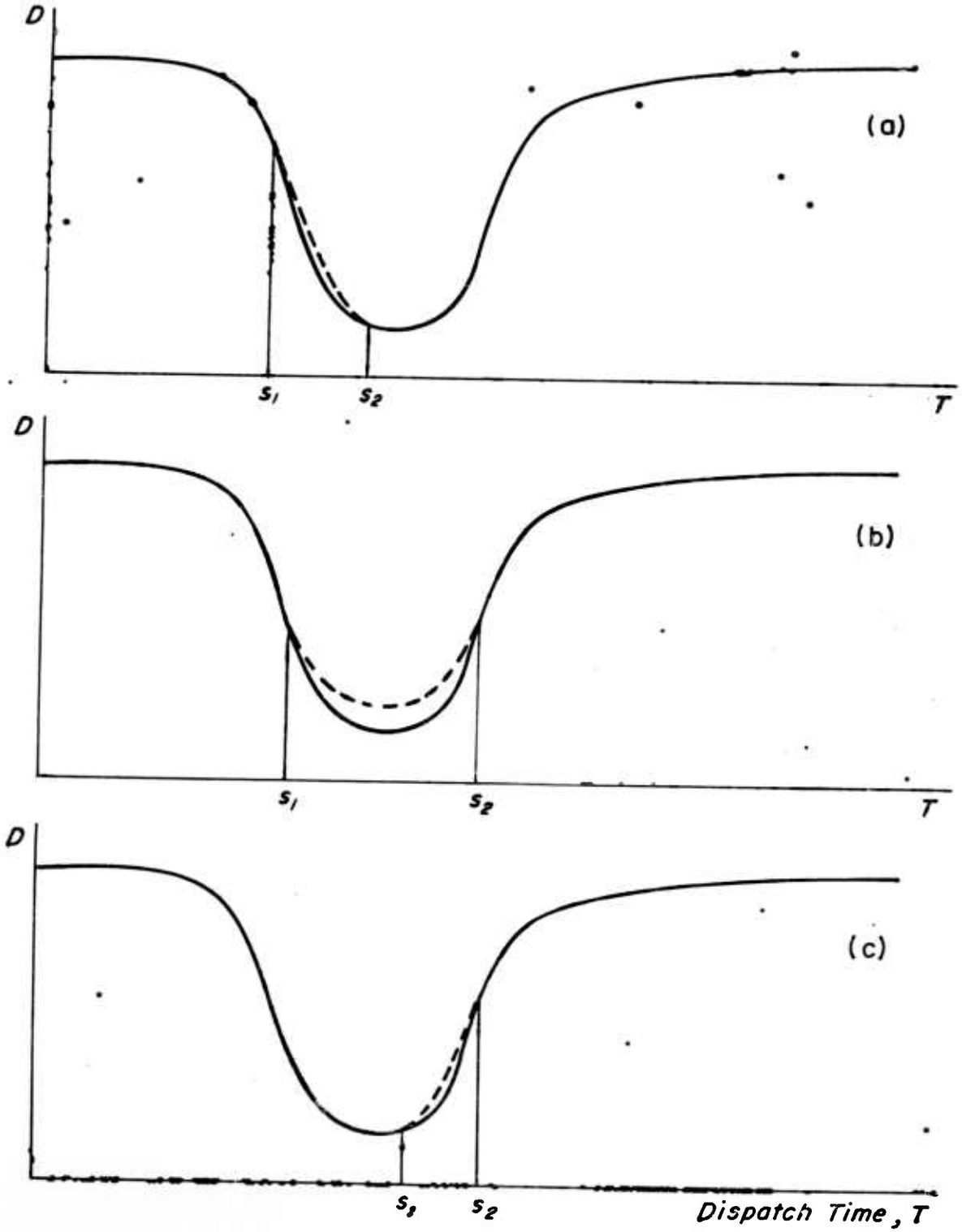


Fig.14-Optimal timing of one dispatch.

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 Section 2 we found that all letter delays were concentrated at the first stage; hence if we replace k in Equations (3.1) - (3.10) by the constant $k(\sum_i a_i)^{-1}$ of Equation (2.36), 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 Equation (3.9) into (3.7a),

$$D^* = 1 - \bar{t} - \frac{k}{4} + \frac{2k-1}{4k}\left(ks_1 - \lambda(s_1)\right) \qquad (3.11)$$

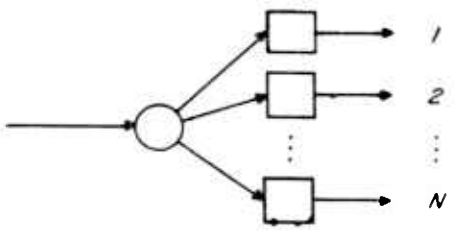
Comparison of this expression for D^* with D in Equation (3.7b) gives a quantitative estimate of the merits of (i) increasing processing rates with present non-optimal dispatch times versus (ii) using optimal dispatch times with present capacity restrictions on processing rates.

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 single dispatch. Secondly, they serve as an introduction to the final parts of Section 3. Thirdly, and perhaps most important of all, they answer some of the dispatch problems which will arise when semi-automatic and fully automatic sorting machines of the future obviate many if not all of the 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 Figure 15. Mail sorted at stage 1 will collect in each of the N storage stages and await dispatch.



Stages: Sorting Storage

Fig.15 - Sorting into N storage stages.

The optimal timing of one intermediate dispatch from each of the N storage stages in Figure 15 is no more difficult to find than the one-branch case considered in Figure 11. Since the dispatch times of each branch are independent of one another, the average delay of letters is given by

$$D = 1 - \bar{t} - \sum_{i=1}^{N} \alpha_i (1 - T_i)\, \Lambda(t_i) \qquad (3.12)$$

where we again interpret t_i as the time of arrival of the last letter making the intermediate dispatch T_i in the i^{th} branch. The optimal dispatch time T_i^* of the i^{th} branch equals T^* in Equation (3.8) or (3.9) since the optimal dispatch times are independent of the constant fraction sorted at stage 1. Hence, the arguments following Equation (3.9) apply; all dispatches are scheduled at the same time and the minimum average delay is still given by Equation (3.11).

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 j^{th} dispatch, we now obtain a simple generalization of Equations (3.6) and (3.7)

$$D = 1 - \bar{t} - \sum_{j=1}^{m} (T_{j+1} - T_j)\, \Lambda(t_j) \qquad (3.13)$$

where T_j is the time of the j^{th} dispatch, $T_{m+1} = 1$ and t_j is now the arrival time of the last letter which makes the j^{th} dispatch. The timing of the optimal dispatches, when sorting rates are not capacity restricted at stage 1, is also a simple generalization of Equation (3.8):

$$\Lambda(T_j) - \Lambda(T_{j-1}) = \lambda(T_j)(T_{j+1} - T_j) \qquad 1 \le j \le m \qquad (3.14a)$$

When we interpret T_0 as zero, Equation (3.14a) represents m equations in

the m unknown dispatch times of say the i^{th} branch.$^{\neq}$ Since $\Lambda(t)$ is a non-decreasing function and $T_{j+1} > T_j$ there is no difficulty in proving the existence of solutions.

When sorting rates are restricted in $S = (s_1, s_2)$ a subset, $m - M$, of the m dispatches will be the solution of Equation (3.14a) while the remaining M will be the solutions of the M equations obtained by substituting $\Lambda(t_i)$ for $\Lambda(T_i)$ on the left hand side of (3.14a)

$$\Lambda(t_j) - \Lambda(t_{j-1}) = \lambda(t_j)(T_{j+1} - T_j) \qquad 1 \leq j \leq M \qquad (3.14b)$$

The number of optimally timed dispatches inside S is equal to the number of solutions of Equation (3.14a) inside S since, as we have shown earlier, there is a one to one correspondence between the solutions of Equation (3.14a) and (3.14b) in S.

Since the times of arrival of the last letter making the i^{th} dispatch can be obtained from Equation (3.2) by substituting t_i for t_1 and T_i for T, we are again able to obtain the solutions for the M optimal dispatches in S in terms of the corner point s_1. These solutions are obtained in exactly the same way as was the solution of Equation (3.9):

$$T_j^* = \frac{1}{M+1}\left\{j - (M+1-j)\left(\frac{\Lambda(s_1)}{k} - s_1\right)\right\} \qquad 1 \leq j \leq M \qquad (3.15)$$

$\neq$ In order to include the effect of many dispatches from many branches we can replace the last term of Equation (3.13) by

$$\sum_{i,j}^{n,m_i} \alpha_i (T_{i,j+1} - T_{i,j})\, \Lambda(t_{ij})$$

where the subscript (i) refers to branch and (j) to dispatch. However, as we have pointed out, we need only study the one-branch case to be able to solve the many branch case.

T_j^* is linear in j and points up the case of special interest where input flow rates are constant. For, if we were to substitute k = constant for $\lambda(t)$ in Equation (3.14a), the optimal dispatch times would divide the interval (0, 1) into m equal parts. Excluding the contribution due to the non-constant flow rates up to time s_1, we have obtained essentially this same result in Equation (3.15) in the interval (s_1, s_2). Again, expressions for the minimum average delay, D^*, are simple generalizations of Equation (3.11) and they need not be re-derived here.

The flow of mail from one of the storage stages is zero except at the j^{th} 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 closely resemble the output from the sorting or processing stage. Each pulse of mail will be small but so long as the dispatches are evenly spread out and not concentrated at one point the output flows from the storage stages should have the effect of replacing one or two high-volume pulses of mail by a 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 (3.8) and (3.14a) 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 (3.13) has the limiting expression

$$\lim_{m \to \infty} D = \lim_{m \to \infty} D^* = \lim_{m \to \infty} \left\{ 1 - \bar{t} - \sum_{j=1}^{m} \frac{1}{m} \lambda \left(\frac{j}{m} \right) \right\} = 0 .$$

[†] or $\alpha_i(\Lambda(t_{i,j}) - \Lambda(t_{i,j-1}))$ if we consider the storage stage in the i^{th} branch.

The Primary and The Secondary

If a post office contained only one major sorting area the dispatch problems which we have studied might go a long way towards understanding and improving letter delays. But there are at least two additional complications which arise: a Secondary may follow the Primary and while dispatches from early storage stages may be variable, dispatches from the final storage stages may be fixed and not under the explicit control of the post office. The last mentioned restriction is often due to the fixed departure schedules of trains and planes. Historically, these schedules have been determined by the management of transportation facilities for reasons which are more closely tied to passenger and freight than to mail service.

Even while dispatch times in the Secondary may be fixed, decision rules must be sought for the release of mail inventories from the Primary storage areas. But before we do this we must attempt to answer an important question which is often asked: Why can't each Primary branch feed 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 artifically interrupted by storage?

These storage stages exist partly by accident and partly by design. 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 analyse the effects of parallel stages which 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 which 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 which could collect the mail until economical and timely transportation was provided.

Quite naturally, transportation facilities which carry mail between sorting areas within a post office differ quite radically from those which 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 Section 4 are those which depend on a system of belts and conveyers to carry the mail from sorting to storage areas. There are still 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.

The problems which now arise in the Secondary sorting areas are primarily due to the nature of the large pulses of mail which 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 instantaneously 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 and labelled or packaged in some form which preserves the distinct categories into which the mail has already been sorted.

The sorting and processing rates which are available in a Secondary are, of course, not infinite. The obvious conclusion is that one need only calculate the amount which 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.

As the reader might suspect, the flow and delay problems which arise in the Secondary can be partially answered by considering a single branch. In Figure 16, 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.

Let us fix a single dispatch time of stage 4 at $t = 1$ and assume momentarily that stage 1 is not capacity restricted. If the capacity sorting rate at stage 3 is c, the largest inventory which can be released from

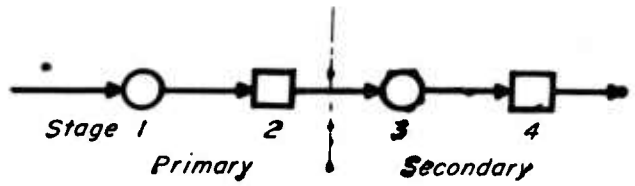


Fig.16—A branch through the primary and secondary

stage 2 and completely processed in stage 3 by the stage 4 dispatch equals $\Lambda(T)$. T is the solution of

$$\Lambda(T) = c(1 - T) \tag{3.16}$$

The argument follows that if T is earlier than the solution of Equation (3.16) 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.

If there are two dispatches from stage 2, the inventory released with the early dispatch, T_1, must be completely processed by the late dispatch time, T_2. Mathematically,

$$\Lambda(T_1) = c(T_2 - T_1) \tag{3.17a}$$

$$\Lambda(T_2) - \Lambda(T_1) = c(1 - T_2) \tag{3.17b}$$

and, in general, for m dispatches from stage 2 the dispatch times are the solutions of:

$$\Lambda(T_j) = c(T_{j+1} - T_1) \qquad 1 \leq j \leq m \qquad (3.18)$$

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

It is useful to study the special case where the input rate to stages 1 and 2 is a constant, $\lambda(t) = k$. In this case the solution of Equation (3.18) for the timing of the j^{th} dispatch from stage 2 is

$$T_j^* = \frac{1 - \rho^j}{1 - \rho^{m+1}} \qquad 1 \leq j \leq m \qquad (3.19)$$

where $\rho = kc^{-1}$, the ratio of the constant arrival rate to the capacity sorting rate at stage 3. The last dispatch in stage 2, T_m^*, also represents the largest fraction of sorted mail which can make the final dispatch from stage 4.†

With a large number of dispatches $T_m^* = \frac{1 - \rho^m}{1 - \rho^{m+1}} \to 1$ for $\rho \leq 1$.

Figure 17 shows how a small number of optimally timed dispatches from stage 2 affects the fraction of mail volumes processed by Secondary dispatch time at $t = 1$.

If the fixed dispatch time of the final stage were not at $t = 1$ but at some other time $X < 1$, Equations (3.16), (3.17) and (3.18) remain correct if we substitute X for 1. The solutions of Equation (3.19) are also correct if we replace T_j^* by T_j^*/X.

The graphical solutions of Equation (3.18), for general $\lambda(t)$ and two dispatches, is shown in Figure 18. T_1 and T_2 are the two dispatch times at 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 inven-

† We are still assuming, of course, that the final dispatch leaves at $t = 1$, after all the mail has entered the post office.

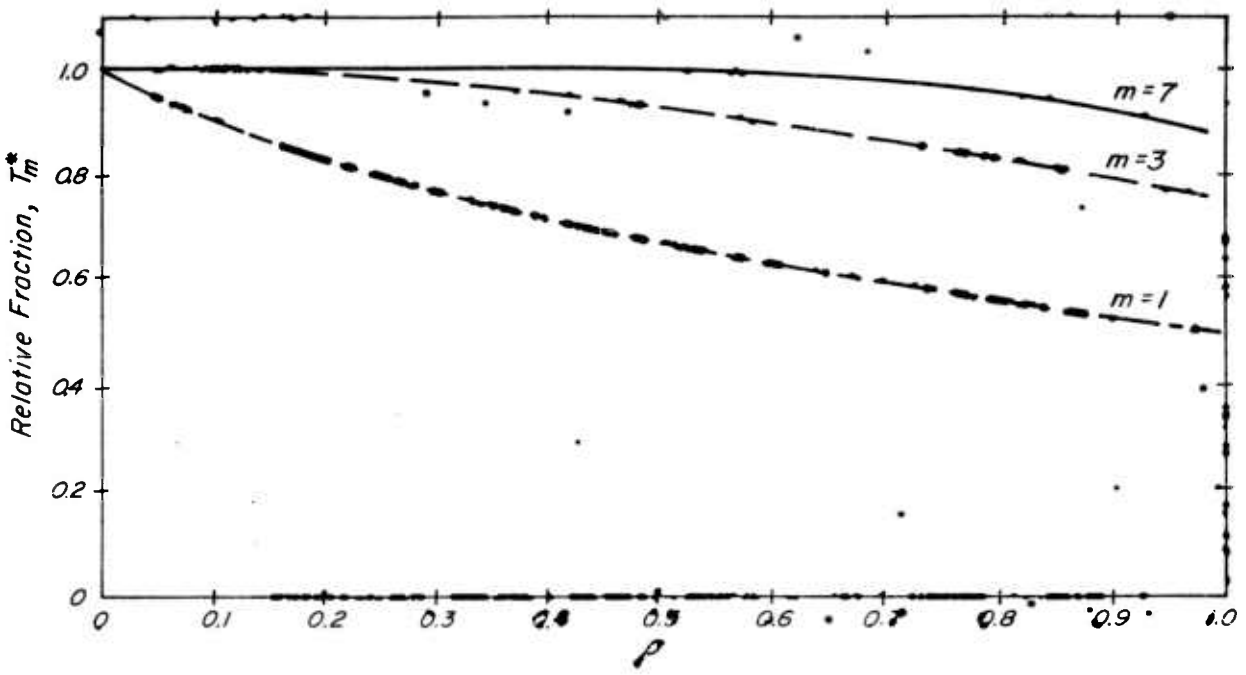


Fig. 17- Relative fraction making final dispatch.
(constant flow rates)

tory sorted in the interval (T_1, T_2) is $A(T_1)$. The amount of inventory sorted in the interval (T_2, X) is $A(T_2) - A(T_1)$. Hence, the difference between the cumulative input curve, $A(X)$, and $A(T_2)$ is the amount of mail which 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 Figure 18.

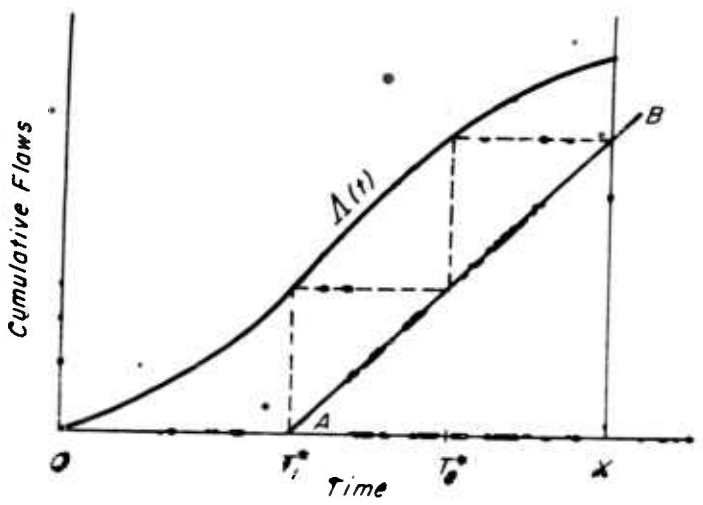


Fig. 18 - A graphical solution of the two dispatch case.

We have shown in (3.16) that these solutions maximize the fraction of mail which makes an early (rather than a late) Secondary dispatch. This would seem to be identical to the optimal policy which 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 which has to wait and the time to the next dispatch.

If, for example, there were only one daily fixed dispatch from the Secondary, the effect of carrying inventories from day to day could be studied by locating final dispatches at $t = 1, 2, 3 \ldots$ Excluding the unusual conditions during the Christmas and New Year season, we can assume that mail is never stored more than one day and, in particular, that there is ample capacity to process the peak flows of one period before those of a

second period arrive. In this case, we need only study two successive days and substitute $1 - t$ $(t < t_1)$ and $2 - t$ $(t \geq t_1)$ for $\tau(t)$ in Equation (3.6a). The average letter delay is equal to a constant term minus $\Lambda(t_1)$, the fraction of mail making the first Secondary dispatch; hence D is always decreased by increasing t_1. If T is the timing of the Primary dispatch, $\Lambda(t_1) = \Lambda(T)$ when $\Lambda(T) < c(1 - T)$; $\Lambda(t_1)$ equals $c(1 - T)$ when $\Lambda(T) > c(1 - T)$. Obviously, the average delay is minimized when the Primary dispatch is the solution of Equation (3.16).

Figures 19 and 20 are plots of the optimal timing of Primary dispatches and the corresponding fractions of mail which make the early Secondary dispatch at time $X < 1$. The ordinates are the times of the dispatches or the fraction of mail making a dispatch while the abscissas represent the capacity sorting rate, c, in a Secondary stage. The numerical calculations were based on the cumulative input flows of Case II in Figure 3.

In Figure 19 we note that the timing of the last Primary dispatch rapidly approaches that of the early Secondary dispatch at X when sorting rates are greater than 2.0. Since c can also be interpreted as the ratio of the maximum mail volumes that can be processed to the total volume actually processed, we were able to identify the fact that manual sorting stages in the Secondary had values of c in the range 1.5 to 14.0! Hence, even with the introduction of semi- and fully-automatic high-speed sorting equipment one should not necessarily expect to find significant changes in the timing of the optimal dispatches. Since our analysis indicated that the largest contribution to average delay arose from the infrequent and ill-timed dispatches from storage, this raises natural questions as to the efficacy of costly high-speed sorting equipment.

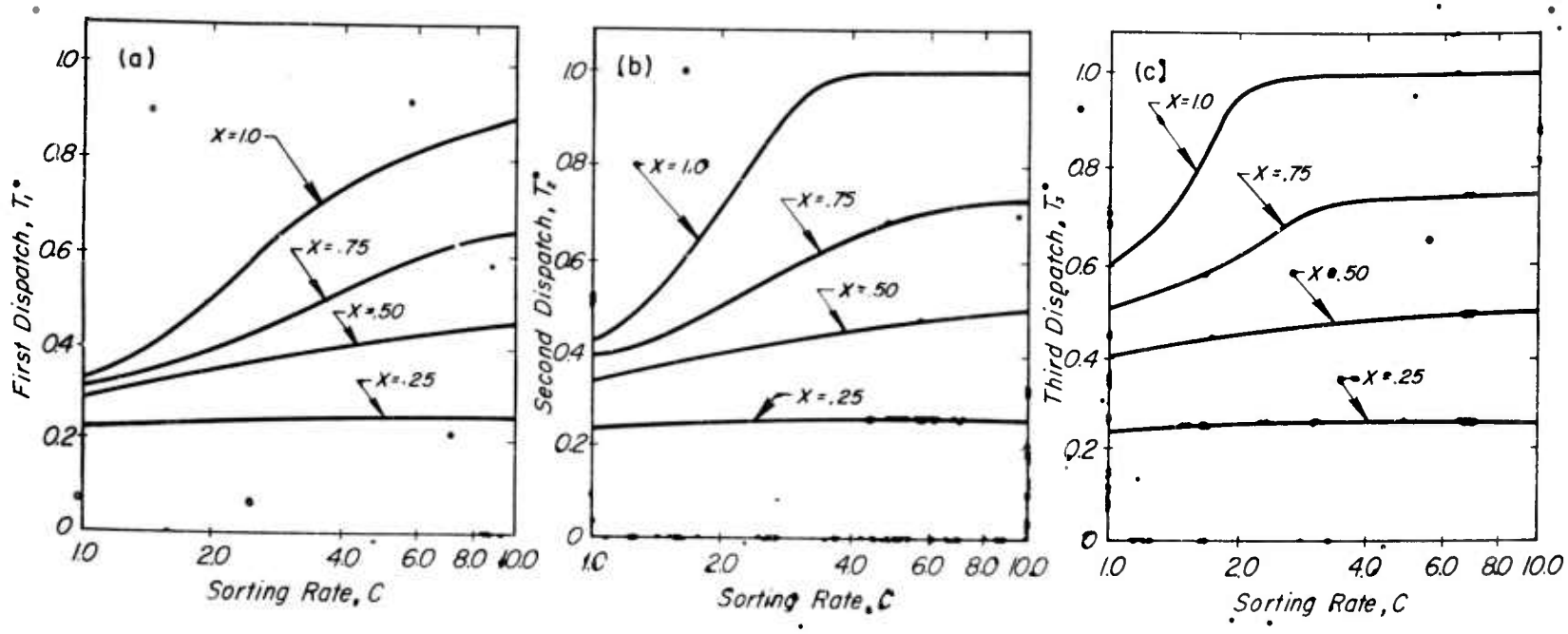


Fig. 19 – Optimal timing of three primary dispatches.

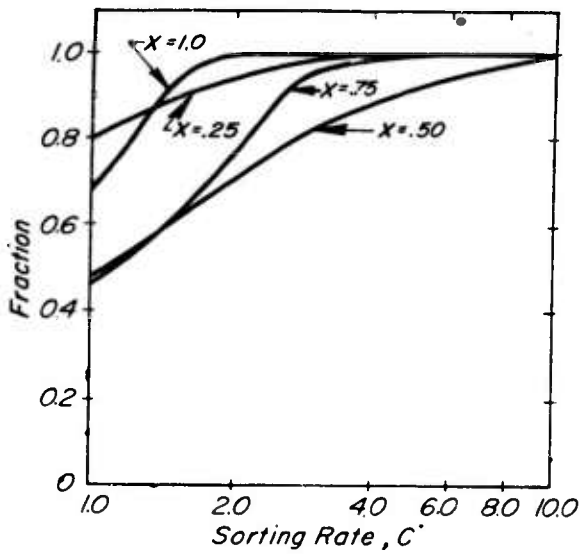


Fig. 20-Fractional mail flow making secondary dispatch at X.

In Figure 20 the fractional scale along the ordinate has been normalized by dividing the fraction of mail which makes a Secondary dispatch by the fraction of mail which would make the same Secondary dispatch if capacity sorting rates were infinite. For this reason the limiting values of this fraction approach unity for large sorting rates. Again, the introduction of high-speed sorting equipment may only bring small increases in the amount of mail dispatched from storage.

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 which have been studied and, to one extent or another, 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.

4. 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 re-routing of all classes of mail through the City of Detroit.

While the Roosevelt Park Annex sorted and routed all classes of mail, the experiments which are described in this section were restricted to sorting and processing operations of first class letter mail. These operations were carried out on the third floor of the Annex. 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

[†] A new post office has since been built to handle these operations.

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 Figures 21 and 22.

At this point is 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 (precanceled 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 slipped directly to the Outgoing Primary; airmail and special delivery mail were removed at this stage and 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 definitions (receiving $\sim 5\%$ of the mail) were classified as "Residue". They were sorted in a Tertiary labelled according to their railroad branch lines.

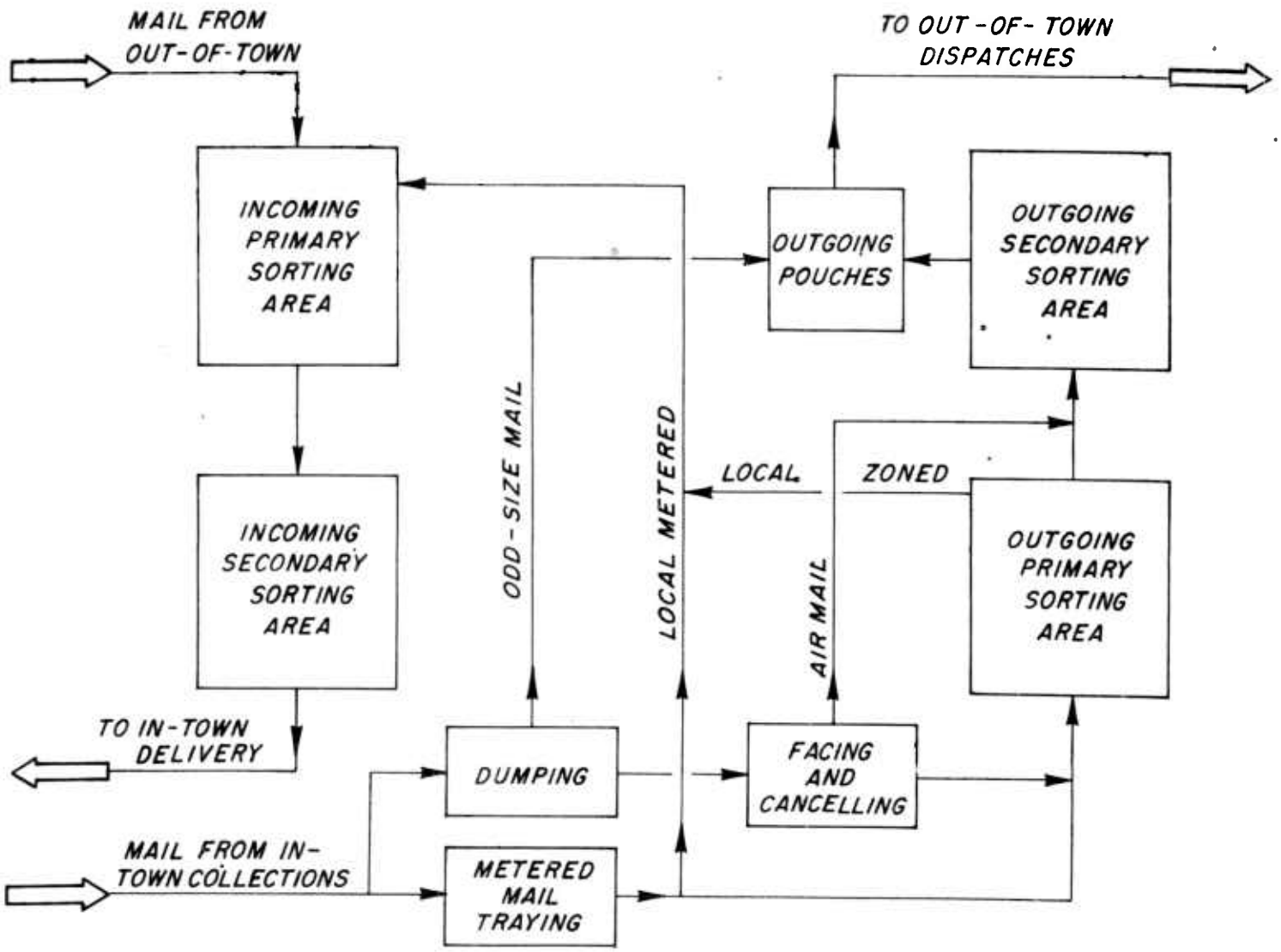


Fig. 21- Mail flows through a post office.