Optimization of Traffic Signal Settings by Mixed-Integer Linear Programming*

Part II: The Network Synchronization Problem

NATHAN H. GARTNER,
JOHN D. C. LITTLE
and
HENRY GABBAY

Massachusetts Institute of Technology, Cambridge, Massachusetts

Results obtained in Part I are extended to include offsets, splits at each intersection, and a common cycle time for the network as simultaneous decision variables. In addition to the deterministic link performance function, the stochastic effects of overflow queues from one cycle to the next are modeled by means of a saturation deterrence function that enters as an additive component in the objective function. Computational results demonstrate the feasibility of using mixed-integer linear programming on problems of realistic size. Sensitivity analysis of cycle time shows it to have a strong influence on network performance.

Part I of this paper[1] has considered the network *coordination* problem, i.e., given a common cycle time and splits at each intersection, determine the optimal offsets throughout the network. The mathematical model for this problem consists of an objective function comprising the individual link performance functions and the network loop equations constraining

* This research was supported (in part) by KLD Associates under DOT Contract FH-11-7924 and (in part) by ARO Contract DAHCO4-73-COO32.

the algebraic sum of the offsets around any loop of the network to an integral multiple of the cycle time.

The link performance function and the loop equations tie together all the traffic signal control variables of the system. This interdependency suggests that a comprehensive optimization program for the network should let all the coupled control quantities be simultaneous decision variables. Modeling and computational difficulties have prevented the achievement of this goal up to now. The result is that all the existing optimization procedures use essentially a sequential decision process, with some of them allowing iterations among the different stages.

In SIGOP*[I-6] a predetermined number of cycle times is scanned. For each cycle, splits are determined locally at each intersection and offsets are optimized by the OPTIMIZ subroutine. The system's performance is then evaluated by a coarse simulation of traffic flow on all the links of the network by the VALUAT subroutine. The optimal set of cycle time, splits, and offsets is selected according to the results obtained from VALUAT. Three deficiencies of this procedure are apparent. First, the offset optimization procedure determines a local, not necessarily global, optimum. Second, splits are determined independently, and third, stochastic effects on link performance are ignored. Thus, in one study using SIGOP,[2] the lower bound on cycle time was consistently selected as the optimal value, where as the inclusion of stochastic effects quite conceivably might have shifted the results upwards.

The Combination Method and TRANSYT use a critical intersection approach for determining the network cycle time, TRANSYT allowing also for evaluation of performance at different cycle times.[3] Since through capacity increases with cycle length, the critical intersection approach is based on analysing the capacity requirements of the most heavily loaded intersection, i.e., the intersection with the highest sum of demand/capacity ratios on conflicting signal phases. A procedure that is used for a single intersection, such as Webster's method or a prespecified degree of saturation on all approaches to an intersection is then used to calculate the cycle length. This may not be optimal in a network situation since the interaction among the flows and the spatial road network structure of the area is disregarded. The dynamic programming procedure,[I-9,4] which is a generalization of the combination method, is in principle capable of searching over all the variables. However, because of the 'curse of dimensionality,' only small networks can be treated in this way and including the splits as decision variables is computationally prohibitive.

In this part of the paper, stochastic phenomena are introduced into the

* Prefix I refers to references and equations in Part I.

basic mode of Part I by regarding the traffic flow on a link as the combination of a periodic component imposed by the preceding signal and a random component arising from variations in driving speeds, marginal friction, turns, etc. The latter causes additional delays owing to the occurrence of an overflow queue at the signal's stop line. The overflow queue represents the number of vehicles that are not able to clear the intersection during the preceding green phase. While this effect is negligible at low degrees of saturation, its predominance at high levels has been established in several studies.[I-18,I-19,5] TRANSYT also uses a random delay component that was calibrated to fit observed traffic behavior.[I-10] The traffic signal network timing problem is next formulated in a general form to reflect the increased scope of optimization.

1. THE TRAFFIC SIGNAL NETWORK TIMING PROBLEM

IN THIS SECTION all the signal parameters (cycle, splits and offsets) are considered variables. We have, therefore, additional constraints to those considered in Section I-1. For each link in the system we have the following relation:

$$r_{ij} + g_{ij} = C. \qquad \forall (i, j) \in L \quad (1.1)$$

Equation (1.1) also implies the relation between settings on opposing links given in Fig. I-1.

In order for the network to handle the given flow, capacity requirements must be met:

$$s_{ij} g_{ij} \geqq f_{ij} C. \qquad \forall (i, j) \in L \quad (1.2)$$

To facilitate pedestrian crossing we may have a minimum red requirement, $r_{ij} \geqq (r_{ij})_{\min}$. Since the gain in capacity with very long cycle times is often insignificant, we should have an upper limit on the cycle length $C_{\max}$. This is also desirable to prevent drivers from becoming impatient or believing the signals are defective. From a practical point of view, including safety considerations, it is also desirable to have a lower limit on the cycle time $C_{\min}$ in addition to capacity considerations.

As the signal's capacity is now affected by the decision variables we must consider also the delay contributed by the stochastic component. This delay is represented by a *saturation deterrence function* (SDF), which is developed in terms of the overflow queue $Q_{ij}(r_{ij}, C)$. [For simplicity we consider only r_{ij} since by (1.1), $g_{ij} = C - r_{ij}$.] The network signal setting program is now stated in its general form as a mixed-integer nonlinear program.

MINLP: *Find* values of ϕ_{ij}, r_{ij}, C *to*

$$\min \sum_{(i,j) \in L} f_{ij} z_{ij}(\phi_{ij}, r_{ij}, C) + Q_{ij}(r_{ij}, C),$$

subject to: Loop constraints, equation (I-1.3), and

$$\left.\begin{aligned} r_{ij} + g_{ij} &= C, \\ g_{ij}s_{ij} &\geqq f_{ij}C, \\ r_{ij} &\geqq (r_{ij})_{\min}. \end{aligned}\right\} \forall (i, j) \in L$$

$$C_{\min} \leqq C \leqq C_{\max},$$

$$r_{ij},\ g_{ij} \geqq 0,$$

ϕ_{ij}, n_i, unrestricted in sign; n_i integer.

Solution of this program by mixed-integer programming[I-11] will be described in two stages. First, offsets and splits will be considered, with cycle time fixed. Then, all three sets of variables, including cycle time, will be taken up simultaneously.

2. VARIABLE SPLIT FORMULATION

IN THIS SECTION we add the split at each node of the network as a decision variable to our formulation of the traffic signal network coordination problem in Part I.

2.1 Link Performance Function-Approximation by Planes

To obtain piecewise linear relations in the (γ, r, z) space described in Fig. I-7 we have to fit planes to the surface $z(\gamma, r)$. This is accomplished by selecting triplets of points. A plane $z = k_1 + k_2\gamma + k_3 r$ can be formed from the points $P_i(\gamma_i, r_i, z_i)$, $i = 1, 2, 3$ by simple geometrical relations. Referring to Fig. 1 we define five planes α_J by the following triplets:

$$\alpha_{\mathrm{I}}: \{P_1, P_2, P_3\} \qquad \alpha_{\mathrm{II}}: \{P_2, P_3, P_4\} \qquad \alpha_{\mathrm{III}}: \{P_i \mid z_i = 0\}$$

$$\alpha_{\mathrm{IV}}: \{P_4, P_5, P_6\} \qquad \alpha_{\mathrm{V}}: \{P_2, P_4, P_6\}.$$

The coordinates γ_i, r_i, z_i are determined as follows:

P_1: $\gamma_1 = -r_{\max}$;	$r_1 = r_{\max}$;	$z_1 = z_{\max}(\text{at } r_1)$;
P_2: $\gamma_2 = C - r_{\max} - p$;	$r_2 = r_{\max}$;	$z_2 = z_{\min}(\text{at } r_2)$;
P_3: $\gamma_3 = \gamma_2$;	$r_3 = r_4$;	$z_3 = z_2$;
P_4: $\gamma_4 = 0$;	$r_4 = C - p$;	$z_4 = 0$;
P_5: $\gamma_5 = C - r_{\min} - p$;	$r_5 = r_{\min}$;	$z_5 = 0$;
P_6: $\gamma_6 = p$;	$r_6 = r_4$;	$z_6 = z_{\max}(\text{at } r_4)$;

where $r_{\max} = C - yp$ is determined by the capacity requirement, equation (1.2). $z_{\max}$ is given by (I-2.10) if $p(1 - y) \leqq r$ and by (I-2.8) if $p(1 - y) > r$. $z_{\min}$ is also given by (I-2.8). To obtain the piecewise linear surface as a function of offsets we have to substitute $\gamma = \tau - \phi$.

Planes α_{I} and α_{II} are parallel to the r-axis by Theorem I-2.5. Therefore $k_3^{\mathrm{I}} = k_3^{\mathrm{II}} = 0$ and we can determine the planes by two points only. Planes $\alpha_{\mathrm{I}}, \cdots, \alpha_{\mathrm{V}}$ enclose a convex subspace that constitutes a piecewise linear

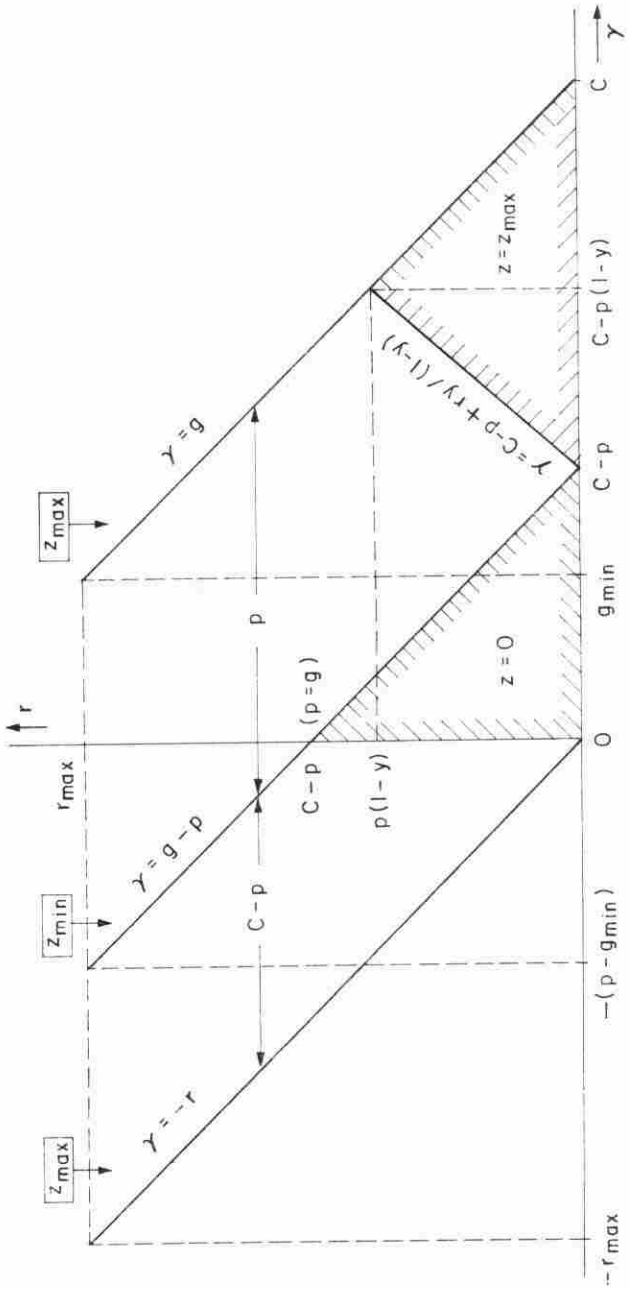

Fig. 1. Planar approximations to link performance function.

approximation to the three-dimensional surface $z(\gamma, r)$. The points defining the planes have been chosen to approximate most closely the bottom part of the valley-like enclosure, i.e., where the optimal solutions are eventually expected. A typical set of $z(\gamma, r)$ that were calculated for a rectangular-shaped platoon and their approximation via the planes described above, are shown in Fig. 2.

An important parameter in determining the $z(\gamma, r)$ surface and its planar approximations is the platoon length p_{ij} on each link. In Section I-2 the platoon characteristics were calculated according to the green splits at the intersection, which were fixed and precalculated according to Webster. Since in the current context the upstream green time is itself a decision variable, a modified procedure is employed. Given the cycle time, we first

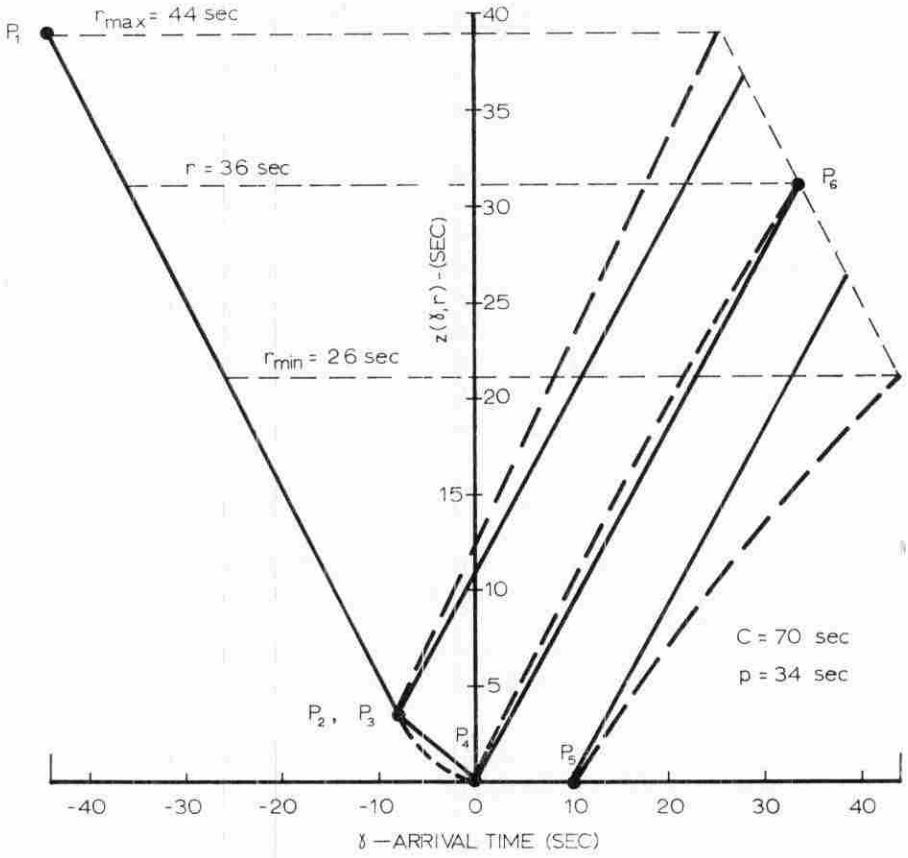

Fig. 2. Projections of (z, γ, r) space.

termine nominal splits at each node, e.g., by Webster's method. These splits determine the platoon lengths according to the procedure of Section I-2. Though the final splits might be slightly different than those assumed initially, the procedure gives satisfactory results.

Since input flows are a prime determinant of the splits for signals on the boundary of the network, *input links* must be included in the program. Traffic on these links is assumed to be unaffected by other controls and therefore is modeled as a continuous flow, i.e., $p = r + g = C$. Stochastic effects will be included later. We can use equation (I-2.8) in this case, thus obtaining for the average delay,

$$z = r^2/2C(1 - y). \tag{2.1}$$

Equation (2.1) is equivalent to the deterministic delay component in Webster's delay formula[I-15], and is piecewise linearized with respect to r.

2.2 Stochastic Effects—The Saturation Deterrence Function

At flows that are close to capacity but still, on the average, below it, occasional fluctuations in the size of the platoon can lead to temporary overflow queues that seriously degrade performance. This is a stochastic phenomenon, which has a consequence that, as average flow approaches capacity, average delay increases gradually at first and then very rapidly. A representation of this effect is needed to prevent green time from approaching its lower bound too closely. It is particularly essential if cycle length is a variable to be determined by the optimization process. This is because of the fundamental trade off between capacity loss at short cycles and the inherently large delays of long cycles.

WORMLEIGHTON[6] has studied the overflow queue in some detail. Following his model, we assume

1. Arriving vehicles come in platoons or other periodic function of time at a rate $q(t)$.

2. Arrivals are a nonhomogeneous Poisson process. Thus the number of vehicles in $(t, t + c)$ is a random variable having a Poisson distribution with mean A_C given by equation (I-2.1).

3. The service provided by the green time is deterministic with rate s(veh/sec) while a queue exists. This means gs vehicles are served in a cycle generating an overflow queue.

If the state of the intersection is examined at the end of green, a bulk service queuing model is defined. Let

$Q(0)$ = number of vehicles in queue at the end of green (start of red); this is an overflow queue.

$A_C = fC$ = average number of vehicles arriving in a cycle.
$S = gs$ = number of vehicles that can be served in one green time (number of release points).
$x = A_C/S = fC/gs$ = degree of saturation.

Wormleighton finds the generating function of $Q(0)$ and calculates $E\{Q(0)\}$ as a function of S and x over $S \in [5, 55]$ and $x \in [0.2, 0.975]$. Within these ranges $E\{Q(0)\}$ varies from essentially zero up to 18 vehicles (Table I).

TABLE I

Expected Overflow Queue, Q

$S = gs$, No. of Release Points per Cycle	x, Degree of Saturation						
	0.20	0.40	0.60	0.80	0.90	0.95	0.975
5	0.00	0.02	0.20	1.15	3.50	8.41	18.36
15	0.00	0.00	0.04	0.70	2.81	7.61	17.50
25		0.00	0.01	0.47	2.41	7.08	16.91
35			0.00	0.34	2.11	6.68	16.45
45				0.23	1.88	6.34	16.05
55					1.68	6.02	15.67

We wish to understand how the presence of overflow queues affects delays as a function of offset. Insight into the situation is obtained by examining the deterministic case. A theorem is proved below to show that under certain circumstances delay can be broken into two components; one depending only on the overflow queue and the other being the delay that would be incurred in the absence of the overflow queue. This motivates the introduction of the overflow queue effect into the link performance function in a simple way.

Consider the sketch of queue vs. time during a cycle in Fig. 3A Let,

$Q(t)$ = number in queue at time t,
$Q(0)$ = overflow queue at start of red,
$Q_0(t)$ = number in queue at t if $Q(0) = 0$,

$$Z = \int_{-r}^{+g} Q(t)\, dt = \text{aggregate delay over cycle (veh/sec)},$$

$$Z = \int_{-r}^{+g} Q_0(t)\, dt = \text{aggregate delay if } Q(0) = 0 \text{ (veh/sec)}.$$

It is assumed, as before, that $q(t) \leqq s$, i.e., the instantaneous arrival rate never exceeds the instantaneous service rate.

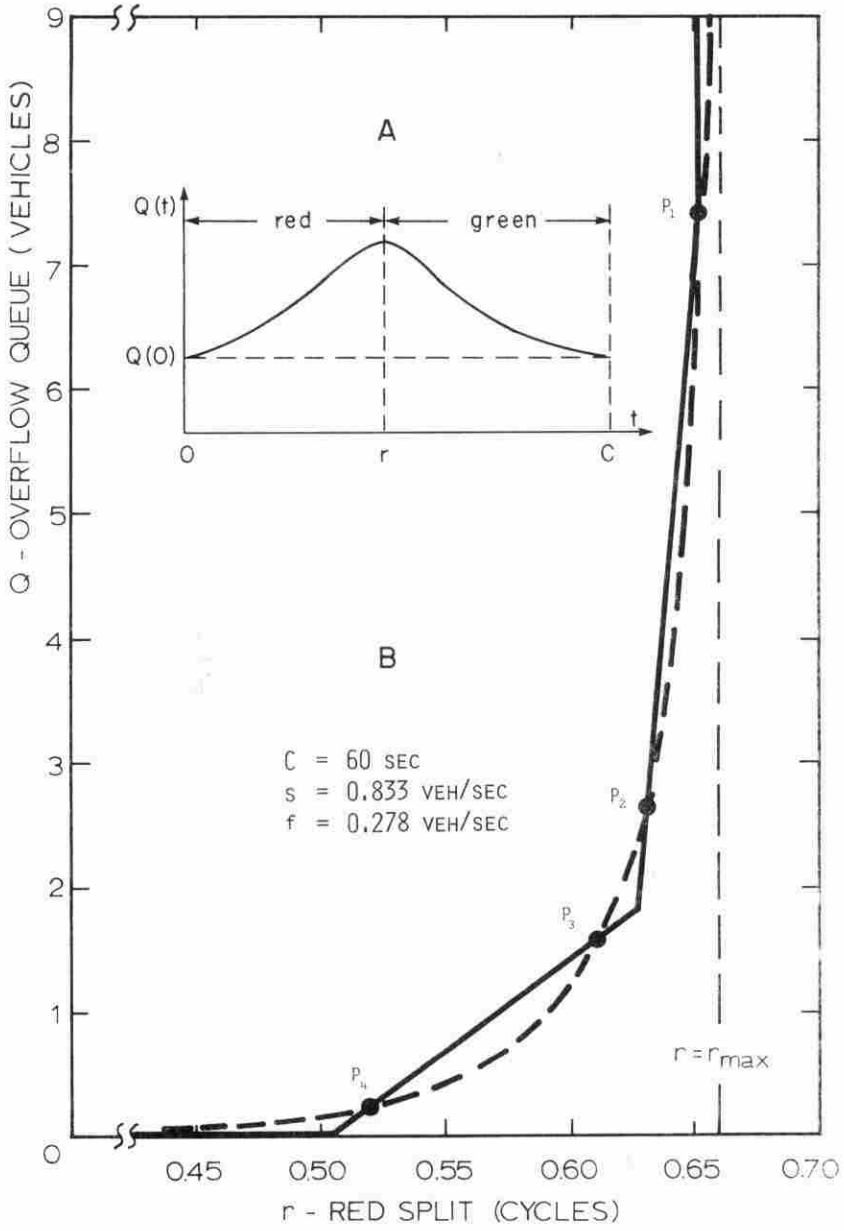

Fig. 3 A. Queue evolvement during a saturated cycle (including overflow queue).
B. Saturation deterrence function and piecewise linear approximations.

THEOREM 2.1. *If the number of arrivals equals the maximum number of vehicles that can be served, then*

(a) *the overflow queue is preserved,* $Q(C) = Q(0)$.

(b) *the aggregate delay is composed of a term equal to the wait of the overflow queue and a term equal to the delay in the absence of the overflow queue*:

$$Z = Q(0)C + Z_0. \tag{2.2}$$

Proof. See Fig. 3. Because $A_C = gs$, all of the green is used for serving vehicles. $Q(t) \geqq Q(0)$ and the terminal queue returns to its starting point. Therefore, $Q(C) = Q(0)$, proving (a).

Next observe that the order of service does not affect the number in queue and so does not affect the aggregate delay. Suppose, therefore, that the overflow queue vehicles are served last. Then they are, in fact, never served and the arriving vehicles are served just as if $Q(0) = 0$. The delay of the vehicles arriving during the cycle is Z_0. The delay of the overflow queue vehicles is $CQ(0)$. The total delay is $Z = Z_0 + CQ(0)$ thereby proving (b).

We use the above result to propose that when average flow is near saturation, it is reasonable to separate the total delay into two components, one being the result of the presence of an average overflow queue, the other being the normal delay of vehicles in the absence of an overflow queue. The term involving the overflow queue is regarded as a saturation deterrence function (SDF), which involves increasing penalties to performance as capacity is approached. In terms of delay per unit of time, the SDF component is simply $Q(0) \equiv Q$ and enters additively into the objective function.

To represent the SDF in a form amenable to mixed integer linear programming, we treat it similar to the LPF. First we use secant approximating lines, which are then transferred into the constraint set with corresponding delays contributing additively to the original objective function. For each link we have:

$$S = sg = s(C - r), \tag{2.3}$$
$$x = fC/sg = fC/s(C - r), \tag{2.4}$$
$$r = C(1 - f/sx). \tag{2.5}$$

To assure a minimum level of service we restrict r to be such that the degree of saturation stays below 0.95. Thus we obtain an upper limit on the red split that appears as a vertical constraining line in the (Q, r) plane. Two additional lines are determined by the two pairs of points (P_1, P_2) and

(P_3, P_4), having degrees of saturation $x_1 = 0.95$, $x_2 = 0.90$, $x_3 = 0.85$, and $x_4 = 0.70$, respectively. Using (2.5) and (2.3) we calculate r and S at each point and then obtain the corresponding Q value from Table I. An example is shown in Fig. 3.

2.3 Solution by Mixed-Integer Linear Programming

MILP: *Find* ϕ_{ij}, r_{ij} *to*

$$\min \sum_{(i,j)\in L} (f_{ij}z_{ij} + Q_{ij}),$$

subject to

$$z_{ij} \geqq z_{ij}^{k}(\phi_{ij}, r_{ij}), \qquad k = 1, \cdots, K(\text{LPF})$$

$$Q_{ij} \geqq Q_{ij}^{h}(r_{ij}), \qquad h = 1, \cdots, H(\text{SDF})$$

$$\sum_{(i,j)\in P(l)} \phi_{ij} - \sum_{(i,j)\in R(l)} \phi_{ij} + \sum_{k\in N(l)} \psi_k(l) = n_l C, \qquad \forall l \in \mathcal{L}_f,$$

$$(r_{ij})_{\min} \leqq r_{ij} \leqq C(1 - f_{ij}/0.95\, s_{ij}), \qquad \forall (i,j) \in L,$$

$$n_l^1 \leqq n_l \leqq n_l^u, \qquad (n_l \text{ integer}).$$

ϕ_{ij} is unrestricted in sign. Both *internal links* and *input links* are included. Results for the test network of Fig. I-9 are given in Table II. An improvement of more than 2 percent in the functional value is observed with respect to the solution in Section I-4.

Section 2.1 indicates a cause-effect relation among the splits, which are decision variables of the program, and the platoon length, which is an input parameter to the optimization procedure. Iterating through the sequence platoon → splits → platoon → splits, the optimized splits were used to recalculate the platoon length. It is shown[I-25] that the initial splits are satisfactory in the sense that there are no appreciable discrepancies in the ensuing platoon lengths produced by the iteration. This is due to the fact that the combination of the SDF values for conflicting approaches at an intersection has a parabolic shape. Therefore, any choice of splits will most likely occur in the neighborhood of the minimum, which is quite closely approximated by Webster.

3. THE NETWORK SYNCHRONIZATION PROBLEM

THE PRECEDING SECTIONS first considered the offset optimization problem (the network *coordination* problem) and then added the splits as decision variables. Now we let the cycle time, common for all the intersections of the network, be a decision variable too. This is termed the network *synchronization* problem.

TABLE II

Results of Computations
(C = 80 sec)

Link No.	Offset sec (cycle)	Green Time sec (cycle)	Avg. Delay z_{ij} (sec)	SDF Q_{ij}(veh)	Degree of Saturation
101	24.0 (0.30)	21.6 (0.27)	2.72	0	0.674
102	28.8 (0.36)	23.2 (0.29)	2.08	0	0.628
103	62.4 (0.78)	29.6 (0.37)	34.64	0	0.492
104	17.6 (0.22)	29.6 (0.37)	1.36	0	0.712
105	46.4 (0.58)	29.6 (0.37)	10.64	0	0.492
106	33.6 (0.42)	30.4 (0.38)	7.2	0	0.693
107	14.4 (0.18)	40.8 (0.51)	0.08	1.067	0.817
108	24.8 (0.31)	40.0 (0.50)	2.56	1.339	0.833
109	18.4 (0.23)	39.2 (0.49)	13.36	0.279	0.714
110	−18.4 (−0.23)	49.6 (0.62)	13.52	0	0.355
111	35.2 (0.44)	33.6 (0.42)	11.92	1.505	0.833
112	44.8 (0.56)	39.2 (0.49)	23.6	0	0.449
113	24.0 (0.30)	47.2 (0.59)	0.0	0	0.565
114	44.0 (0.55)	40.8 (0.51)	27.36	0	0.654
115	41.6 (0.52)	40.0 (0.50)	8.32	0	0.700
116	24.8 (0.31)	31.2 (0.39)	4.48	2.035	0.897
Input Links					
117		49.6 (0.62)	9.92	0	0.565
118		49.6 (0.62)	9.92	0	0.565
119		21.6 (0.27)	26.4	0	0.674
120		30.4 (0.38)	20.08	0	0.479
121		29.6 (0.37)	23.04	2.410	0.901
122		33.6 (0.42)	18.64	0	0.524
123		37.6 (0.47)	19.36	1.784	0.887
124		29.6 (0.37)	21.92	0	0.712

Optimal Integer Solution: (0, 1, 1, 1, −1, −1, −1, 0)
Functional Value: 59.6269 veh-sec/sec

3.1 The Optimal Cycle Time in a Network

Experience of researchers and practitioners in urban traffic control has shown that the cycle time may well be the most important among the decision variables.[3,7] Not only does it tie together the offsets in conjunction with integer variables but it also provides for the necessary capacity to serve the traffic demand at each intersection. The net serving capability of an intersection is determined by the fraction of the cycle time that is effective green, i.e., capacity is a function of

$$(1/C)\sum_i g_{ij} = 1 - L_i/C, \qquad \forall i \in N. \tag{3.1}$$

Since L_i, the total lost time at node i, is a fixed quantity at a particular intersection, the net capacity increases with cycle time. This exposes a trade-off. On the one hand, an increase in cycle time usually increases red times on individual phases and, consequently, the average waiting time as expressed by the LPF. On the other hand, an increase in cycle time increases capacity, thus reducing stochastic delay that is due to the randomness in arrivals of vehicles, as modeled by the SDF.

The interplay between the LPF and the SDF in the objective function for variable cycle time has been studied for the test network of Section I-4 (Fig. I-9). The MILP formulation of Section 2.3 was run for various fixed cycle times in the range of 40 sec to 120 sec to optimize offsets and splits. A graphical illustration of the relation is presented in Fig. 4. It becomes evident that the optimal cycle time for the network constitutes a least-cost equilibrium point between delays caused by deterministic effects and delays contributed by stochastic effects. While the first delays usually

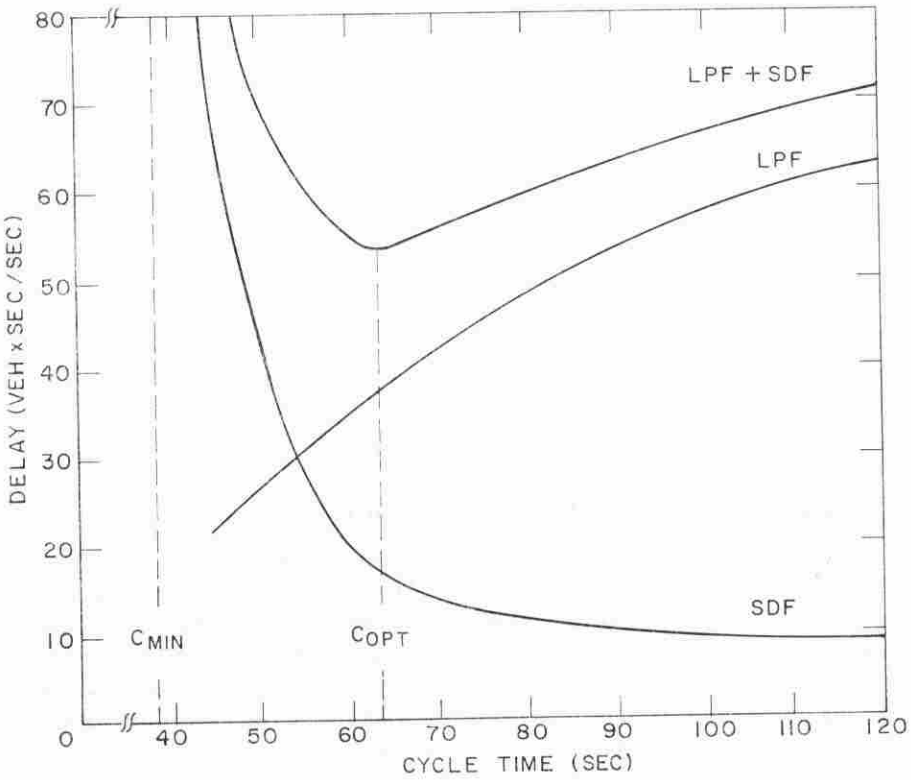

Fig. 4. Variation of delays with cycle time in a synchronized network.

increase with cycle length (though at a decreasing rate), the latter decrease with it owing to the decrease in the degree of saturation. They approach asymptotically from above the minimal cycle time for the network, which is the theoretical minimal cycle time for the most heavily loaded intersection if all flows were strictly deterministic. These characteristics are analogous to the behavior of delay with respect to cycle time at a single intersection and were studied by Wardrop,[I-14] Webster,[I-15] and other researchers.[8,9] However, the implications regarding signal settings in a network are different, because of the added complexity of the network structure, and a single intersection analysis would virtually never give the optimum settings for the network.

3.2 Variable Cycle Formulation

To introduce the cycle time as a decision variable, the formulation of the optimization problem has to be modified. Examining the loop constraints, equation (I-1.3), we observe that for variable cycle time there is a nonlinearity in the term $n_l C$. By dividing each term of this equation by C we introduce such terms as ϕ/C, r/C, and $1/C$. This suggests a change of variable $\phi' = \phi/C$, $r' = r/C$, $w = 1/C$, i.e., we transform variables from seconds to dimensionless fractions of a cycle time. (*In the discussion below we drop primes on variables, understanding that they are measured in fractions of a cycle time.*)

Following the above transformation, p is measured in fractions of a cycle too. We assume that the ratio of platoon length to cycle length (in seconds) remains invariant as we allow the cycle time to vary. This assumption is reasonable since the platoon length on any link (i, j) is considered to be proportional to the upstream green that varies almost linearly with cycle time. Equation (I-1) now becomes $\gamma = \tau w - \phi$. Thus, w enters linearly into the constraints representing the LPF. The LPF for the input links is independent of C and requires no such transformations (τ is still given in absolute time, sec).

The reciprocal cycle w is also introduced into the SDF. As g and r are now given in fractions of cycle time, (2.3) and (2.4) now become

$$S = sgC = s(1 - r)/w, \tag{3.2}$$

$$x = f/sg = f/s(1 - r). \tag{3.3}$$

Solving for r and w we obtain

$$r = 1 - (f/sx), \tag{3.4}$$

$$w = f/Sx. \tag{3.5}$$

We may now transform the triplet of points (Q, S, x) into (Q, r, w), which coincides with our decision space, and use planar approximations for the latter. Since the overflow queue is more sensitive to variations in split than in cycle time, the planes representing the SDF (appearing as constraints in the MILP formulation) are of the form $Q + ar + bw \geqq 0$ with a-values that are considerably higher than the b-values.

The original objective function is

$$\min \sum_{(i,j)} (f_{ij} z_{ij} + Q_{ij})$$

with z_{ij} in seconds. Introducing C as a decision variable we need a LPF representation that is invariant with cycle time. Changing z_{ij} to fractions of a cycle, the objective function becomes

$$\min \sum_{(i,j)} (f_{ij} d_{ij} + Q_{ij}),$$

where $d_{ij} = z_{ij}/w$ is now the average delay per vehicle in seconds. No change is necessary in the variable Q_{ij} that is due to the SDF. A new nonlinearity is introduced by the expression z/w. In order to maintain the MILP formulation, we approximate d by a set of planes forming a convex region and, as before, transfer them into the constraint set. The surface is closely approximated in the domain in which we expect the solution to lie. Only extreme portions of the feasible region are excluded, where we are quite sure a solution is unlikely to occur. The complete MILP formulation is given below.

3.3 Solution by Mixed-Integer Linear Programming

MILP *Find* ϕ_{ij}, r_{ij}, w *to*

$$\min \sum_{(i,j) \in L} (f_{ij} d_{ij} + Q_{ij}),$$

subject to:

$$z_{ij} \geqq z_{ij}^k(\phi_{ij}, r_{ij}, w), \qquad k = 1, \cdots, K \text{(LPF)}$$

$$Q_{ij} \geqq Q_{ij}^h(r_{ij}, w), \qquad h = 1, \cdots, H \text{(SDF)}$$

$$d_{ij} = d_{ij}^m(z_{ij}, w), \qquad m = 1, \cdots, M \text{(representation of } z/w)$$

$$\sum_{(i,j) \in F(l)} \phi_{ij} - \sum_{(i,j) \in R(l)} \phi_{ij} + \sum_{j \in N(l)} \Psi_{j(l)} = n_l, \qquad \forall l \in \mathcal{L}_f,$$

$$r_{\min} \leqq r_{ij} \leqq 1 - f_{ij}/0.95 s_{ij}, \qquad (i, j) \in L,$$

$$w_{\min} \leqq w \leqq w_{\max},$$

$$n_l^1 \leqq n_l \leqq n_l^u, \qquad (n_l \text{ integer})$$

ϕ_{ij} unrestricted in sign.

The test network of Fig. I-9 is now optimized for offsets, splits and cycle time simultaneously. The results are given in Tables III, IV, and V below.

TABLE III

Test Network Results and Statistics

			Time (min)	Functional (veh ×sec/sec)
Rows	248	Continuous Optimum	0.10	40.0006
Columns	121	First Integer Solution	0.24	59.6186
Variables	369	Optimal Integer Solution	0.25	53.8010
Integer Variables	8	Optimality Proved	0.40	
Elements	1027			
Density	1.12			

TABLE IV

Integer Values

Integer Variable	First Integer Solution	Optimum Integer Solution
N1	1	1
N2	1	1
N3	1	1
N4	1	1
N5	−1	−1
N6	−1	−1
N7	0	0
N8	−1	0
Cycle (sec)	65	63.8
Frequency (sec^{-1})	0.01565	0.01569

The optimal solution in this case is a considerable improvement over the solutions obtained previously, because of the decisive role played by the cycle time in the total optimization. By considering all the decision variables simultaneously (objective function = 53.80), network performance was improved 11.6 percent with respect to settings obtained when the decision variables were considered sequentially (objective function = 60.87). To check the quality of the planar approximations to the objective function the test network was re-solved with the optimal value of $w = 0.01569$ sec^{-1} ($C = 63.8$ sec) kept fixed. The linear approximation to the nonlinear objective function is circumvented in this case. The same integer set resulted, with all other variables being very close to the variable-cycle optimal solution. The functional value was 53.1272, i.e., a difference of only 1.25 percent.

Since in practice flows change frequently, the test network was further used to analyze the sensitivity of delay with respect to flows. All flows in the network were changed by the same percentage and the MILP formulation of Section 2.3 was used to optimize offsets and splits for various cycle times. The results are shown in Fig. 5. The decisive role played by the

cycle time in reaching optimum operating conditions is illustrated here even more emphatically than before. Again, similarity with single intersection behavior is apparent.[1-15, 8] A clear conclusion is that it makes good sense to try and adapt cycle length to actual traffic conditions. How to do this in an optimal manner, taking into account the transient phenomena involved in the process, requires further research.

4. CONCLUSIONS

THE STUDY REPORTED here demonstrates the feasibility of using mixed-integer linear programming to optimize traffic signal settings for practical-size road networks. In addition to the test network solutions given in this paper, the method was used for a portion of the Urban Traffic Control

TABLE V

Results of Computations

Link No.	Offset sec (cycle)	Green Time sec (cycle)	Avg. Delay z_{ij} (sec)	SDF Q_{ij} (veh)	Degree of Saturation
101	28.0 (0.44)	17.2 (0.27)	4.84	0	0.674
102	29.3 (0.46)	17.2 (0.27)	2.23	0	0.674
103	38.2 (0.60)	24.2 (0.38)	12.43	0	0.479
104	25.5 (0.40)	21.7 (0.34)	5.99	0.943	0.775
105	32.5 (0.51)	21.7 (0.34)	3.95	0	0.535
106	31.2 (0.49)	24.2 (0.38)	5.74	0	0.693
107	21.0 (0.33)	30.6 (0.48)	3.12	1.660	0.868
108	25.5 (0.40)	30.0 (0.47)	3.00	1.930	0.887
109	44.6 (0.70)	28.7 (0.45)	8.09	1.255	0.778
110	19.1 (0.30)	37.6 (0.59)	2.87	0	0.373
111	33.8 (0.53)	24.9 (0.39)	12.43	2.244	0.897
112	30.0 (0.47)	28.7 (0.45)	4.94	0	0.489
113	33.8 (0.53)	37.6 (0.59)	4.72	0	0.565
114	30.6 (0.48)	33.1 (0.52)	14.60	0	0.641
115	24.9 (0.39)	30.6 (0.48)	10.26	0.690	0.729
116	24.2 (0.38)	24.9 (0.39)	4.14	2.244	0.897
Input Links					
117		37.6 (0.59)	9.43	0	0.593
118		33.1 (0.52)	12.11	0	0.673
119		21.7 (0.34)	18.04	0	0.535
120		24.2 (0.38)	16.12	0	0.479
121		23.6 (0.37)	18.29	2.714	0.901
122		24.9 (0.39)	16.06	0	0.564
123		30.0 (0.47)	15.42	1.930	0.887
124		24.2 (0.38)	17.14	0	0.693

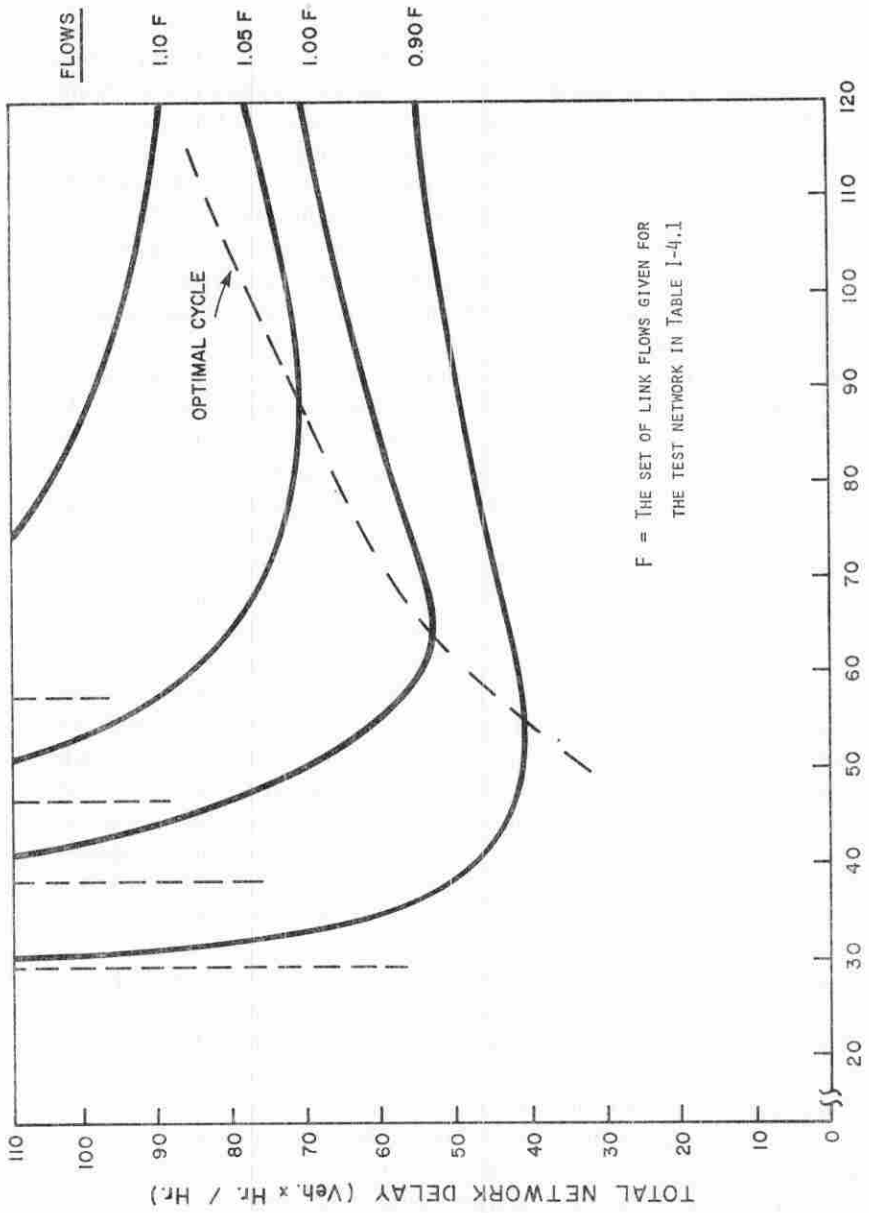

Fig. 5. Sensitivity of network delay with respect to cycle time and flows.

System/Bus Priority System (UTCS/BPS) in Washington, D. C., containing 20 nodes, 63 links, and 21 independent loops, and for an arterial with 11 signals in Waltham, Massachusetts.[1-25] The study also indicates the importance of having all the control variables of the system as simultaneous decision variables. Performance may be significantly degraded when using a sequential decision process. Of particular importance is the cycle time, which is strongly affected by the flows in the network.

The branch-and-bound procedure employed by the MPSX-MIP code[1-11] has proven to be quite efficient in handling the traffic signal network problem. This is largely due to the fact that the proportion of integer variables in the problem is relatively small. The availability of a general-purpose code such as this, which avoids the development of ad hoc branch-and-bound algorithms, has greatly increased the scope of optimization possibilities. It has been our experience that the first set of feasible integer solutions obtained by this method contains either the optimal solution or a solution that is very good. These solutions are obtained relatively fast, computation time varying almost linearly with the number of integer variables. Searching for the optimal solution, or proof of optimality for the best solution obtained initially, requires fathoming all branches of the branch-and-bound tree. This takes longer, the search time varying roughly exponentially with the number of integer variables. Integer programming is an active area of research and it is likely that further improvements in algorithms will occur in the future.[10]

The current study also suggests some directions for further research and development efforts. It would be useful to have a traffic signal network optimization procedure set up in an interactive system. This would enable the traffic engineer to study the performance of the road network and its control system in the design stages; to investigate the sensitivity of the performance with respect to critical parameters (such as capacity at critical intersections); and to evaluate alternative proposals for improvements, or operational changes, in the network (such as widening of approaches, changes in circulation patterns, etc.).

The sensitivity of network performance with respect to flows indicates the potential benefit of traffic responsive systems. Although recent attempts at using such strategies[11,12] have not yet proven to be superior to existing techniques, work is continuing in this area.[13] There seems to be a need for an adequate representation of the transient phenomena in the traffic process that eventually will be tractable by dynamic control techniques. Another promising area is to use traffic control measures to affect route choice of trip makers, so that overall system performance is improved. It is assumed at present that route choice is fixed, resulting in a constant average flow on each link, irrespective of controls imposed on that link. This is not always true. Development of a model that incorporates both

traffic controls and route choice could be used to search for a system-optimized flow distribution in the network.[14-16]

REFERENCES

1. Part I of this paper, "The Network Coordination Problem," *Trans. Sci.* **9,** 321–343 (1975).
2. J. A. Kaplan and L. D. Powers, "Results of SIGOP-TRANSYT Comparison Studies," *Traffic Eng.* **43** (12) (1973).
3. J. Holroyd, "The Practical Implementation of Combination Method and TRANSYT Programs," Transport and Road Res. Lab. Report LR 518, Crowthorne, 1972.
4. N. Gartner and J. D. C. Little, "Generalized Combination Method for Area Traffic Control," *Transportation Research Record No.* 531, 58–69 (1975).
5. W. Jacobs, "A Study of the Averaged Platoon Method for Calculating the Delay Function at Signalized Intersections," B.A.Sc. Thesis (unpublished), Department of Civil Engineering, University of Toronto, 1972.
6. R. Wormleighton, "Queues at a Fixed Time Traffic Signal with Periodic Random Input," *Canadian Opns. Res. Soc. J.* **3,** 129–141 (1965).
7. F. A. Wagner, F. C. Barnes, and D. R. Gerlough, "Improved Criteria for Traffic Signal Systems in Urban Networks," NCHRP Report 124, Highway Research Board, 1971.
8. A. J. Miller, "Settings for Fixed-Cycle Traffic Signals," *Opnal. Res. Q.* **14,** 373–386 (1963).
9. G. F. Newell, "Approximation Methods for Queues with Application to the Fixed-Cycle Traffic Light," *SIAM Rev.* **7,** 223–240 (1965).
10. A. M. Geoffrion and R. E. Marsten, "Integer Programming Algorithms: A Framework and State-of-the-Art Survey," *Management Sci.* **18,** 465–491 (1972).
11. J. Holroyd and D. I. Robertson, "Strategies for Area Traffic Control Systems: Present and Future," Transport and Road Res. Lab. Report LR 569, Crowthorne, 1973.
12. D. W. Ross, et al, "Improved Control Logic for Use with Computer-Controlled Traffic," Final Report-NCHRP Project 3-18 (1), Stanford Research Institute, Menlo Park, California, November 1973.
13. E. B. Lieberman, "Algorithms for Computer Traffic Control: Current State of the Art," Workshop on Computer Traffic Control, University of Minnesota, March 1974.
14. D. Brand, "Substituting Operational Control for Physical Capacity in Urban Transportation," *Transportation* **1** (3), 247–264 (1972).
15. R. E. Allsop, "Some Possibilities for Using Traffic Control to Influence Trip Distribution and Route Choice," *Transportation and Traffic Theory* (D. J. Buckley, ed.), American Elsevier, New York, 1974.
16. N. Gartner, "Area Traffic Control and Network Equilibrium," Proc. Intern. Symp. on Traffic Equilibrium Methods, Université de Montreal, November 1974.

(Received, March 1975; revised, June 1975)