Spillback from an Exit Ramp of an Expressway

D. C. GAZIS, IBM Watson Research Center,
Yorktown Heights, New York

A discussion is given of the problem of control of an oversaturated system comprising an expressway, a highway, and an exit ramp leading from the expressway to the highway. A traffic light is assumed to control the intersection of the exit ramp and the exit highway. When this intersection becomes oversaturated, the queue along the ramp may spill back into the expressway causing a reduction of its throughput. Any improvement in the service rate of the exit ramp can be effected at the expense of causing some additional delay to the traffic on the exit highway. The operation of the traffic light serving the intersection of ramp and highway is determined, which minimizes the delay of the vehicles served by the entire system.

•ONE very common feature of congestion is the progressive deterioration of various sections of a roadway system due to the "spillback" from one section to its neighbors. Spillback is the result of queueing at certain points coupled with the troublesome fact that automobiles have nonzero length, and sometimes appreciably so. Given this reality, spillback could only be avoided by providing ample parking space for the queueing vehicles. In practice, such parking space is limited or even nonexistent. The question arises whether or not judicious management of the inevitable queues might decrease the aggregate delay to the users of the entire system.

In two previous papers (1, 2), examples were given of oversaturated systems in which the aggregate delay could be reduced by an appropriate allocation of the green time of the intersection signals throughout the period of oversaturation. The previous theory (1, 2) is used here for the treatment of the problem of optimization of an oversaturated system involving an expressway, 1, an exit ramp, 2, and the exit highway, 3 (Fig. 1). The intersection of 2 and 3 is controlled by a traffic light, 4, as in the case of an observed real situation. It is assumed that this intersection is oversaturated during a rush period. A queue may then build along the exit ramp, 2. When the length of this queue exceeds the storage capacity of the ramp, it spills back into the expressway. The spillback ties up at least one lane of this expressway. In practice, it ties up probably more than one lane, because drivers desiring to use the exit ramp may drive for a while along the lane next to the right lane and then slow down and try to find an opening into the queue. At the same time, some through traffic is invariably trapped in the right lane and fights its way out, very likely reducing the efficiency of the neighboring lane in this process. In any event, a substantial reduction of the throughput of the expressway is caused which frequently results in queueing along this expressway.

In what follows, this spillback problem is treated as one of optimization of an oversaturated system involving three traffic streams along 1, 2, and 3. The control parameter is the split of the green of the traffic light, 4, the operation of which is to be optimized during the rush period.

Paper sponsored by Committee on Vehicle Characteristics and presented at the 44th Annual Meeting.

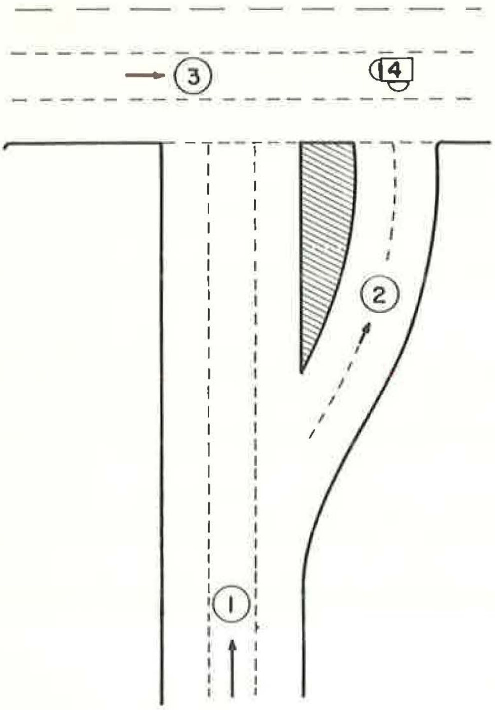

Figure 1. Configuration of the system comprising an expressway, 1, an exit ramp, 2, and an exit highway, 3; the intersection of ramp and highway is controlled by a traffic light, 4.

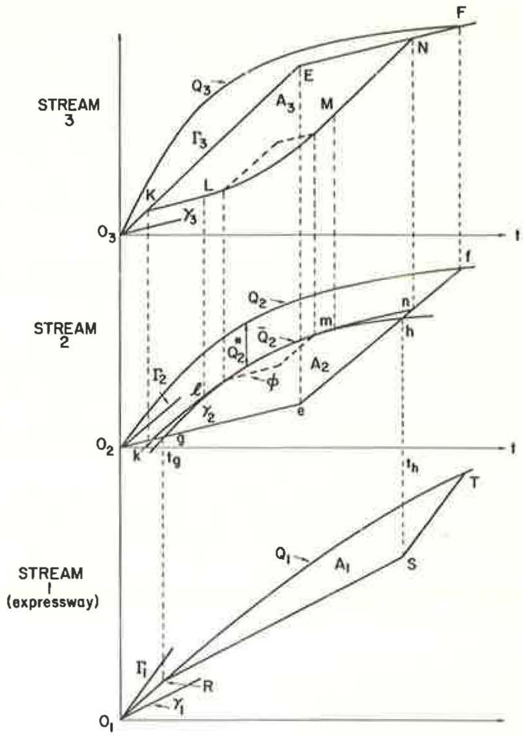

Figure 2. Optimum control of the system when spillback can be avoided altogether.

SOLUTION OF THE OPTIMIZATION PROBLEM

The cumulative demand curves Q_1, Q_2, and Q_3, of the streams 1, 2, and 3 are shown in Figure 2. The maximum and minimum service rates for streams 2 and 3 are determined from the operation of the traffic light 4. Thus, Γ_2, γ_3 correspond to allocation of maximum green to stream 2, whereas γ_2, Γ_3 correspond to maximum green for stream 3. The light cycle is, for the moment, assumed constant. Also shown are two service rates for stream 1, Γ_1 and γ_1. The former is obtainable when the expressway is unobstructed, and the latter is the reduced expressway throughput in case of spillback. It is assumed that Q_1 can be adequately served by the normal service rate Γ_1.

More often than not the saturation flows s_2 and s_3 are such that

$$s_2 < s_3 \tag{1}$$

According to the theory (1), the optimum operation of light 4 alone would be a two-stage operation involving the service curves O_3 EF and O_2 ef, with the highway stream 3 receiving preferential treatment. However, in the present case one must take into account that the intersection 4 is not isolated, and a large enough size of the queue of stream 2 will cause additional delays on the expressway. Let us draw the curve

$$\overline{Q_2} = Q_2 - Q_2^* \tag{2}$$

where Q_2^* is the maximum acceptable queue which does not cause spillback. Let the curve $\overline{Q}_2$ intersect the service curve O_2 ef at points g and h. This means that if one accepts the service curve O_2 ef for stream 2, he will cause spillback during the time

$t_g \le t \le t_h$. The result will be a delay for stream 1 proportional to the area between Q_1 and the service curve RST, which will be denoted by A_1.

Assuming now that we operate the light 4 so that the queue along 2 remains smaller than or equal to Q_2^* at all times, any such service curve of stream 2 must be between the curves Q_2 and $O_2k\ell mnf$. The portions $k\ell$ and mn of the latter curve are tangent to the curve $\bar{Q}_2$ and correspond to service rates Γ_2 and γ_2, respectively. The curve $O_2k\ell mnf$ corresponds to the minimum possible service rate of the stream 2 which prevents spillback, with full utilization of the green light in both directions 2 and 3. The complementary service curve of stream 3 is O_3KLMNF. Choosing these two service curves rather than the curves O_3EF and O_2ef involves an increase of the aggregate delay at intersection 4 equal to the difference between the areas A_2 and A_3 which are contained between the pairs of curves (KEN, KLMN) and (ken, $k\ell$mn), respectively. It may be seen that any trade-off of delay between streams 2 and 3 involves quantities proportional to the saturation flows s_2 and s_3. This is so because the trade-off is accomplished by taking green time from stream 3 and giving it to stream 2. The utilization rate of this green time is then reduced from s_3 to s_2 cars per second of green. Accordingly, the ratio A_2/A_3 is given by

$$A_2/A_3 = s_2/s_3 \tag{3}$$

The total change in the aggregate delay of all three streams is given by

$$\delta = A_1 + A_2 - A_3 \tag{4}$$

or, in view of Eq. 3,

$$\delta = A_1 + A_2\left(1 - \frac{s_3}{s_2}\right) \tag{5}$$

A net reduction of delay results if δ is positive. In this case it pays to adopt the strategy of keeping the queue along the ramp below the critical value Q_2^*. If δ is negative, then spillback is not as damaging as it appears, at least in terms of total delay, which is minimized by an optimum operation of the traffic light 4, assumed isolated (1). However, it may still be desirable to prevent congestion on the expressway for safety reasons which may override delay considerations. If this is the case, one may accept a small negative delay trade-off, δ.

Assuming that the delay criterion is the dominant one, we find that a critical constant rate, q_1, of demand along the expressway exists, which is related to A_2 according to a relationship obtained by setting δ in Eq. 5 equal to zero. Thus,

$$\frac{(\Gamma_1 - \gamma_1)(q_1 - \gamma_1)}{\Gamma_1 - q_1}\,\frac{\tau^2}{2} = A_2\left(\frac{s_3}{s_2} - 1\right) \tag{6}$$

where the left-hand side of Eq. 6 is equal to A_1, and

$$\tau = t_h - t_g \tag{7}$$

Solving Eq. 6 for q_1,

$$q_1 = \frac{2A_2(s_3/s_2 - 1)\Gamma_1 + \gamma_1(\Gamma_1 - \gamma_1)\tau^2}{2A_2(s_3/s_2 - 1) + (\Gamma_1 - \gamma_1)\tau^2} \tag{8}$$

If the demand rate is smaller than q_1, then spillback is the lesser of two evils, since it corresponds to minimum total delay.

By similar arguments we may investigate the possibility of allowing spillback during a portion of the interval $(t_h - t_g)$. If the rate of demand along the expressway falls sufficiently below Γ_1, it may be profitable to adopt a strategy such as that corresponding to the dashed line ϕ in the middle diagram of Figure 2, and the complementary service curves for streams 1 and 3 (the last one not shown in Fig. 2). This policy will introduce some additional delay to the stream 2, and because of spillback it will also delay some vehicles in the stream 1. It will reduce somewhat the average delay to stream 3. The net change can be computed by an expression similar to Eq. 5, if the exact shapes of Q_2 and Q_1 are known. Finally, if the demand rate along the expressway falls below γ_1, it is always profitable to allow spillback.

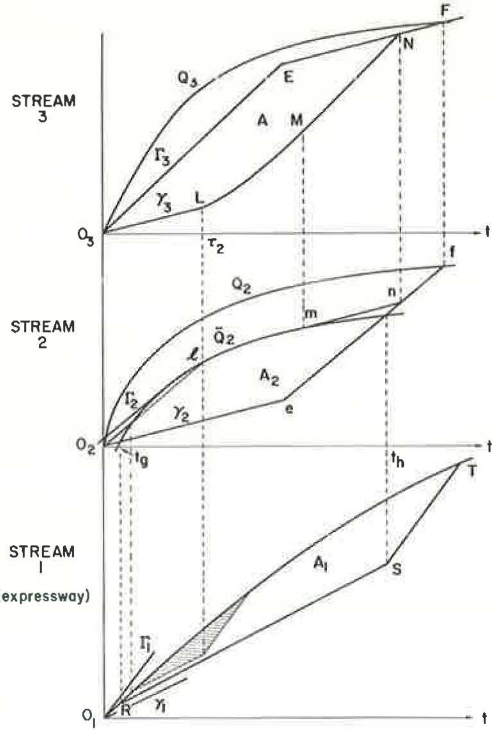

Figure 3. Optimum control of the system when spillback is unavoidable for at least a short period between the onset of oversaturation and the time τ_2.

So far, we have discussed only the case when it is possible to prevent spillback during the entire rush period, if so desired. This is not the case if the tangent to the curve $\bar{Q}_2$ with slope equal to Γ_2 intersects the abscissa axis to the left of O_2, as shown in Figure 3. In this case spillback is inevitable, due to a very fast rise in demand along the exit ramp. The best one can do is maintain maximum service for the stream 2 until the spillback is eliminated at time τ_2. This alternative is to be compared with that corresponding to the service curves O_3EF, O_2ef, and O_1RST. The net change in total delay, δ, is again given by Eq. 5, where A_1 and A_2 now denote the total delay to streams 1 and 2 minus the inevitable one shown by the shaded areas of Figure 3. If δ is positive, then spillback should be prevented after τ_2.

The preceding discussion has certain similarities with the examples of References 1 and 2 and certain differences. As in those examples, the solution given is a deterministic one depending on the demand during the entire rush hour rather than the instantaneous sizes of the queues. Also, the need for anticipating the critical behavior of queues, on the basis of available data regarding recurrent demands, is shown in Figures 2 and 3. Thus, if spillback is to be avoided, one must sometimes act before the queue along the exit ramp attains the critical size Q_2^*. Thus, the optimum strategy calls for maximum service of this queue starting at t_k (Fig. 2) and at 0, i. e., the onset of oversaturation (Fig. 3).

One special feature of the present problem is that the size of the queue along the exit ramp affects the service rate of the expressway 1. This was not the case in the problems of References 1 and 2 where it was pointed out that the asymptotic behavior of the demand curves, near the end of the rush period, might be sufficient for determining the optimum operation of the traffic lights. In the present problem, however, the exact shape of $Q_2(t)$ is needed. Moreover, the solution is more sensitive to fluctuations of demand along 2. In any case, the discussion given can be used as a guide for designing an adaptive control system which takes into account fluctuations of demand. For example, if spillback is to be avoided, the system must keep the size of queue along 2 below the critical size Q_2^* at all times.

VALUE OF PARKING SPACE

Let us try to get a gross estimate of the value of an increase of the parking space along the exit ramp, or equivalently of the critical queue size Q_2^*. An increase of Q_2^* by one car permits a reduction of the area A_2 (Fig. 2) by

$$\delta_2 \approx (t_h - t_g) \tag{9}$$

Assuming that spillback is to be avoided, the increase in Q_2^* will result in total reduction of delay for streams 2 and 3 equal to

$$\delta_{23} = (t_h - t_g)\left(\frac{s_3}{s_2} - 1\right) \tag{10}$$

Consider an example: assume $s_3/s_2 = 2$, $t_h - t_g = 1$ hour, and a cost of delay equal to $1.50 per hour. Also, assume that oversaturation occurs approximately once a day or, say 300 times per year. During a 30-yr amortization period, the increase in capacity by one car will result in a reduction of delay which may be valued at $13,500. A highway planner may take such an estimate into account in deciding whether to increase the capacity of an exit ramp or not. It should be pointed out, however, that the return per unit increase of Q_2^* diminishes as one approaches the maximum vertical distance between the curve Q_2 and the line O_2ef (Fig. 2). The decrease in rate of return is equal to the decrease of $(t_h - t_g)$, or roughly linear.

Incidentally, a substantial decrease in delay may also be accomplished by a drastic reduction of the average length of the automobile. This will be the case, provided that the overall performance of the automobiles is not affected by the reduction of their size, a conjecture which will be easily refuted by Detroit.

OPTIMUM LIGHT CYCLE

Up to this point, the light cycle at intersection 4 has been assumed constant. It should be interesting to find the value of the light cycle which optimizes the overall performance of the system in terms of delay. The light cycle influences delay in the following ways:

1. A long cycle decreases the delay by increasing the utilization rate of the cycle, assuming that the lost time due to acceleration and clearance is essentially independent of the light cycle length.
2. A long cycle increases the additional per cycle delays due to intermittent service. These delays are proportional to the sawtooth areas of Figure 4.
3. A long cycle decreases the effective parking capacity, Q_2^*, of the exit ramp. This is so because Q_2^* is the actual capacity of the exit ramp minus the extra queue length built during the red phase of the cycle along 2. A decrease of Q_2^* causes additional delays as seen in the preceding section.

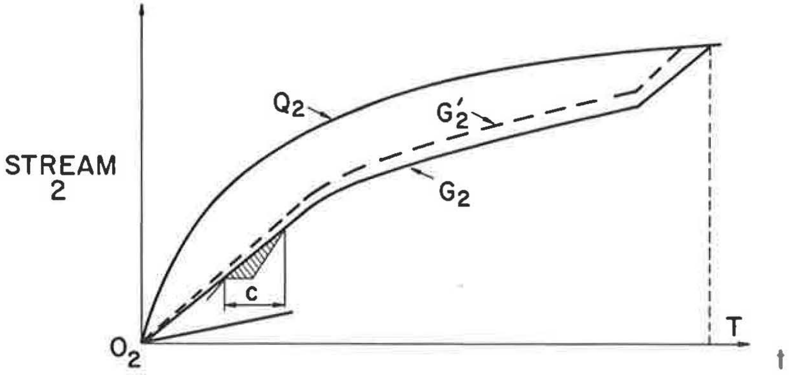

Figure 4. Influence of a variation of the light cycle on the aggregate delay of stream 2.

The optimum cycle is such that small variations about its value produce essentially zero variation of delay due to the three previous factors. An estimate of the optimum light cycle is obtained as follows:

Let the cycle be denoted by c, the total lost time per cycle by L, and the percentage of effective green allocated to direction 2 by p(t), where t is the time. The service rates along 2 and 3 are

$$\bar{\gamma}_2(t) = (1 - L/c)\, s_2\, p(t) \tag{11a}$$

$$\bar{\gamma}_3(t) = (1 - L/c)\, s_3[1 - p(t)] \tag{11b}$$

The decrease of delay due to an increase of the light cycle, Δc, is approximately equal to

$$(\Delta D)_a = \Delta c \left\{ \frac{LT}{2c^2} \int_0^T [\, s_2 p + s_3 (1 - p)\,]\, dt \right\} \tag{12}$$

where T is the duration of the rush period. It is assumed that p(t) and T are essentially unaffected by a small change Δc. An approximate value of the integral in Eq. 12 is $(s_2 + s_3)T/2$, assuming a more or less symmetric distribution of the values of p(t) about the value $1/2$, during T. Using this approximation,

$$(\Delta D)_a \approx \frac{LT^2}{4c^2} (s_2 + s_3)\, \Delta c \tag{13}$$

The delay corresponding to the sawtooth area of Figure 4 is

$$a_2 = \frac{1}{2}\, p(1 - p)\, s_2\, (c - L)\, c \tag{14a}$$

per cycle, for stream 2, and

$$a_3 = \frac{1}{2}\, p(1 - p)\, s_3\, (c - L)\, c \tag{14b}$$

per cycle, for stream 3. An increase in c produces an increase of this delay equal to

$$(\Delta D)_2 = \Delta c \left\{ \frac{s_2 + s_3}{2} \int_0^T p(1 - p)\, dt \right\} \tag{15}$$

We need an estimate for the integral of Eq. 15. The integrand is equal to 0.25, for p = 0.5 and 0.09 for p = 0.1. In view of the fact that the optimum control calls for extreme values of p, we shall assume (a better estimate may be obtained if one has, from a trial solution, a good approximation for the function p(t)) for the integral, the value 0.12 T, in which case

$$(\Delta D)_b = 0.06\, (s_2 + s_3)\, T \Delta c \tag{16}$$

Finally, the decrease in Q_2^* due to an increase dc is equal to

$$\Delta Q_2^* = \Delta c\, [\, p(1 - p)\, s_2\,] \tag{17}$$

Hence, according to the preceding section, assuming

$$t_h - t_g \approx T \tag{18}$$

we find an increase in total delay

$$(\Delta D)_c = \Delta c \left\{ (s_3 - s_2) \int_0^T p(1 - p)\, dt \right\} \approx \Delta c \left[0.12 (s_3 - s_2)\, T \right] \tag{19}$$

Now setting

$$(\Delta D)_b + (\Delta D)_c - (\Delta D)_a = 0 \tag{20}$$

we obtain an equation for c, namely,

$$- \frac{LT^2}{4c^2} (s_2 + s_3) + (0.06)\, (s_2 + s_3)\, T + (0.12)\, (s_3 - s_2)\, T = 0 \tag{21}$$

which yields

$$c \approx \left[\frac{LT\, (s_2 + s_3)}{0.24 (3 s_3 - s_2)} \right]^{1/2} \approx 0.2 \left[LT \frac{\rho + 1}{3\rho - 1} \right]^{1/2} \tag{22}$$

where $\rho = s_3/s_2$. For example, assuming $\rho = 2$, T = 1 hour, and L = 4 sec, we find $c \approx 3$ min.

It will be noted that although the exact value of the coefficient multiplying the square root in Eq. 22 will vary after a more accurate computation of the integrals of Eqs. 12 and 15, the dependence of c on the square root of T and L remains as an intrinsic feature of the present theory. It should be remarked that the capacity of the exit ramp imposes an upper limit on the light cycle which may be of primary importance in the case of a very short ramp. Thus, the queue buildup during red must not exceed the actual ramp capacity. If this capacity is Q_{max}, then

$$Q_{max} \geq p(1 - p)\, s_2\, (c - L) \tag{23}$$

Hence,

$$c \leq L + \frac{Q_{max}}{s_2\, p(1 - p)} \tag{24}$$

For example, assuming p = 0.2, Q_{max} = 10 cars, s_2 = 0.3 cars/sec, and L = 4 sec, we find that c must be at most 3 ½ minutes.

The preceding discussion assumes, of course, that the queue 2 can be served critically by an appropriate choice of p.

CONCLUDING REMARKS

The discussion is based on the assumption that the system comprising the expressway, the ramp, and the highway is isolated from other oversaturated regions. If this is not the case, one must consider an enlarged oversaturated system which can be considered isolated. For example, the stream 3 along the highway may contain a large amount of traffic coming from an exit ramp of the direction of the expressway opposite to direction 1. In this case, both exit ramps may be likely to produce spillback if the expressway is heavily traveled both ways. One general rule, in such a case, is that queueing can be permitted where there is greater parking capacity. A more detailed investigation is needed to determine where spillback, if inevitable, must be allowed in order to minimize the delay in the entire system.

Perhaps an explanation is due regarding the meaning of the demand curve $Q_2(t)$ used in this paper. This curve represents all the cars which would have demanded ramp service at time t, had the approach to the ramp been completely open. In practice, if spillback takes place, a large number of these cars will be mingled with through traffic in the queue formed along the expressway. Therefore, care must be exercised in ascertaining the appropriate value of $Q_2(t)$ by observing the composition and length of the queue along the expressway.

REFERENCES

1. Gazis, D. C., and Potts, R. B. The Oversaturated Intersection. Proc., Second International Symposium on Traffic Theory (to be published).
2. Gazis, D. C. Optimum Control of a System of Oversaturated Intersections. Operations Research, 12, 815-831 (1964).