ABSTRACT
A simple model is described which makes possible approximate calculation of delays encountered by calls served by two groups of interacting servers. The particular interaction studied is typical of control circuits in switching systems. Methods are given for finding saturation load and for determining the size and holding time of a single server group which has approximately the same delay characteristics.

1. INTRODUCTION
We are concerned here with a type of queuing problem which often arises in the analysis of the control systems of common-control switching systems. These control systems have usually been considered as groups of servers (stages) which serve calls sequentially. Exact models of sequential service systems are generally too complicated to lend themselves to mathematical analysis, even of an approximate nature, except in a few cases as, for example, the model treated by Burke. The models which can be analyzed exactly involve (infinite) waiting lines between stages and hence are too divergent from the physical reality of control systems to be useful in studying their capacity. Thus we seek an approximate model which is close enough to reality and allows sufficiently accurate mathematical analysis for us to have confidence in the applicability of the results.

Lee has suggested an approach to the problem indicated above which is based on a modified two-stage tandem queue with feedback. The present approach is quite different from the traditional tandem queuing model or from that of Lee in that the calls are not processed sequentially but in parallel. It appears to be applicable to a broad class of control systems in both electromechanical and electronic switching systems.

The particular problem which suggested the present model is the calculation of the distribution of dial tone delays given by dial tone markers and originating registers in the No. 5 Crossbar Switching System. Thus, the model, although it may find broader applications, will be discussed in terms of markers and registers. This discussion will employ the customary terminology of the telephone traffic literature.

2. THE MODEL - PARALLEL SERVER GROUPS
We consider calls arriving with exponentially distributed, independent, interarrival intervals. To be served, a call must find both a marker and a register idle; otherwise it waits in a queue. Once a marker and register have started processing a call they act independently of each other. Holding times are exponentially distributed with different means for each group. (Registers normally have holding times much longer than marker holding times).

Let $M =$ Number of Markers
$m =$ Number of markers busy at any instant
$R =$ Number of Registers
$r =$ Number of registers busy at any instant
$h_m =$ average holding time of marker
$h_r =$ average holding time of register
$s = h_r/h_m$

Basic to our analysis is the load that can be handled by the system under conditions of steady-state saturation. This is the situation where more calls are offered than can be processed and where a queue has existed for a long time. When saturation exists we can identify three types of
states of the servers:

1. \( m = M \) and \( r = R \)
   - The rate of call acceptance by the servers is zero in this state.

2. \( m < M \) and \( r = R \)
   - Calls are accepted at the rate at which registers become idle. Letting \( h_r \) equal the holding time of a register, this rate is \( R/h_r \).

3. \( m = M \) and \( r < R \)
   - Calls are accepted at the rate at which markers become idle. Letting \( h_m \) equal the holding time of a marker, this rate is \( M/h_m \).

The states of the model and the transition rates between adjacent states can now be described as in Figure 1. Letting \( f(m, r) \) be the probability of state \((m, r)\), we have,

\[
\frac{r}{h_r} f(M, r) = \left( \frac{M}{h_m} \right) f(M, r-1)
\]

or

\[
f(M, r) = M f(M, r-1)
\]

and similarly,

\[
m f(m, R) = \left( \frac{R}{s} \right) f(m-1, R)
\]

The states from \((0, R)\) to \((M, R)\), or from \((M, 0)\) to \((M, R)\), have transitions which are identical to those of an Erlang loss model. It is simple to express all state probabilities in terms of \( f(M, R) \) and then set the sum equal to unity. The state probabilities are conveniently expressed in terms of:

\[
B_m = \sum_{x=0}^{M} \frac{(R/s)^x}{x!} / \left( \frac{R}{s} \right)^M / M! = B(M, R/s)^{-1}
\]

\[
B_R = \sum_{x=0}^{R} \frac{(Ms)^x}{x!} / \left( \frac{Ms}{R} \right)^R / R! = B(R, Ms)^{-1}
\]

\[
D = B_m + B_R - 1
\]

where \( B(c, a) \) is the Erlang loss formula with \( a \), the offered load and \( c \), the number of servers. (Tables of values are generally available.)

All state probabilities can now be computed. In particular, the probability that all markers and all registers are busy is

\[
f(M, R) = D^{-1}
\]

The means are:

For registers \( E(r) = R(B_{R-1} + B_{M-1})/D \)

For markers \( E(m) = M(B_{R} + B_{M-1})/D \)

Each call is handled in both groups so that the group loads must be in the same ratio as the holding times, that is:

\[
E(m) = E(r)/s
\]

Thus, because \( B_{R-1} < B_R \) and \( B_{M-1} < B_M \), the capacity of interacting groups of this type under a load large enough to produce saturation is always less than that of either group alone. Some results are shown graphically in Figure 2 for one marker and numbers of registers as a function of \( s \).

It is of interest to look at what we will call the maximum system efficiency that can be reached in a marker-register system. Maximum system efficiency occurs when the occupancies of the two groups are equal, that is when \( s = R/M \). At any other value of \( s \), one of the two groups will be below maximum system efficiency. A plot for up to 5 markers and 120 registers is shown in Figure 3.
Figure 4 shows how the efficiency of the lower loaded group drops as the ratio \( M_s/R \) departs from unity. This curve is only approximate; each combination of \( M, R \) and \( s \) produces slight deviations from the curves. It is likely that the curve is sufficiently accurate for practical purposes given the usual difficulties of insuring that the proper holding times are used.

Fig. 4. Effect on Efficiency of Lesser Loaded Group of Departure from Maximum System Efficiency Loading.

3. DELAY APPROXIMATION FOR TWO PARALLEL INTERACTING GROUPS

State equations describing the behavior of the queue in the model described above are not easily solved at loads below saturation. To get some idea of the delay characteristics, a simple, single server group equivalent of the interacting groups is proposed as an approximation to the model described above. The equivalent number of servers in the approximation is determined by the following heuristic considerations. In saturation when all registers are busy (and markers are idle), calls are served as they would be by a single group of \( R \) registers. When all markers are busy (and registers are idle), they are served as they would be by \( M \) markers. When both markers and registers are all busy, we arbitrarily assume the number of servers to be \( RM/(R+M) \). Taking the weighted average we get an equivalent number of servers:

\[
c_e = \frac{R(B_R-1) + M(B_R-1) + RM/(R+M)}{D}
\]

An equivalent holding time \( h_e \), is that which results in the same number of calls being carried at full occupancy of the equivalent group as are carried at saturation load in the original model, that is:

\[
h_e = \frac{c_e}{E(r)} h_r
\]

This approximation has the property of being correct at the extremes of ratios of markers to registers. In the case where markers are the limiting factor in call carrying capacity and registers are practically never all busy, \( c_e \) will equal \( M \) and \( h_e \) will equal \( h_m \). Similarly the model is accurate when registers alone limit the call capacity.

Figures 5 and 6 show plots at maximum system efficiency of equivalent single group size and equivalent holding time for up to 6 markers and a ratio of holding times up to 40. The reason for the small difference in equivalent holding time with different marker group sizes is not evident.

Fig. 5. Single Group Equivalent to Interacting Marker and Register Groups - Group Size at Maximum System Efficiency.

If the equivalent holding time were exponentially distributed, delay distributions would be derived by standard methods. This is not the actual case, although the assumption of exponential distribution appears conservative. The ratio of the second moment of the intervals between call acceptances to the exponential second moment can be used as a factor to reduce the delays calculated under the exponential assumption. (A method of estimating the second moment of the interservice time is given in the appendix.) A comparison between exact calculation from state equations and the approximate method is given in Table I for 2 registers and 1 marker and for 6 registers and 2 markers. It is seen that the probability of delay is well estimated but the delay distribution needs further correction. Fortunately, at more realistic group sizes the main contributor to changing the distribution shape, the state of all markers and registers busy, is less prominent than in the example. As a result a service time distribution correction for this cause seems normally unnecessary.
TABLE I

Comparison of Approximation with Numerical Calculation
(M=1 Marker, R=2 Registers, s=2, c_e=1.2 Servers, Max. Eff. = .714)

<table>
<thead>
<tr>
<th>Marker Load</th>
<th>Probability of Delay Calculation</th>
<th>Approximation</th>
<th>Average Number of Calls Waiting Calculation</th>
<th>Approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1</td>
<td>.108</td>
<td>.108</td>
<td>.0128</td>
<td>.0135</td>
</tr>
<tr>
<td>.2</td>
<td>.231</td>
<td>.235</td>
<td>.0663</td>
<td>.0700</td>
</tr>
<tr>
<td>.3</td>
<td>.365</td>
<td>.370</td>
<td>.199</td>
<td>.198</td>
</tr>
<tr>
<td>.4</td>
<td>.508</td>
<td>.52</td>
<td>.490</td>
<td>.51</td>
</tr>
<tr>
<td>.5</td>
<td>.658</td>
<td>.67</td>
<td>.67</td>
<td>.67</td>
</tr>
<tr>
<td>.6</td>
<td>.814</td>
<td>.82</td>
<td>.317</td>
<td>.32</td>
</tr>
</tbody>
</table>

(M=2 Markers, R=6 Registers, s=3, c_e=2.91, Max. Eff. = .81)

<table>
<thead>
<tr>
<th>Marker Load</th>
<th>Probability of Delay Calculation</th>
<th>Approximation</th>
<th>Average Number of Calls Waiting Calculation</th>
<th>Approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>.7</td>
<td>.193</td>
<td>.19</td>
<td>.110</td>
<td>.116</td>
</tr>
<tr>
<td>.9</td>
<td>.313</td>
<td>.32</td>
<td>.292</td>
<td>.337</td>
</tr>
<tr>
<td>1.1</td>
<td>.465</td>
<td>.47</td>
<td>.739</td>
<td>.872</td>
</tr>
</tbody>
</table>

4. ACCOUNTING FOR "WASTE USES"

In practice in No. 5 Crossbar, a call which finds that all registers are busy and one or more markers are idle cause a marker to go through a cycle, usually less than a normal holding time, during which it cannot be seized by a new call. The effect of this "waste" marker use is to lower the saturation load of the system, although it should be observed that many waste uses occur when the marker could not have been performing any useful work because all registers were busy.

To introduce this feature directly into the model appears to take it beyond the range of simple analysis. Instead, we introduce an approximation that the waste usage is added to register load instead of to the marker load. This is roughly equivalent to assuming that register releases during all-register-busy periods are delayed by marker waste uses but that this delay affects only the registers, adding to the work time of the registers. As long as such time is small it does not appear to introduce material error when included in this fashion. The size of the correction can be derived by finding from the saturation load model the proportion of register seizures affected by waste usage.

From the state probabilities of the model without waste uses, the average number, B, of markers idle (hence in waste usage operation) when a register releases, given that at least one marker is idle, is M \((B_\text{y}/B_\text{y}-1)-(\lambda/\mu)\). The proportion, \(P\), of register seizures occurring when at least one is idle is \(R(B_\text{y}-1)/(R(B_\text{y}-1)+M_\mu(B_\text{y}-1))\). Neglecting queueing for idle markers, we get the average time added to register work time as \(h_\text{w}P\). Plots of the average proportion of a marker holding time which should be added to register holding time at maximum system efficiency show remarkably little variation with changing number of markers. Accordingly only the curve for one marker (which turns out to be the familiar Erlang loss formula) is drawn in Figure 7.

5. INTERACTION OF ONE GROUP WITH MANY

We can now expand our model to include cases in which one group interacts with several others which are independent of each other. Using our marker-register analogy, this is the case of two different groups of registers, TOUCH-TONE and dial pulse, which work with a single group of markers. We let \(p_d\) equal the proportion of calls destined for dial pulse registers and \(p_t\) the proportion destined for TOUCH-TONE registers, where each register group can be considered separately. Looking at a single marker we note that if a marker uses a full holding time on each seizure the average time between associations with a particular group is the average marker holding time divided by the proportion of calls to that group. If we can assume a waste use equal to \(h_\text{w}\), average time is unaffected by the all busy state of the register groups and we can define an effective marker holding time, \(h = h_\text{w}/P_d\) for TOUCH-TONE calls, and \(h = h_\text{w}/P_t\) for dial pulse calls. Equivalent waste-use holding time will be equal to equivalent marker holding time. The model is then equivalent to two independent models of the form described in the previous section with waste marker uses and lengthened marker holding times.

A number of cases were computed numerically from the state equation to check out the approximation at saturation load. In particular there is concern that the groups of markers are not quite independent. Some results are given in Table II where it is seen that the approximation is best when registers are the prime source of congestion. For the cases studied, there is a tendency for the saturation load to be underestimated.
TABLE II

<table>
<thead>
<tr>
<th>Proportion Dial Pulse</th>
<th>Number TT Registers</th>
<th>Estimated Max. Load</th>
<th>Computed Max. Load</th>
<th>Number DP Registers</th>
<th>Estimated Max. Load</th>
<th>Computed Max. Load</th>
</tr>
</thead>
<tbody>
<tr>
<td>.75</td>
<td>18</td>
<td>16.4</td>
<td>16.4</td>
<td>91</td>
<td>96.8</td>
<td>87.2</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>18.5</td>
<td>18.7</td>
<td>101</td>
<td>94.3</td>
<td>94.7</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>19.3</td>
<td>20.0</td>
<td>111</td>
<td>100.0</td>
<td>101.5</td>
</tr>
<tr>
<td>.50</td>
<td>43</td>
<td>40.1</td>
<td>40.0</td>
<td>61</td>
<td>58.0</td>
<td>57.7</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>43.3</td>
<td>43.5</td>
<td>68</td>
<td>62.3</td>
<td>63.1</td>
</tr>
<tr>
<td></td>
<td>53</td>
<td>45.5</td>
<td>47.6</td>
<td>75</td>
<td>65.2</td>
<td>68.4</td>
</tr>
<tr>
<td>.25</td>
<td>66</td>
<td>62.3</td>
<td>62.3</td>
<td>19</td>
<td>18.7</td>
<td>18.1</td>
</tr>
<tr>
<td></td>
<td>73</td>
<td>66.5</td>
<td>67.3</td>
<td>21</td>
<td>20.7</td>
<td>19.8</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>70.0</td>
<td>71.5</td>
<td>23</td>
<td>22.0</td>
<td>21.5</td>
</tr>
</tbody>
</table>

Two Register Groups with One Marker Group: Comparison of Approximate Maximum Load Calculations with Numerical Calculation

3 Markers
h_M = .3 (Waste use hold time = .2)
h_RTT = 9.5
h_RDP = 13.5

6. VARIATIONS IN HOLDING TIME DISTRIBUTIONS

In comparing the above model with simulations it becomes apparent that the assumption of exponential holding time distribution is not very good when the marker holding time is constant and only one register group is involved. It is hypothesized that the exponential distribution is reasonable when more than one group is involved because the equivalent marker holding time is made up of a variable number of work times. To handle the case of single register groups and constant holding time, one can go back to the procedure used earlier to find an equivalent single server model. There, proportions of time for M servers, for R servers and for RM/(R+M) servers were found. We propose that one can apply the same weightings to the second moment of the holding times. With constant holding time this is \( l \cdot h^2 \) and with exponential holding time it is \( 2 \cdot h^2 \). Letting \( K \cdot h^2 \) be the second moment of the holding time, at high occupancies, probability of delay \( d \) can be found approximately by adjusting delays from exponential holding time delay curves by \( K \cdot h^2 \).

7. CONCLUSION

A model has been proposed which approximately describes the delays given to calls served by a variety of configurations involving two groups of servers. The approximation consists of two parts: Calculation of saturation load and calculation of delays by determination of an equivalent single server group.

Delay data from overloaded offices and from simulations have been compared with calculations based on the above approximations. Preliminary results indicate fair agreement suggesting that a practical method of predicting delay has been found for marker-registers groups.

References:

APPENDIX

Estimation of Second Moment of Interval Between Successive Call Acceptances in Marker-Register Model at Steady State Saturation

After a call is accepted for service, see Figure 1 of main text, four separate, mutually exclusive series of events may occur before the next acceptance:

1. All registers are busy and markers are idle. Other markers may release before a register.
2. All servers are busy and one or more markers release before a register.
3. All markers are busy and registers are idle. Other registers may release before a marker.
4. All servers are busy, and one or more registers release before a marker.

The probabilities of these series of events, their durations and moments are given below with:

\[ h_r = \text{register holding time} \]
\[ h_d = \text{average delay to first server} \]
\[ h_d = h_r/(R+Ms) \]
\[ R = \text{number of registers} \]
\[ M = \text{number of markers} \]
\[ s = \text{ratio of register to marker holding time} \]

325/5
The average interval and its second moment are obtained by taking the sum of the respective quantities weighted by the corresponding proportions.