NEW OVERLOAD ISSUES IN A DIVESTED ENVIRONMENT

L. J. Forys
Bell Communications Research
Red Bank, New Jersey, U.S.A.

ABSTRACT

The new economic environment resulting from divestiture has put continued emphasis on the Operating Telephone Companies (OTCs) to reduce capital expenditure and efficiently utilize existing equipment. Of particular importance is real time capacity. We examine the capacity of local exchanges which have incorporated the new overload strategy described by the author at the 10th ITC using an approximate analytic model, simulation and field data. Particular attention is paid to the effects of nonstationary traffic inputs. An improvement of about 5% is reported. Other issues examined by the paper deal with the impact that service problems experienced by one Interexchange Carrier can affect other traffic. An analytic model is developed to quantify the impact in special cases and the need for selective overload strategies is indicated.

1. INTRODUCTION

The divestiture of the "Bell System" on January 1, 1984 has of course had enormous consequences in a number of areas. The new economic environment, for example, has resulted in increased stress on capital expenditure and efficient use of resources by the Operating Telephone Companies (OTCs). This has encouraged continued examination of the capacities of a number of traffic sensitive components, in particular, the real time capacities of local exchanges. At the 10th ITC we presented results (see ref. [1]) for a newly implemented overload control for local exchanges. At that time we alluded to a possible increase in capacity for No. 1 ESS* local exchanges. In this paper we will examine this issue more closely by presenting a model which accounts for nonstationary traffic variations. This model incorporates the thinking presented at the 9th ITC (see ref. [2]) and is applicable not only to the analysis of No. 1 ESS* exchanges, but to a wide class of local switching exchanges. Our analysis shows that a substantial capacity increase is possible by incorporating the new overload strategy. It indicates that the new strategy is much more robust to traffic variations than previous strategies.

The model results are verified both by simulation and recently analyzed field data. The agreement between theory and practice is shown to be excellent. Although designed originally for real time overload protection, the new strategy will also be invaluable for handling hardware overloads such as digit receivers, where the same considerations prevail.

Another important impact of divestiture on overload controls has been the introduction of "equal access" into SPC exchanges. Because of equal access, a number of competing Interexchange Carriers (ICs) will be using common resources at both local exchanges and access tandems. Although the recently introduced overload strategy handles excessive real time loads (as well as receiver overloads) in an efficient manner, it may not provide sufficiently selective response to causes. In this paper we will discuss a number of overload scenarios caused by service difficulties experienced by a given IC and demonstrate how these service difficulties can impact the service of other ICs. This motivates the need for more selective overload controls which identify the cause of the service impairments and takes appropriate actions to control the cause without impacting other traffic.

The paper is organized as follows. In Section 2 we begin by reviewing the improved overload strategy and how it differs from previous controls. In Section 3 the issue of nonstationary traffic and its impact on real time capacity is examined, and previous results summarized. An approximate analytic model is described in Section 4 which determines "worst case" traffic patterns for the improved strategy. The results of the model are verified and extended using simulations. This is presented in Section 5. Finally, in Section 6 recent field results are analyzed which confirm our predictions and indicate substantial (about 5%) real time capacity increases can be realized.

The need for selective overload controls to mitigate the impact of IC specific service problems is discussed in Section 7 and the results of possible scenarios quantified in 8. Finally, in section 9 we discuss how network management type controls which are present in many equal access exchanges can be used to mitigate the problem.
2. REVIEW OF OVERLOAD STRATEGIES

Until fairly recently, new originations for service were handled in a first-in-first-out (FIFO) manner or random order of service (sometimes a combination of both) by nearly all local exchanges. Unfortunately, these methods of servicing do not properly account for customer behavior patterns and can lead to significant reduction in the system's ability to properly complete calls. At the 10th ITC, Berard et al. [3] presented the results of a field study which investigated the behavior of originating customers when subjected to dial tone delays. The results show that as many as 40% of the customers started dialing and an additional 45% abandoned or flashed before dial tone when dial tone was randomly delayed by 3 seconds. These would impact the switching system in adverse ways since the system would interpret the "early dialers" as partial dials and the flashes as false seizures. Sizable real time would be "wasted" processing these ineffective attempts who would then often reattempt their bids, thus further inflating the load.

The overload strategy described by the author in [1] consciously incorporates customer behavior phenomena in its design. New originations are served in a last-in-first-out (LIFO) manner unless they are still in the off hook state after a period of 20 to 30 seconds. At this point, the "patient" customer is treated as a new attempt and moved to the start of the queue. Customer activity such as flashing or abandoning is disregarded when in queue. This strategy has the effect that most customers experience no dial tone delays under overloads and hence will most likely be properly processed. The 20 to 30 second "time out" is introduced to insure that the early dialers (or "flashers") have completed dialing and have abandoned their requests. The automatic reattempt after 20 to 30 seconds insures that patient customers will be given priority. This strategy was shown to give vastly superior throughput performance under overloads by both analytic models, simulations and field studies. It was conjectured in [1] that the strategy would also give improved real time capacity. We examine this point in what follows. It is important to note that not only does the above LIFO strategy give superior performance for real time overload control, but that it has important application in controlling hardware overloads as well. For example, a version of LIFO has been shown to give superior performance in controlling digit receiver overloads as well since the same customer behavior phenomenon are pertinent. Access to many switching networks may also benefit from these considerations.

3. EFFECTS OF NONSTATIONARY TRAFFIC PATTERNS

At the 9th ITC, we discussed the results of the affects of nonstationary traffic patterns (see reference [2]). This paper pointed out the presence of significant nonstationary Poisson traffic patterns present in many exchanges. It also addressed the question of how to simply characterize these traffic patterns. The characterization would have to be simple enough to allow measurement of only a few parameters of the process on an ongoing basis. The characterization chosen was to use a mean and variance of call arrivals in small (10-30 second) time intervals for a fixed time interval, 15 minutes. The process is modelled as being piecewise stationary Poisson process. Thus, if \( \lambda_i, \ i = 1, ..., n \) is the mean intensity of a piecewise stationary process with peakedness factor \( z \), then the expectation of the measured sample variance over an interval of length \( T \) is given by (see [5]):

\[
(2z - 1)\lambda T + \frac{T^2}{n-1} \sum_{i=1}^{n} (\lambda_i - \lambda)^2
\]

where

\[
\lambda = \frac{1}{n} \sum_{i=1}^{n} \lambda_i
\]

Unfortunately, the peakedness and mean do not uniquely characterize a nonstationary Poisson process. There are a multitude of such processes which produce the same mean and peakedness. Each of these processes would presumably influence a traffic sensitive system in different ways.

The method used to characterize the traffic was to select the process which would produce the "worst case" results for the system of interest. In our case, the system investigated was a single server FIFO queuing system which served to model many local exchanges with the "pre-LIFO" overload strategy. The traffic model which resulted from this analysis was simply approximated by a two level Poisson process where the values of the two levels and their duration were determined by the peakedness and mean. It was noted that the "optimum" solution was robust in that small deviations from the optimal produced similar results. It was hypothesized that it would take an unusual traffic pattern in order that the effects of traffic variation be minimal for queuing systems. It was also stated that blocking systems would presumably be less sensitive. We will pursue this point in the next section.

Based on our results, a trial measurement was introduced into versions of a No. 1 ESS system. Instead of measuring a sample variance, a transformation was used termed the "variability ratio" defined below:

\[
VR = \frac{V(M) + 1}{2}
\]

where \( V(M) \) is given by

\[
V(M) = \frac{n}{n - 1} V/M + 1
\]
Here, \( \nu \) is the measured biased sample variance (see eq. (9)) and \( m \) is the sample mean of the traffic process. Our experience with the trial measurement indicated that the variability ratio was almost always less than 2.

4. DETERMINATION OF "WORST CASE TRAFFIC PATTERNS"

As in [2] we will use an approximate analysis to guide us in determining a worst case traffic pattern with which to analyze the capacity of systems with the LIFO/Time-Out strategy. The behavior of this overload strategy under heavy loads causes a new origination request to either be served almost immediately, or after a sufficiently long delay so that the "early dialer" has abandoned. Thus the system appears quite similar to a "blocked calls cleared" system with possible reattempts.

To simplify the analysis, we will consider a discrete time approximation to the system. We consider a finite time interval \( T \) which is subdivided into \( n \) equal subintervals. The time varying underlying call process is modeled as a vector \( \mathbf{N} \), whose \( k^{th} \) component is \( N_k \). The term \( N_k \) represents the number of calls arriving in the \( k^{th} \) subinterval. We assume that the system can process at most \( M \) calls in any subinterval. If more than \( M \) calls arrive in any subinterval, the excess will be "blocked" and a fraction, \( p \), of these will reattempt immediately in the next subinterval. In reality, calls which are "blocked" will of course reattempt at some random time in the future. However, by choosing the subinterval length equal to a mean abandonment time (say 30 seconds), we hope to capture the essentials of the phenomenon.

With these assumptions, the total number of calls, including reattempts, arriving in the subinterval \( k \) is given by \( C_k \):

\[
C_k = N_k + p(C_{k-1} - M)^+ \quad (5)
\]

The real time capacity of local exchanges is governed by a constraint on the percentage of calls which have delays exceeding a certain threshold, for example 3 seconds. In our model we will equate the number of calls exceeding 3 seconds by those calls which are "blocked". \( N_B(T) \) will denote the number of calls blocked in \([0,T]\) and \( N_S(T) \) will denote those served. Then, the fraction of calls blocked, \( P_B(T) \) is given by:

\[
P_B(T) = \frac{N_B(T)}{N_B(T) + N_S(T)} \quad (6)
\]

Since \( P_B(T) \) is a strictly monotonically increasing function of \( N_B(T) \), to determine a "worst case" traffic pattern it suffices to maximize the total number blocked in \([0,T]\). In accordance with our assumptions, this is simply:

\[
N_B(T) = \sum_{j=1}^{n} (C_j - M)^+ \quad (7)
\]

Let us denote by \( m \) and \( \nu \) the mean and variance of the underlying calls process (we use the "biased" sample variance for consistency with our work on the FIFO control strategy). Thus,

\[
m = \frac{1}{n} \sum_{j=1}^{n} N_j \quad (8)
\]

\[
\nu = \frac{1}{n} \sum_{j=1}^{n} (N_j - m)^2 \quad (9)
\]

We want to therefore maximize the expression in equation (7) subject to the constraints of equations (8) and (9).

We solve the problem in two steps. For a fixed integer \( r \), \( 1 \leq r \leq n \), we first find the underlying traffic pattern which produces the most total blocking in \( r \) subintervals, and then maximize with respect to \( r \). Without loss of generality, the blocking can be assumed to occur in a contiguous set of subintervals. Also, the first subinterval in such a contiguous set can be assumed to be first subinterval of \([0,T]\). Thus, we seek to maximize \( N_B^{r} \) defined by:

\[
N_B^r(T) = \sum_{j=1}^{r} (C_j - M) \quad (10)
\]

Using equation (5) recursively, we can determine that:

\[
N_B^r(T) = \sum_{j=1}^{r} A_j N_j - M \sum_{j=0}^{r-1} (r-j)p^j \quad (11)
\]

where,

\[
A_j = \frac{1 - p^{j+1}}{1 - p} \quad (12)
\]

for \( j = 0, 1, \ldots, r-1 \) and \( \hat{p} \neq 1 \)

As a final step to setting up the appropriate optimization problem, we will "relax" the variance constraint in equation (9) by imposing an inequality. We will solve this "relaxed" problem and show that the solution obtained satisfies the equality condition as well. Thus our optimization problem is:

maximize \( N_B^r(T) \) \quad (13)

subject to
\[ M(N_1, \ldots, N_n) = mn \quad (14) \]
\[ V(N_1, \ldots, N_n) \leq vn \quad (15) \]
where,
\[ M(N_1, \ldots, N_n) = \sum_{j=1}^{n} N_j \quad (16) \]
\[ V(N_1, \ldots, N_n) = \sum_{j=1}^{n} (N_j - m)^2 \quad (17) \]

We can solve this problem by using classical Calculus of Variation methods. Thus, the "Lagrangian" is given by:
\[ L = N_j \{ \lambda_1 (M(N_1, \ldots, N_n) - mn) + \lambda_2 [V(N_1, \ldots, N_n) - vn] \} \quad (18) \]

Equating the partial derivatives with respect to \( N_1, \ldots, N_n \) equal to zero one obtains:
\[ \lambda_1 (N_j - m) = \frac{(A_{r,j} + \lambda_1)}{2} \quad j = 1, \ldots, r \quad (19) \]
\[ \lambda_2 (N_j - m) = \frac{\lambda_1}{2} \quad j = r+1, \ldots, n \quad (20) \]

Summing the above equations, and using equation (8) we obtain:
\[ \lambda_1 = \frac{\sum_{j=0}^{r-1} A_j}{n} \quad (21) \]

Squaring both sides of equations (19) and (20), summing and using (21) we obtain:
\[ 4\lambda_2^2 \sum_{j=0}^{r-1} (N_j - m)^2 = \sum_{j=0}^{r-1} A_j^2 - \left( \sum_{j=0}^{r-1} A_j \right)^2 / n \quad (22) \]

With a little algebra one can show that for \( p \neq 0 \), the right hand side of equation (22) is positive, and so \( \lambda_2 \) cannot be zero. Therefore, the condition defined by (21) is sufficient for the mean constraint to be satisfied. Moreover, if,
\[ \lambda_2^2 = \frac{\sum_{j=0}^{r-1} A_j^2 - \left( \sum_{j=0}^{r-1} A_j \right)^2 / n}{4vn} \quad (23) \]
then the variance inequality constraint is satisfied as well. Clearly, if equality holds in (23), then equality will hold for the constraint.

Let \( \lambda_2 \) denote the negative value of \( \lambda_2 \) corresponding to equality in (23), and let \( H(L) \) denote the Hessian matrix of the Lagrangian. Performing the appropriate differentiations results in:
\[ H(L) = 2\lambda_2^2 I \quad (24) \]

Here, \( I \) is the identity matrix. Thus \( H(L) \) is positive definite and the usual sufficiency conditions (Kuhn-Tucker) for a local maximum are fulfilled. Because of the convexity of \( N_j \{ T \) and the constraint set, the solution is also a global maximum. The maximum point must also lie on the boundary of the constraint set and so the solution provides for equality of the variance constraint.

By solving the respective equations we observe that:
\[ N_j - N_{j+1} = \frac{-p^{r-j}}{2\lambda_2} \quad j = 1, \ldots, r \quad (25) \]
\[ N_j - N_{j+1} = 0 \quad j = r+1, \ldots, n \quad (26) \]

From (25) we see that the "worst case" call stream is strictly monotonically decreasing \( (\lambda_2^2 \leq 0) \) for the first \( r \) intervals and remains constant for the remainder of the time interval \([0,T] \). Also, the rate of decrease increases with \( r \).

The solution obtained has the following physical interpretation. Since we assume that blocked calls reattempt in the next interval, calls arriving earliest in the interval are provided with the largest number of subintervals in which to reattempt. Thus from the point of view of maximizing the total number of blocked calls, the largest possible number should arrive in the first subinterval, the second largest in the second subinterval, etc. .

The optimization problem is completed by computing numerically the value of the maximum blocking for each \( r \), and simply selecting the \( r \) which produces the most blocking.

Finally, we note that for "flat" traffic, i.e. \( v = 0 \), we obtain:
\[ P_B(T) = \frac{(m-C)^2}{1-p} \left( \frac{n - P(1-p)}{1-p} \right) \quad (27) \]
for \( p \neq 1 \)

This value was used as a benchmark to determine the relative degradation of capacity with increasing
variances. Note that, while the case \( p = 1 \) is unrealistic, a simple closed form solution exists to our problem in this event:

\[
P_0(T) = \frac{n(n + 1)}{2} \left(m - C\right)^+ \tag{28}
\]

5. SIMULATION RESULTS

The "worst case" deterministic traffic was used as the mean arrival rate for a piecewise stationary Poisson process. This was then used as an input to a detailed simulation of a No. 1 \( ESS^TM \) and the results compared to both our analytic approximation and to the results obtained by using the worst case traffic model in reference [2] for the FIFO control strategy.

The values of \( T, N \) and \( p \) were chosen to be 900 seconds 30 and 8 respectively. Thus the subinterval sizes were 30 seconds long. Using \( M = 27,500 \) calls per quarter hour and a variability ratio of 2, the results of the analytic model and simulation were compared. With \( m = 27,500, 26,750 \) and 26,250 calls per quarter hour, the analytic model predicted blockings of 23.5%, 18% and 13% respectively. The simulation values obtained were 26.5%, 15% and 7.5%. The disparity between the results can be considered mild in view of the simplicity of the analytic model. Note also that there is close agreement for higher loads since it is in this region that the system more closely resembles a blocking system.

We should emphasize that we are interested in comparing relative changes in performance, and that the analytic results are to be used as guides to selecting stochastic inputs to a simulation.

In figure 1 we plot the load-service relationships obtained from our simulation for a range of variability ratios.

The high day criterion used to determine real time capacity in local SPC exchanges is 20% dial tone delay in excess of 3 seconds. With this criterion we see that for a variability ratio of 2, the reduction of capacity is 2% compared with stationary Poisson traffic. As we indicated earlier, variability ratios in excess of 2 are uncommon.

These results are dramatic when the results of the FIFO strategy are compared. The results from [2] were used to determine a suitable "worst case" traffic model. It is a two level time varying Poisson process. This was used as input to the same simulation, but with the FIFO strategy. In this case, a capacity reduction of 10% for the same parameters was determined. The differences in relative performance are not so surprising. Blocking type systems generally are less sensitive to traffic mean changes than are queuing type systems.

6. FIELD DATA

Dial Tone Delay data were gathered for all No. 1 \( ESS^TM \) local Bell Operating Company exchanges for the period 1979 through 1984 inclusive. The exchanges were divided into categories according to whether the systems had the LIFO strategy or the pre-LIFO (or FIFO) strategy. Offices having hardware shortages and not just real time shortages were eliminated from the study. The measured % of Dial Tone Delay was plotted versus the measured number of E-E cycles for various types of exchanges. An E-E cycle is the number of times the central processor cycles through its base level activities (see [6]). It was found that insufficient information existed to analyze offices not equipped with an auxiliary Signal Processor. For Signal Processor equipped exchanges however, the field study results compared favorably with the simulation results. Figure 2 indicates the simulation results with various variability ratios. (Recall that a variability ratio of more than 2 is unusual.)

Figure 3 depicts the results of the field investigation. A statistical curve fit to the data is indicated, with the LIFO (-1) indicating the curve fit when one data point was deleted. Note that the scales of figures 2 and 3 differ by a factor of 4. The simulation curves for a variability ratio of 2 are an excellent approximation (a slight upper bound, as expected) to the field data!
It was determined that a LIFO switch must slow down at least an additional 4600 E-E cycles per hour before it reaches 20% dial tone delay as compared with a NOLIFO (or FIFO) switch.

In figure 4 we plot simulation results indicating the relationship between E-E cycles and load for a range of variability ratios for the LIFO case.

Note that there is nearly a linear relationship between E-E cycles and load except at the high values of load. This is the region where overload control actions limit the rate at which originations can be served by the system. Using the linear relationship for various values of parameters, it was determined that the LIFO strategy yields a 3-7% capacity improvement over the pre LIFO strategy. The precise improvement is dependent on the amount of overhead in the office.

There are many scenarios whereby service difficulties experienced by one IC can possibly affect the service of other ICs unless specific corrective actions are taken. These scenarios are produced by examining the resources the ICs share in common and focusing on specific types of service difficulties. Among resources shared in common are real time, digit receivers, switching network, transmitters (outpulsers), memory and in some cases even trunks.

Among the possible service difficulties that might impact the equal access switch are the failure of the trunk group to the IC, the failure of an IC switch or other problems internal to the IC's network.

We will examine a subset of the possibilities and quantify the magnitude of the impact and the new overload control issues it presents. Specifically we will...
focus on what happens when the IC switch which is accessed by a trunk group from the equal access switch experiences a total failure. We will assume that the equal access switch has no knowledge of the failure.

As we indicated before, one item which can be shared by many ICs (as well as other local interoffice calls) is the common group of transmitters. If the transmitters are Multifrequency transmitters, their holding time is typically 2.5 seconds. In the event of the IC switch failure, the equal access switch will seize a trunk to the IC, and wait for a "wink" signal which is used to denote that outpulsing of digits may begin. Since the IC switch has failed, no wink will be forthcoming and the transmitter will be held until a time-out threshold is exceeded. A typical time-out mechanism is that if more than 4 to 8 seconds have elapsed before detection of wink, that transmitter may be preempted (reused) by another call requiring that type of transmitter. If no call arrives to preempt the transmitter, then the transmitter will time-out in 16 seconds and the timed-out call will be sent to reorder. The reason for the range of 4 to 8 seconds is that the determination of which transmitters are preemptable is done only every 4 seconds by switches such as a No. 1 ESS™. Other time-out mechanisms will produce similar results and so we will concentrate on this mechanism. There is normally no queuing for transmitters. Thus, calls destined to the failed IC will on average, hold the MF transmitters for 6 seconds or more.

Not only is the holding time thus increased for these calls, but after preemption or time-out, the calls will often reattempt. In addition, calls not destined for the failed switch will find that there are no free or preemptable transmitters available and they will be sent to reorder, and will often reattempt.

8. ANALYSIS OF TRANSMITTER BLOCKING

We will assume that the reattempt probability for all calls is a constant, p. Without loss of generality we can assume that calls which are timed-out after 16 seconds in effect have the same holding time as a preempted call. This will not affect the number of calls blocked or preempted since a transmitter can only time-out if there are no calls arriving to preempt it. Again, without loss of generality, we assume that all calls to the failed IC will receive a uniform holding time between 4 and 8 seconds (the minimum preemption interval) which of course yields a mean holding time of 1/4p = 6 seconds. All calls not destined to the failed IC will receive a normal holding time with mean 1/\mu_0 = 2.5 seconds.

We assume that reattempts occur sufficiently far in the future so that they can be modelled as an independent Poisson process. Thus, the entire stream is Poisson. In a blocked calls cleared system offered Poisson traffic, the blocking is independent of holding time distribution, depends only on its mean, and is given by the classical Erlang B formula. Thus we can derive a nonlinear equation to be solved which involves one unknown, the total offered load. This can be solved numerically using the method of successive approximations.

This analysis indicates that the impact can be dramatic. If the fraction of MF calls to a failed IC is only 10%, the overall blocking will be 20% (at ABSBH loads) if the reattempt rate were 8.

9. POSSIBLE CONTROL STRATEGIES

Two general approaches to mitigating these effects are to provide additional equipment or to require that failed IC switches send a failure indication. A similar alternative is to monitor completion ratios to each IC network. If the completion ratio were zero, these calls could be blocked from attempting to seize a transmitter. If the ratio were nonzero, caution must be exercised. "Call gapping" controls could then be used to limit the rate at which calls would be sent to the IC in question. It is more difficult to quantify the effects of IC failures on queuing systems. Because of the effects of abandonments, different techniques must be employed. An approach to this kind of problem has recently been proposed by Sze in [4]. In [4], the problem of non exponential holding times is also considered.

ACKNOWLEDGEMENTS

We wish to thank J. M. Bennett of Bell Communications Research for his analysis of the field data and E. H. Lipper of AT&T Bell Laboratories for his work on the worst case traffic model.

REFERENCES


