A MODEL FOR ESTIMATING THE REAL-TIME CAPACITY OF CERTAIN CLASSES OF CENTRAL PROCESSORS

N. Farber
Bell Telephone Laboratories, Inc.
Holmdel, New Jersey USA

ABSTRACT
This paper describes a model which can be used to estimate the call-carrying capacity of certain classes of central processors. The model assumptions have been verified using data from simulations of two such systems, designed and developed by the Bell Telephone Laboratories, as well as data obtained from field studies conducted on several operational processors. Pertinent results of these studies are reported.

1. INTRODUCTION
A model has been developed which can be used to estimate the call-carrying capacity of a real-time telephone switching office controlled by a stored program which processes a set of work lists in a predefined order, subject to interruption in order to execute higher priority (input/output) tasks. The assumptions used for the model were derived as a result of simulation and field studies conducted to investigate the processor performance of two such systems designed and developed by the Bell Telephone Laboratories (No. 1 ESS and No. 1 TSPS), but the concepts used and the results obtained can be generalized to apply to other real-time systems of similar design.

2. FUNCTIONAL DESCRIPTION OF SYSTEMS ModeLED
The model assumes that the switching office contains a stored program controlled processor which operates at two levels of control: base level and interrupt level. The base level control program is responsible for servicing a set of work lists (called hoppers) in a predefined order. At regular intervals, the base level program is interrupted and control is passed to the interrupt level control program. Interrupt level tasks pertain primarily to input/output functions such as the scanning of trunks, connecting of network links as prescribed by the base level programs, etc. When an input task determines that further processing is required for a given call, an entry is placed in the appropriate hopper (for example, if the line scanning task determines that a particular subscriber line has gone off-hook, an entry is placed in the Line Origination Hopper, referencing the relevant line). The base level programs, when servicing that entry, determine the appropriate call processing actions to be taken. This may result in a request to the interrupt level program that network connections (or disconnections) be performed. The request is added to an output work list which is processed by an output task program (during an interrupt period) according to a scheduled frequency. Completion of a particular output task (such as the connection of a digit receiver to a subscriber line) results in another hopper entry informing the base level program that the requested activity has been completed.

Clearly then, as the traffic volume being processed increases, more and more entries are being placed in the hoppers, as a result of which the time to service every hopper in the work list sequence, which we shall refer to as the cycle time, gets longer and longer. Thus, the cycle time distribution is an important measure of the speed with which the processor is responding to call processing requests. We shall, however, be concerned primarily with the cycle rate, i.e., the number, E, of cycles completed per unit time, for reasons which will become apparent. (If, over a period of time T, the average cycle time is T, then E = T/t.)
We also assume that there are service requirements* which the processor must satisfy, so that we define processor capacity as the maximum call volume which can be processed without violating any of the stipulated service criteria. We thus distinguish between the capacity of a real-time processor and the maximum throughput of the processor, the latter not being the concern of this paper.

3. THE COMPOSITION OF REAL-TIME

The model employed in the estimation of central processor capacity was developed by considering the decomposition of real-time into its four basic components: call dependent and call independent tasks executed at base level and interrupt level. This decomposition leads to a useful characterization of central processor response to offered load, and suggests a convenient formulation of the capacity estimation problem.

3.1 NONLOADED SYSTEMS

Consider a processor which is engineered to handle large traffic loads, but to which no traffic is submitted during a particular hour. If we were to categorize the processor work effort during such a no-load hour, we would find essentially two types of work being done:

1) Interrupt programs obtain control every m ms in order to perform specified I/O functions. Certain of these functions (such as trunk and line scanning) must be performed even when there is no traffic being handled. Thus, regardless of loading conditions, a portion of the hour, J₀, is consumed processing traffic independent I/O tasks. During a no-load hour, J₀ represents the total I/O real time consumption, since traffic dependent I/O functions will not be performed.

2) For the remainder of the hour the processor would be executing scheduled maintenance tasks, if any, and would be cycling through the hoppers looking for base level call processing tasks, of which there would be none. Call this proportion of the hour L₀, the amount of time spent doing base level overhead during a no-load hour. Clearly, L₀ = 1 - J₀.

Figure 1 illustrates the decomposition of the no-load hour. Given L₀, we can now compute E₀, the number of times we complete a cycle during this hour, assuming we know k₀, the average amount of base level time required to complete one cycle when there is no load in the system. Obviously,

\[ E₀ = \frac{L₀}{k₀}. \]  (1)

k₀ (the base level overhead, in hours, per cycle) is independent of traffic conditions, i.e., the overhead portion of a cycle will always average k₀ for a given processor generic program.

3.2 LOADED SYSTEMS

Suppose a traffic load is submitted to the same office considered in 3.1 and we similarly decompose one hour of real-time. There will now be four ingredients in the decomposition, as illustrated in Figure 2.

Figure 2

Because certain maintenance tasks are not performed during every cycle, k₀ can vary slightly from cycle to cycle.

---

* Service requirements are usually specified as constraints governing the time to perform various call processing functions.
The same portion of the hour $J_i$ will be spent processing traffic independent I/O work as in 3.1.

A portion of the hour $J_{cp}$ will be spent processing traffic dependent I/O work (connecting digit receivers, disconnecting channels, etc.) as required by call processing.

A time $L_{cp}$ will be spent in base level, processing traffic dependent work (e.g., called party number analysis, requests for hardware connections, billing, timing, etc.), i.e., servicing the entries in the set of work lists.

The remainder of the hour, $L$, will be spent in base level looking for entries in the work lists. Then,

$$ L = 1 - J_0 - J_{cp} - L_{cp}. \quad (2) $$

If we let $E$ be the number of cycles during the hour, then we again obtain the relationship:

$$ E = L/k_0 \quad (3) $$

where $k_0$ is the same quantity as in 3.1. Obviously, $L < L_0$, so that $E < E_0$.

Suppose we find that we processed $N$ calls during the hour. If $k$ denotes the average amount of real time required to process a call (base level and I/O level), then

$$ kn = J_{cp} + L_{cp}. \quad (4) $$

From equations (2), (3), and (4) we obtain the relationship

$$ J_0 + kn + Ei_0 = 1, \quad (5) $$

so that

$$ E = 1/J_0 - kn/L_0 \quad (6) $$

That is, if $k$ remains fixed, the number of E's per unit time decreases linearly with increased call rate. It is easily seen that, when $N = 0$,

$$ E = 1/L_0 - L_0/J_0, \quad (7) $$

which checks with Figure 1.

The quantity $k$ depends on the kind of traffic being processed, so that equation (6) represents a family of lines for any office with parameters $J_0$ and $L_0$. If the proportion of various kinds of traffic are known, $k$ can be estimated by simulation or by direct count of processor program execution cycles. A simple procedure for estimating $k$ from actual data is described in reference [1]. The quantity $J_0$ depends primarily on the amount and type of hardware in the office. It is traffic independent only in the sense that, once the hardware configuration is specified, $J_0$ does not vary with load. However, if a particular office were to be expanded to handle additional traffic, $J_0$ would change according to the alterations in hardware facilities. $L_0$, on the other hand, depends only on the particular stored program which the office utilizes, and would change only if program modifications were introduced.

4. THE CAPACITY ASSUMPTION

The capacity of a real-time system is defined to be the maximum load at which all service criteria imposed upon system activities are satisfied.

If we restrict our attention to the relationship between service and delays which work list entries encounter, ignoring factors such as availability of hardware items, software items, link network paths, etc., then equation (6) suggests a convenient characterization of the relationship between central processor activity and customer service. If, for example, we consider the limiting case in equation (6), when $kn = 1 - J_0$, so that $E = 0$, then we have a situation where no processing time is allotted for base level overhead. Practically, this can never occur since base level work cannot be performed without expending base level overhead effort to locate the work. Theoretically, however, as $kn$ approaches $1 - J_0$ the average cycle time of the processor increases without bound, causing excessive increases in the time required for the central processor to progress from one hopper to another. Such a slowdown in the cycle time clearly results in a deterioration of service.9 The dependence of quality of service on the cycle time suggests the following formulation of the capacity assumption:

For a given set of parameters $J_0$, $k$, and $k$, and a given set of service criteria, there exists a value $E_{min}$ such that the service criteria will be satisfied if and only if $E \geq E_{min}$.

The capacity in calls per hour of a particular office (an "office" being defined by the triple $(J_0, k, E)$) is determined by using the value $E_{min}$ in equation (6), and solving for $N_{max}$, i.e.,

$$ N_{max} = \frac{1-J_0-kE_{min}}{k} \quad (8) $$

where $k_0$, $J_0$, and $k$ are expressed in hours.

5. DETERMINATION OF $E_{min}$

Suppose there are $m$ service constraints which must be satisfied at and below capacity loads. For the $i^{th}$ constraint, there will be a value $E_i^{min}$ such that the $i^{th}$ constraint will be satisfied if and only if $E \geq E_i^{min}$. The value $E_i^{min}$, then, must be such that

$$ E_{min} = \max _i E_i^{min} \quad (9) $$

so that all constraints are satisfied when $E \geq E_{min}$. If the $m^{th}$ constraint corresponds to $E_{min}$, i.e., $\max _i E_i^{min} = E_{min}$, then we call the $m^{th}$ constraint the dominant service constraint.

In general, the dominant constraint can vary from office to office. With respect to No. 1 ESS, for example, the dominant constraint is that pertaining to the proportion of calls which

9 For example, as the cycle time increases, so does the time between consecutive processing periods of the line origination hopper in No. 1 ESS. Since entries in this hopper pertain to lines waiting for dial tone, the result is increased dial tone delay.
receive dial tone delay (DTD) greater than 3.0 seconds, assuming that there are dial tone requests to be processed. It is obvious, however, that the DTD constraint cannot be dominant in any ESS office which processes only incoming traffic, since in fact there will be no dial tone requests for such an office.

With respect to the No. 1 ESS, prior to recent program modifications, the dominant constraint corresponded to the delay to dial pulse receiver connection, provided the office processed calls originating at noncommon control step-by-step local offices. Otherwise, the dominant constraint pertained to the average delay to connect a call to an operator.

Emin, therefore, may be dependent on the traffic characteristics of the load being processed, if only with respect to identifying the dominant constraint. It remains to consider those factors, independent of traffic characteristics, which might affect the value of Emin.

5.1 CUSTOMER SERVICE AND HOPPER DELAY

Associated with each service constraint are key hoppers* for which processing delays (incurred by entries in these hoppers) affect the particular service indicator involved in the constraint. (Delays due to hardware and/or software unavailability may also affect the indicator, but we are concerned here only with processor-caused delays.) The delay seen by entries in a particular hopper is a function of the average rate, \( H \), with which the hopper is visited. \( H \) is then related to \( E \) by the expression \( H = fE \), where \( f \) is the frequency per cycle with which the hopper is visited, the value of \( f \) being determined by the structure of the work list sequence. (In the No. 1 ESS, for example, \( f = 15 \) for a class A hopper, \( f = 8 \) for a class B hopper, \( f = 4 \) for a class C hopper, etc.).

It is this relationship between \( H \) and \( E \) which lends credence to the tacit capacity assumption that service is a function of \( E \). The fact that the hopper visitation rate for every hopper is linearly related to \( E \) justifies using \( E \) as the measure of processor response to load in the capacity model.

It is true, however, that customer service is not completely determined by the obvious hopper visitation rate, since the variation of the hopper visitation rate can also affect the service measurement. Since \( H = fE \) may be viewed as a statistical equation, \( Var \ (H) = f^2 \ Var \ (E) \), and it follows that customer service is dependent on the variance of \( E \) as well as on the average value of \( E \).

Assuming, then, that \( E_{\text{min}} \) is to be determined from simulation studies, it is necessary to include as independent variables in the experimentation all office parameters which influence cycle time variability. In general, \( E_{\text{min}} \) for a given system can be expected to be a function of the office parameters \( (J_0, k_o, \text{and } k) \) rather than a unique value.

That this is indeed the case has been illustrated by the results of the following study of the relationship between the value of \( E_{\text{min}} \) and the office parameter \( J_0 \) (proportion of time spent processing traffic independent I/O work).

5.2 Emin VS. TRAFFIC INDEPENDENT I/O PROCESSING

Suppose we view the processor as a single server accommodating two types of work sources: (1) traffic independent I/O and (2) call processing. The traffic independent I/O may be considered to be a "smoothed" work source (constant arrivals relatively constant holding time), while the call processing source is highly "peaked". The variability of cycle times will depend on the relative proportions of the two sources being processed.

Consider two offices, one with parameters \((J_0, k_0, k)\) and one with parameters \((J_0', k_0', k)\), where \( J_0 < J_0' \). Let \( N_1 \), the number of calls processed in office 1, and \( N_2 \), the number of calls processed in office 2, be such that \( E \) is the same for both offices; i.e.,

\[
\frac{(1 - J_0')}{J_0'} - \frac{kN_1}{k_0} = \frac{(1 - J_0^2)}{J_0^2} - \frac{kN_2}{k_0^2}.
\]

This implies that the average cycle time \((1/E)\) is the same for both offices. But office 1, which is processing more peaked traffic than office 2 \((N_1 \text{ must be greater than } N_2 \text{ to satisfy the equation})\), will have more variability about the average cycle time, and hence will have a longer and heavier tail (Figure 3). Since service is related to the tail of the cycle time distribution as well as to the mean cycle time, then office 1 will exhibit worse service than office 2. In order to equalize service, then, the average cycle time for office 1 must be less than that for office 2. The implication is that the value of \( E_{\text{min}} \) for office 1 must be greater than the value of \( E_{\text{min}} \) for office 2 if the grade of service is to be equivalent in both offices.

5.3 EXPERIMENTAL RESULTS

A simple experimental design was constructed to test the hypothesis that \( E_{\text{min}} \) was a function of \( J_0 \), the proportion of traffic independent I/O processing. A range of values of \( J_0 \) was selected for testing, and for each value of \( J_0 \) a No. 1 ESS simulation program was used to determine the number of E's for which 5 percent of the originating traffic received dial tone delay of three seconds or more. The results of this experiment are charted in Figure 4. The conclusion to be drawn seems clear: The percentage of time spent processing traffic independent I/O work does influence \( E_{\text{min}} \) in the manner hypothesized; i.e., as \( J_0 \) increases, \( E_{\text{min}} \) can be decreased.

* During subsequent discussion, "hopper" and "key hopper" will be used interchangeably; i.e., we restrict our attention to the key hoppers.
Similar studies were conducted with the No. 1 TSPS simulator, the results of which are plotted in Figure 5. As can be seen, the qualitative effect of varying $J_0$ is as hypothesized.

One point should be introduced here in the interests of adding insight to the concepts at hand. The simulation programs (both ESS and TSPS) treat all I/O as if it were traffic independent. Thus, for example, when constructing a TSPS simulation run with 35 percent traffic independent I/O, 10 percent traffic dependent I/O, and 35 percent traffic dependent base level, a somewhat conservative view was adopted. The corresponding simulation run was made to produce 35 percent traffic independent I/O and 45 percent traffic dependent base level, i.e., the traffic dependent I/O was simulated by considering it to be the same as traffic dependent base level. Since traffic dependent I/O work is probably less variable than traffic dependent base level work, the simulation results will tend to overestimate $E_{\text{min}}$.

This hypothesis suggests that the model represented by equation (6) should, in the interests of completeness, be amended to distinguish between the call processing interrupt level effort and the call processing base level effort. This can be easily done by letting

$$k = k_J + k_L$$

(10)

where $k_J$ and $k_L$ are the average amounts of time required to process a call on interrupt level and base level respectively. Equation (6) is thus extended to include the distinction between the two types of call processing, so that

$$E = \frac{(1-J_0)}{\bar{Q}_0} \left( \frac{1}{\bar{T}_J} + \frac{1}{\bar{T}_L} \right) N$$

(11)

It is possible that $E_{\text{min}}$ depends on the composition of $k$, i.e., the relative contributions of $k_J$ and $k_L$. Because the ESS and TSPS simulators cannot simulate the traffic dependent I/O effect ($k_J$), this possibility has not been explored. It is unlikely, however, that the effect on $E_{\text{min}}$ would be too significant since simulation studies have shown that $E_{\text{min}}$ is relatively independent of the value of $k$ in any event. The value of $J_0$ is by far the most important contributor in determining the value of $E_{\text{min}}$. 

426/5
6. VERIFICATION OF THE CAPACITY MODEL

The linear relationship between the cycle rate and call volume has been verified by a number of studies, using both simulation data and data trials conducted on operational ESS offices [1]. Verification of predicted $E_{\text{min}}$ values and of the $J_0$ effect was not possible until recently when several operating ESS offices began to experience real-time congestion and severe dial tone delays. (The data from these offices also verified that the linearity relationship holds true right up to capacity conditions.)

In November of 1968, the No. 1 ESS signal processor* office in Columbus, Ohio was cut over and immediately began to exhibit the effects of real-time congestion. Figure 6 shows the office's DTD characteristics as a function of the number of E's per quarter hour. It can be seen that the five percent delay point occurs at about 3500 E's per quarter hour, which agrees fairly well with the linear extrapolation of Figure 4 ($J_0$ is about 5.7 percent for the Columbus office).

More recently, the Beverly Hills nonsignal processor office experienced some real-time congestion and dial tone delay data was obtained for this office. Figure 7 shows the Beverly Hills DTD characteristics. The 5 percent point can be seen to occur at about 2300 E's per quarter hour. Since the value of $J_0$ at Beverly Hills has been determined to be about 38 percent, we again see close agreement between the simulation prediction (Figure 4) and reality, with the simulation estimate being slightly conservative as we expected.

* A signal processor office essentially consists of two processors, one responsible for much of the I/O work and the other for base level work. A nonsignal processor office has one processor doing both types of work.

7. EFFECT OF OVERLOAD CONTROL ON $E_{\text{min}}$

Both the No. 1 ESS and the No. 1 TSPS utilize automatic overload control strategies which are designed to prevent the system from reaching a condition, under extremely heavy loads, where the cycle times become so long that the system's call completion capability is decreased. Under present design conditions, the load control activity consists of monitoring the cycle times and restricting the processing of new calls originating when clusters of long cycle times occur. These load control actions are essentially anticipatory; that is, they begin to restrict processing when the cycle times suggest that overload conditions will arise if the present processing rate is not reduced. Thus, the decision to reduce the processing rate is a statistical one and quite frequently reductions will be made when overloads do not materialize.

If a service constraint is associated with a controlled hopper, as is the case with No. 1 ESS (Line Service Request Hopper) and No. 1 TSPS (Trunk Seizure and Answer Hopper), then this will effect the identification of the dominant constraint. A constraint associated with a controlled hopper is more likely to be the dominant constraint since the processor speed of response must be fast enough to overcome any artificial delay caused by the load control scheme.

This fact is illustrated by the No. 1 ESS. As has been pointed out, the dial tone delay constraint is dominant for any ESS office which processes line originations. The presence of the load control strategy, which regulates the hopper which holds line origination requests, causes the DTD constraint to dominate. If the load control mechanism were removed, the time to establish a talking path would be the constraining service requirement.

As an additional consideration, if the control strategy involves a statistical procedure, as do the ESS and TSPS strategies, the efficiency of the parameters used by the strategy will depend on the variability of the processor response to load. Referring to Figure 3, for example, a particular set of control parameters...
for the ESS strategy will be more restrictive for office 1 than for office 2. This fact suggests that the control parameters for a particular ESS office should be made to depend on the amount of traffic independent I/O processing which the office must perform.

8. SUMMARY AND CONCLUSIONS

The linear relationship between the number of calls processed and the cycle rate provides a simple and effective measure of processor response to load. The problem of estimating the processor capacity then becomes one of identifying the dominant service constraint and determining the value of $E_{\text{min}}$ for that constraint.

This procedure has been used to estimate the capacity of the No. 1 ESS and the No. 1 TSFS. Simulation studies have shown that the minimum cycle rate required to guarantee the satisfaction of all service constraints depends on the composition of real-time and therefore varies from office to office. It is hypothesized that this relationship between minimum cycle rate and the components of real-time results from the different effect each component has on the variability of the system response to load.

In light of this, it follows that when conducting simulation studies to determine system capacity, care must be taken to include as independent variables all system parameters which affect the variance of the system response distribution.

BIBLIOGRAPHY
