Peakedness in Switching Machines: its Effect
and Estimation

H. Heffes and J.M. Holtzman
Bell Telephone Laboratories, Holmdel, New Jersey, U.S.A.

ABSTRACT

Peakedness has been shown to have a degrading effect upon performance of switching machines. This paper reviews the effect of peakedness upon a class of electromechanical switching machines (describable by GI/M/1 systems) and focuses on a method of estimating it.

One method of estimating peakedness is based on the measured delay; one determines which peakedness could have caused the delay. This method is desirable from the point of view of directly measuring the effect of peakedness upon performance. However, it can become unreliable when the delays are low, as they normally should be.

Another method (discussed in detail), applicable at low loads, is based on within-hour samples of usage. By determining the variance-mean ratio of the number of busy servers in a GI/M/N queueing system, a peakedness estimation procedure is defined. Peakedness as a function of holding time plays a role in interpreting the results. Statistical accuracy of the procedure is discussed.

1. INTRODUCTION

The importance of peakedness upon blocking in trunking theory and practice has been recognized for some time [1,2]. Recently attention has been given to the effect of peaked traffic on delays in, and capacities of switching machines. Capacity is defined here as the maximum load that can be offered for which a delay criterion is met.

The study of peakedness in switching machines reported here was originally motivated by discrepancies found between capacities predicted by theory, assuming Poisson input, and the measured capacities. For example, it was observed that in many crossbar tandems the sender attachment delay (generally called SADR after the Sender Attachment Delay Recorder) often exceeded that predicted by Poisson theory and that the capacities were correspondingly lower than expected. Investigation of possible causes of the discrepancies showed peakedness to be a significant contributor for those switching machines that receive overflow traffic. The degrading effect of peakedness in delay systems was shown in [4].

The effect of peakedness will be shown in Section II. These results apply quantitatively to some electromechanical switching machines. Peakedness has a similar degrading effect upon some electronic switching machines although we shall not specifically discuss these here.

A new sender engineering procedure was developed and put into practice for crossbar tandems to take peakedness into account in the determination of sender requirements. The procedure, discussed in Section II, characterizes the capacity of an office by providing an estimate of peakedness (actually, an effective peakedness which also takes into account within-hour load variations, which has an effect similar to peakedness). The peakedness is estimated from the measured SADR. However, when the SADR is too low or zero, the technique becomes unreliable.

Since the objective is expressly to keep the SADR low, it was desired to develop an alternate technique of estimating peakedness applicable at low loads and not dependent upon SADR.

A method of estimating the peakedness of a traffic stream offered to a queueing system, based on within-hour samples of usage, was investigated. The pursuit of this approach was stimulated by a suggestion of W. S. Hayward. In particular, the method infers peakedness from the sample variance-mean ratio of the number of busy servers. We thus call it the V/M method. The rationale for using the sample mean and variance of busy servers to estimate peakedness is that for a large number of servers and low input load, the ratio V/M would be expected to approximate the peakedness (defined as the ratio V/M when the number of servers is infinite). Since nothing has yet been published on this method, we shall devote much of this paper to it. Section III shows how to infer the peakedness from the measured V/M using curves of V/M vs. M. The curves are interpreted in Section IV. An important consideration in the interpretation is the dependence of peakedness upon holding time. The accuracy of the method is considered in Section V.

2. EFFECT OF PEAKEDNESS UPON DELAYS AND AN ESTIMATION PROCEDURE BASED ON DELAYS

As mentioned in the Introduction, an analysis of queueing system with peaked inputs is given in [4]. Two models, shown in Figure 1, are considered. The first, Figure 1a, has Poisson traffic with rate λeq offered to Neq trunks, each with (mean) holding time μ−1 (independent exponential distributed random variables). The overflow is offered to N servers, each with exponential holding time μ−2.

Note that we have μ−1 ≠ μ−2. That is, the physical mechanism for peaked traffic being offered to a switching machine is overflow from trunk groups. (Note that the overflow traffic offered to a switching machine is actually carried by a trunk group into the machine. There is thus a smoothing effect on the carried trunk group; this is analyzed in [6]).

An approximation to Figure 1a, which is simpler to solve, is given in Figure 1b, where the overflow process is represented by an interrupted Poisson process (i.p.p.) [5]. Comparison between the two analyses is given in [4].

The effect of peakedness shows up clearly on load service curves (plots of service vs. offered load). A set of
such load service curves is shown in Figure 2. Prob(delay >2.5 sec) is plotted vs. ccs/sender. The (SADR) delay criterion for crossbar tandem senders is Prob(delay>5 sec) <0.005. There is an approximately 0.5 sec fixed delay so that Prob(delay>2.5 sec) has been calculated (using an approximation to the queueing analysis of [4]). It is seen that for a given offered load, the delays increase substantially as the peakedness z increases. Consequently, the capacity of the sender group is degraded. For example, in the case described by Figure 2, the capacity is reduced by about 14 percent as z increases from 1 to 2 and by about 25 percent as z increases from 1 to 3. The physical mechanism for this capacity reduction, as it is also for trunk groups, is that peaked streams offer the traffic in a more bunchy or variable arrival pattern which is more difficult to handle by the service system. The capacity reduction can be shown to increase as the size of the server group decreases.

Traffic variability is also caused by (systematic) within-hour time variations. This variability is due to the fact that the incoming rate of the offered load during the busy hour generally varies. The effect of time variations is similar to that of peakedness. Because the load-service curves are nonlinear, the probability of delay resulting from averaging of the delay at a high load and the delay at a low load is higher than the probability of delay for the average load. Thus, when senders are provided based on the delay expected for stationary Poisson traffic and the traffic is actually time varying, the senders will not be able to handle the traffic within the service constraint. Capacity reduction due to within-hour time variations increases with the size of the server group.

Both peakedness and time variations cause reductions in capacity which are difficult to distinguish between when only hourly measurements are taken. Thus, procedures for incorporating peakedness should be robust for both peakedness and time variations.

Since both peakedness and time variations reduce capacity, a procedure was developed and implemented to estimate an equivalent peakedness, comprising both peakedness and time variations, for crossbar tandems and put into practice. The practice infers a peakedness which could have caused the measured SADR.

Two types of exceptional measured SADR points can occur:

(i) The inferred peakedness is not physically reasonable

(ii) The SADR is too low to obtain a statistically reliable estimate of peakedness.*

Item (i) indicates to the user that some other phenomenon (possibly nontraffic related, e.g., an equipment problem) is significant. It thus can be a useful diagnostic tool. Item (ii) is fundamentally limiting to this type of peakedness estimate where the design criterion is to keep delays very small. This procedure defaults to a nominal value of peakedness when SADR is too low.

3. THE V/M TECHNIQUE

A method of estimating peakedness of traffic offered to a queueing system, applicable at low delays, was developed which is based on the variance-mean ratio of number of busy servers.

The estimation procedure can be described briefly here by referring to Figure 3, which applies to a (full availability) queueing system with independent exponential servers and with peaked input of peakedness z. M and V are the (equilibrium) mean and variance of the number of busy servers. Figure 3 has plots of V/M vs. ccs/server for different values of z. Thus, if we can measure V and M, we can infer z. The development leading to these curves follows.*

We are thus interested in calculating the variance-to-mean ratio (V/M) of the number of busy servers in the queueing system at an arbitrary instant of time in equilibrium. In this section we give results which characterize V/M as a function of the input parameters, particularly peakedness. We consider a renewal process offered to a group of N independent exponential servers in a queueing system.

Let F(t) denote the interarrival time distribution function* and Φ(s) the corresponding Laplace-Stieltjes transform. The peakedness of the process (z) is defined as the variance-to-mean ratio of the number of busy servers at an arbitrary instant of time in equilibrium when this process is offered to an infinite trunk group with independent, exponentially distributed holding times with mean 1/μ.++ The peakedness is given by

\[ z(μ) = \frac{1}{1-Φ(μ)} - \frac{1}{μ} \]

where \( λ \) is the mean arrival rate of the input

\[ \lambda = -\Phi'(0). \]

We offer this process to a queueing system with N independent, exponentially distributed service times with mean 1/μ. The service times are independent of the input stream and there can be no idle server in the presence of a waiting customer.\$

Using the results of Chapter 2 of [7], it can be shown that the variance-to-mean ratio of the number of busy servers in the queueing system is given by

\[ \frac{V}{M} \left( \mu_0, \phi(s) \right) = U_0 + U_1 + N(1-U_0) - M \]

where

\[ M = \frac{1}{\lambda^2} \]

is the erlang load offered to the N servers, and U_j satisfies

\[ Φ(Jμ_j) = 1 - Φ(Jμ_j) - \frac{N \left( 1 - Φ(Jμ_j) \right) - 1}{N(1-μ)} \]

with initial condition

\[ U_0 = 1 - \frac{A}{1-μ} \]

* The shape of these curves will be explained in Section IV.
++ P(0+) is assumed to be zero.
\$
At this point it is not necessary to assume more about the service discipline (e.g., first-come, first-served or random service).
Expressions for A and \( w_1 \), which are also parameters of the delay distribution for a first-come first-served discipline, are given in Chapter 2 of [7].

The specific renewal process of interest to us is the overflow process in Figure 1a. Observe that the Poisson process with rate \( \lambda_{eq} \) is offered to a group of \( N_{eq} \) trunks with mean service times \( u_1^{-1} \). The overflow is offered to a group of \( N \) servers each with mean service time \( u_2^{-1} \). For simplicity in subsequent sections, we replace this process with the interrupted Poisson process (i.p.p.) [9], as shown in Figure 1b.

Thus, given the input load and peakedness, we used the second of the models in Figure 1 to determine \( \phi(s) \). After evaluating \( A \) and \( w_1 \), one determines \( z_1 \) from (III-4) with initial condition (III-5). This leads to an evaluation of \( V/M \) from (III-2). In Section V we turn our attention to the statistical accuracy of the procedure for estimating \( z \) from individual samples of usage.

Before proceeding, we consider, in the next section, representative numerical results for \( V/M \) versus \( M \) with \( z \) as a parameter and discuss the observed behavior.

4. INTERPRETATION OF SHAPE OF V/M CURVES - \( z \) AS A FUNCTION OF \( \mu \)

We consider a two moment match i.p.p. input to a group of 40 servers (\( N = 40 \)) with a mean holding of 10 seconds \( (u_2^{-1} = 10) \). For the purpose of defining peakedness we use a mean trunk holding time of 180 seconds \( (u_1^{-1} = 180) \).

In Figure 3 we show the V/M curves for values of input peakedness ranging from \( z = 1 \) (Poisson) to \( z = 7 \). The abscissa is normalized load in ccs/sender = \( \frac{\lambda}{N u_2} \), \( 30 = \frac{36 M}{N} \).

In order to explain the shape of the V/M curves we back up and look at the behavior of peakedness with the mean service rate (\( \mu \)) on the infinite trunk group used in the peakedness calculation. As mentioned earlier, the peakedness of a renewal process with \( \phi(s) \) being the Laplace-Stieltjes transform of the interarrival time distribution, is given by (III-1)

\[
z(u) = \frac{1}{1-\phi(u)} - \frac{1}{\mu},
\]

where

\[
\frac{1}{\mu} = -\phi'(0).
\]

We observe that

\[
\lim_{\mu \to \infty} z(u) = 1
\]

The numbers are representative of what one would expect if we consider the 40 servers to be senders and \( \frac{1}{\mu} \) to correspond to the mean duration of a call.

† Observe that this result applies to deterministically spaced arrivals for which

\[
z(u) = \frac{1}{1-e^{-\mu T}} - \frac{1}{\mu T}
\]

with \( T \) the interarrival time although in most practical trunking cases, \( z(u) \) is close to \( z(0) = \frac{1}{2} \).

Since \( \phi(u) \to 0 \) as \( u \to \infty \). A physical explanation of this is that as \( u \to \infty \) the holding time \( \frac{1}{\mu} \to 0 \) so the probability of having more than one trunk busy becomes vanishingly small. Letting \( x(t) \) be the number of busy trunks on the infinite trunk group, we have

\[
E[x(t)] = E[x_1(t)] = P(x(t) = 1).
\]

Thus,

\[
z = \frac{\text{Var}(x(t))}{E(x(t))} = \frac{P(x(t)=2) - P(x(t)=1)}{P(x(t)=1)} \to 1 \text{ as } \frac{1}{\mu} \to 0.
\]

For the i.p.p. with switch parameters \( \lambda, w, v \), we have

\[
\phi(z) = \frac{\lambda(u+w)}{s + (\lambda+v+w)s + \lambda u}
\]

thus giving

\[
z(u) = 1 + \frac{v}{(1+w+\lambda)(\lambda+w)}.
\]

It is seen that \( z(u) \) monotonically decreases from \( z(0) \) to its limiting value 1. It should be noted that in most practical trunking situations we are operating near \( z(0) \).

We are now in a position to explain the shape of the \( V/M \) curves. We observe that for low load (initial part of curves) and large \( N \) the sender group looks like an infinite trunk group with \( u_2 \ll u_1 \) and \( \frac{V}{M} \ll z(\mu_2) < z(\mu_1) \) \( \approx z(0) \) of the input process. This is evidenced in Figure 3. As a matter of fact, we can use this reasoning to come up with a good approximation to \( V/M \) for low load and large \( N \). Consider a load of 5 ccs/sender into the 40 senders and

\[
\frac{V}{M} \approx z(\mu_2) = 1 + \frac{\lambda \gamma}{(\lambda+w+\mu_1)(\lambda+w)}
\]

For Poisson traffic \( \gamma = 0 \) and \( V/M \approx 1 \). For the other extreme we consider \( z = 7 \). Evaluation of the switch parameters \((h)\) gives \( \lambda = 2.87, \mu = .169 \) and \( \gamma = .578 \), which leads to (using (IV-8)):

\[
\frac{V}{M} \approx z(\mu_2) = 3.66
\]

which is in excellent agreement with the exact result shown in Figure 3. Thus for low loads and large \( N \), (IV-7) with \( u = \mu_2 \) serves as a good approximation to \( V/M \).

For the two moment match i.p.p. we can relate \( \lambda, w \) and \( \gamma \) to the mean and peakedness \( z(\mu) \) of the offered load.

By doing this, (IV-7) becomes

\[
z(\mu_2) = z(\mu_1) + \frac{1}{u_2} \left( 1 + 1 + 3z(\mu_1) - 1 \right)
\]

This is in excellent agreement with the exact result shown in Figure 3. Thus for low loads and large \( N \), (IV-7) with \( u = \mu_2 \) serves as a good approximation to \( V/M \).
where $M$ is the load $(\lambda/\mu_0)$. We note that for either $\mu_1 = \mu_2$ or $z(\mu_1) = 1$ we have (as expected) the result that $z(\mu_1) = z(\mu_2)$. Figure 4 is a plot of $z(\mu_2)$ versus ccs/sender. We note that for either $\mu_1 = \mu_2$ or $z(\mu_1) = 1$ we have (as expected) the result that $z(\mu_1) = z(\mu_2)$.

Note that on Figure 3 there are loci of points for which the

$$P(\text{delay} > 2.5 \text{ seconds}) = .005$$

and

$$P(\text{delay} > 2.5 \text{ seconds}) = .02. \uparrow$$

We see that for loads to the left of the 0.5 percent delay loads that $V/M$ is very well approximated by $z(\mu_2)$. Furthermore, for the 2 moment match i.p.p. we can easily solve for $z(\mu_1)$ from (IV-9) to obtain an approximate expression for $z(\mu_1)$ in terms of $V/M$. This results in the following approximation for low loads:

$$z(\mu_1) = 1 + \frac{1}{6} \left\{ -\frac{\mu_2}{\mu_1} M + 6 - \frac{V}{M} \right\} + \sqrt{\frac{\mu_2}{\mu_1} M + 6 - \frac{V}{M}} + 12 \left( \frac{V}{M} - 1 \right) \frac{\mu_2}{\mu_1} (1+M)^{1/2}$$

(IV-10)

5. ACCURACY OF ESTIMATE OF $z$

The previous sections assumed knowledge of $V$ and $M$. In practice, one would approximate $V$ and $M$ with a sample variance $SV$ and sample mean $SM$ based on $n$ samples.

Using the $SV$ and $SM$ instead of $V$ and $M$ introduces statistical errors which we will now analyze. In this analysis, we will assume independence of samples (the overall approximation has been checked by a simulation).

In this study, we found it adequate to work with a simple estimate of the standard deviation $\sigma(z)$, where $z$ is the estimate of $z$. Such an estimate is derived simply by looking at $\sigma(SV/SM)$ and seeking that fraction it represents of the distance between $V/M$ curves for different $z$'s.

* As the traffic intensity increases.

$\uparrow$ These probabilities are computed for first-come first-served queuing discipline.

Specifically,

$$\sigma(z) = \frac{\sigma(SV/SM)}{\left[ \frac{V}{M} \right]^{1/2}}.$$  (V-1)  

This type of estimate would be worth considering if most of the statistical variations were in $SV$ rather than in $SM$, i.e., most of the statistical fluctuations represented ordinate changes in the $V/M$ curves. To simply test this, we compared

$$\sigma_a(SV/SM)^2 = \sigma(SV/SM)^2 + \frac{V}{M} - 2.$$  (V-2)

(AN approximation obtained from (10.17) of [8]), with

$$\sigma_a(SV/SM)^2 = \sigma(SV/SM)^2 + \frac{V}{M} - 2.$$  (V-3)

(V-3) simply assumes negligible variation of the sample mean. $\nu_i$ is the $i$'th central moment of the number of busy servers; see (V-11), (V-12).

For further investigation of (V-1) we considered the two moment match i.p.p. where we can use (IV-10) for the rising portion of the $V/M$ curve with $SV$ and $SM$ replacing $V$ and $M$, as an estimate of $z$, i.e.,

$$\hat{z}(\nu_1) = 1 + \frac{1}{6} \left\{ -\frac{\mu_2}{\mu_1} SM + 6 - \frac{SV}{SM} \right\} + \sqrt{\frac{\mu_2}{\mu_1} SM + 6 - \frac{SV}{SM}} + 12 \left( \frac{SV}{SM} - 1 \right) \frac{\mu_2}{\mu_1} (1+SM)^{1/2}.$$  (V-4)

* This is consistent with assuming that $\hat{z}$ is the linear interpolate $\hat{z} = z_2 + a(z_1 - z_2)$ with

$$a = \frac{\frac{V}{M} z_2 - \frac{SV}{SM}}{\frac{V}{M} z_2 - \frac{SV}{SM}} , \frac{V}{M} z_1 < \frac{SV}{SM} < \frac{V}{M} z_2.$$

343-4
Since \( E(\hat{z}(\mu)) = \hat{z}(\mu_1) + o(n^{-1}) \) (see p. 231 of [8]), we are interested in the standard deviation, \( \sigma(\hat{z}(\mu)) \). Using (10.12) of [8], we have the following approximation:

\[
\sigma^2(\hat{z}(\mu)) = \frac{\frac{\sigma_2(\mu) \sigma_2(SM)}{3M}}{\sigma_2(SM)} + \frac{\frac{\sigma_2(\mu) \sigma_2(SV)}{3\sigma_2(SM)}}{2 \sigma_2(SV)} + \frac{\frac{\sigma_2(\mu) \sigma_2(SV)}{3}}{\sigma_2(SV) \sigma_2(SM)} \cdot \text{cov}(SM,SV) \tag{V-5}
\]

where the partial derivatives are obtained from (IV-10). From p. 230 or [8], we have (approximately)

\[
\begin{align*}
E(SV) &= V, \tag{V-6} \\
E(SM) &= M, \tag{V-7} \\
\sigma^2(SM) &= \frac{V}{n}, \tag{V-8} \\
\sigma^2(SV) &= \frac{S^2}{n}, \tag{V-10}
\end{align*}
\]

where \( s_i \) is the ith central moment of the number of busy servers, i.e.,

\[
\begin{align*}
\beta_3 &= E(k^3) - 3(V+M^2)M + 2M^3, \tag{V-11} \\
\beta_4 &= E(k^4) - 4E(k^2)M + 6(M+V^2)M^2 - 3M^4 \tag{V-12}
\end{align*}
\]

with

\[
\begin{align*}
E(k^2) &= M[U_0 + 3U_1 + 2U_2] + 6S^2(1-U_0) \tag{V-13} \\
E(k^4) &= M[U_0^3(1-U_0) + U_0 + 7U_1 + 12U_2 + 6U_3] \tag{V-14}
\end{align*}
\]

Comparisons of (V-5) with (V-1) and of (V-2) with (V-3) yielded the following. First of all, (IV-10) is as expected, a good approximation to the input \( z \) up until the curves flatten out. In these regions \( \hat{z}(z) \) is reasonably close to \( \hat{z}(\mu) \). Furthermore, \( \sigma_2(SV/SM) \) tracks \( \sigma_2(SV/SM) \) fairly well over varying parameters showing that it is the variance of \( SV \) which is most prominent. In particular, \( \sigma_2(SV/SM) \) is closest to \( \sigma_2(SV/SM) \) where \( \hat{z}(z) \) cannot be used as a benchmark thus indicating that \( \hat{z}(z) \) is a reasonable approximation over the entire range.

We also used a simulation to check the approximations used in the statistical analysis of the model. Figure 5 shows (V-1) normalized and plotted for \( N = 60, n = 360, V_1 = 1/180, \mu_2 = 0.1 \) and for \( z = 1 \) and \( z = 4 \). Also shown are some simulation results with approximate 95% confidence bands. It is seen that both the analysis and simulation indicate that with SADR below 2 percent, the percentage error in estimating \( z \) is roughly 10 percent or less. The analysis appears to slightly underestimate the errors as compared to the simulation. That this is due to the independence assumption is supported by simulation results for \( n = 60 \) where agreement is even better (decreasing \( n \) approximately raises \( \hat{z}(z) \) by the square root of the ratio of the \( n \)'s). * The simulated values of V/M agreed very well with calculated values for both \( n = 360 \) and \( n = 60 \).

6. CONCLUSION

The performance degrading effect of peakedness had been shown via a GI/M/N queueing analysis. This is applicable to some electromechanical machines. A more complex analysis is needed for some other machines, e.g., electronic switching machines. Other work has shown that peakedness is important in these machines also.

The philosophy of estimating \( z \) from usage measurements differs from using delays to estimate it in the following way. The usage is closely tied to \( z \), in fact, from the very definition of \( z \). On the other hand, delays are a direct measure of performance and it is desirable to provision based on actual performance. However, when the delays are low, as they normally should be, the delay-based \( z \)-estimate loses statistical reliability.

Since the V/M method does not directly measure performance, but rather a quantity related to performance, the robustness of the method to model assumptions is of interest. A sensitivity investigation made evident a degree of robustness for the method (robustness to nonmeasurable factors; measurable factors affecting the V/M curves can be taken into account). A robustness is obtained if the sender provisioning process is consistent with the \( z \)-estimation method. For example, results indicated that the 2-moment match i.p.p. yielded both higher delays and higher V/M's than a 3-moment match i.p.p. with the same rate and peakedness.† Thus, if we used a procedure based on the 2-moment match i.p.p. and the actual stream was a 3-moment match i.p.p., we would associate a higher-than-true \( z \) with the actual stream but correctly put in additional servers. More work is warranted relative to robustness to deviations in assumptions used in deriving the V/M method.

It is also possible to estimate peakedness of a traffic stream based on the variance-mean ratio of number of arrivals. However, this is even further removed from the actual delay performance and suffers from not being at all based on the response of the system.

Finally, we remark that the V/M method requires within-hour usage samples, which are not standard in many systems.

7. ACKNOWLEDGMENTS

We would like to thank those who were involved in the original crossbar tandem capacity study: S. G. Gannon (who investigated factors causing the discrepancies mentioned in Section I), R. P. Marzec and R. E. Miller (who worked on the procedure of Section II). The comments of M. S. Hall, C. A. LaPadula, J. J. O'Connor, and R. T. Burnett of A.T.&T., were also useful.

* Also, the dependence effect appears to increase with load for \( n = 360 \), which is consistent with the increased correlations with load reported for delay systems in [9] (see Figure 16 of [9]).

† Peakedness does not yield a full characterization of a traffic stream; it can only be used as an approximation for a class of streams.
REFERENCES


FIGURE 4
PEAKEDNESS -LOAD CHARACTERISTIC

FIGURE 5
COMPARISON OF ACCURACY RESULTS WITH SIMULATION