Analysis of Rate Assignments in a Rate-Based Flow Control Scheme

Jane M. Simmons

Submitted to the Department of Electrical Engineering and Computer Science in

Partial Fulfillment of the Requirements for the Degree of

Master of Science

at the

Massachusetts Institute of Technology

February 1990

Thesis Advisor: Robert Gallager

ABSTRACT

On high-speed networks, the round trip delay from source to intermediate node and back is relatively large compared with packet transmission times; thus, the window-based method of flow control does not react quickly enough to prevent congestion and large delay. An alternative method is to employ a rate-based flow control scheme, in which each node is assigned a maximum rate and an average rate of transmission. The purpose of this thesis is to develop a criterion for evaluating how well a given assignment of rates will perform in preventing congestion in the network. The criterion to be used is the probability that the instantaneous aggregate arrival rate at any one node is greater than the capacity of the outgoing link of that node. In order to simplify the development of an expression for this probability, an on/off model is used for the sources, and only a subset of the network is considered - a set of sources feeding into one intermediate node with an outgoing link. This thesis examines how the assigned rates and the burstiness of the sources affect this probability, and analyzes the effect of adding a new source to the system.

1. INTRODUCTION

The purpose of flow control in data networks is to throttle the users under heavy load conditions in order to prevent congestion and large delays from developing. The control should be implemented in a manner which is fair to the users, and which does not unnecessarily restrain them when the network is lightly loaded. This thesis will specifically deal with flow control between sources and intermediate nodes, but much of it also applies to flow control across a broader network.

The most widely used form of flow control is based on windows, where the rate at which a source sends packets depends on both the size of a window and on the rate of feedback from the network to the source. On high-speed networks, the round trip delay from source to an intermediate node and back is relatively large compared with packet transmission times, so that the windowing method does not react quickly enough to prevent congestion and large delay. Additionally, high-speed networks will need to accommodate a wide range of source transmission speeds and windows do not perform well under these conditions. Section 2.1 further details the problems with windows on high-speed networks.

An alternative form of flow control is to assign a maximum transmission rate and an average transmission rate to each source. The network, with knowledge of the burstiness and desired rates of all nodes that feed into any link, assigns the rates such that the probability of congestion developing is very small. This method essentially provides a guaranteed rate without dependence on feedback from the intermediate nodes, and thus is more appropriate than windows for use on high-speed networks. Section 2.2 describes more fully the operation of rate-based flow control.

The purpose of this thesis is to develop a criterion for evaluating how well a given assignment of rates will perform in preventing congestion in the network. The criterion to be used is the probability that the instantaneous aggregate arrival rate at any one node is greater than the capacity of the outgoing link of that node. In order to simplify the development of an expression for this probability, an on/off model is used for the sources, and only a subset of the network is considered - a set of sources feeding into one intermediate node with an outgoing link. The model is fully described in Section 3. The rationale for using this overrun probability and the derivation of an expression for it are provided in Section 4. Section 5 examines how the assigned rates and burstiness of the sources affect this probability. For a given average traffic rate, this provides insight into the tradeoff between allowing a source to be very bursty with a high maximum data rate, and forcing it to be less bursty with a low maximum transmission rate. The effect on the overrun probability when a new source is added to the system is also analyzed.

2. BACKGROUND

2.1 Problems With Windows on High-Speed Networks

In window-based flow control, each source has a maximum limit W on the number of unacknowledged packets, which can be outstanding at a given time. In one implementation, the window of size W is established between the source and the destination, and is referred to as an end-to-end window. Alternatively, a window could be established between each successive node along the data path (i.e., between the source and the first intermediate node, between the first and second intermediate node, etc.); this method is referred to as node-by-node windows. In the discussion below, the node-by-node window method is assumed; more specifically, the window mechanism between the source and the intermediate node is examined.

As shown below, windows do not perform well on high-speed networks, largely due to the relatively long delay before the source receives feedback. With end-to-end windows, the source depends on feedback from the destination, whereas with node-by-node windows, the source receives feedback from the intermediate node to which it directly transmits. The delay in waiting for a response will obviously be longer for end-to-end windows. Thus, the issues of congestion that arise due to delayed feedback, as detailed below for node-by-node windows, are an even more serious problem with end-to-end windows.

After a source has sent the maximum number of packets, it must wait for an acknowledgement (ACK) from the intermediate node before any new packets can be sent. Assume that the round trip delay from source to intermediate node and back is τ, where τ includes any propagation and processing delays. If the desired average transmission rate of the source is R, the window size W must be greater than $(R)(\tau)$ in order for the source not to be unnecessarily throttled under light load conditions. When the data is actually sent, it is transmitted at the maximum capacity r of the source's access link, where r is assumed greater than R.

On the type of high-speed networks envisioned for the future, typical values for R, r, τ, and packet length might be 50 Mbit/s, 100 Mbit/s, 50 msec, and 1 Kbit, respectively. This would force a minimum window size of W = 2,500 packets. Consider an example where there are 10 sources with these parameters, and all of them have a full window's worth of data to send. Assume the intermediate node to which they are transmitting has an outgoing link with capacity 200 Mbit/s.

Assuming that all sources start to transmit at time 0, data will arrive at the intermediate node at a rate of 1 Gbit/s; thus, its queue will grow at a rate of 800 Mbit/s, for a duration of 25 msec (the sources' windows will be exhausted after 25 msec of transmitting at 100 Mbit/s). Thus, the intermediate node must have a large buffer (at least 20 Mbits); otherwise, data will be lost.

Over the next 25 msec., the queue at the intermediate node dissipates to 15 Mbits. During this time, the sources receive feedback, in the form of delayed ACKs, indicating that they are transmitting too fast for the intermediate node. Starting at time 50 msec., ACKs will be received at a rate of 20 Mbit/s, which will then become the effective transmission rate of each source. If all of the sources have more data to send, the aggregate arrival rate at the intermediate node will be 200 Mbit/s. The queue size will remain 15 Mbits, and there will be a large queuing delay of 75 msec.

The intrinsic problem is that the same size window that is used under light load conditions is also being used in the case of heavy congestion. Thus, by choosing the window size to accommodate the sources under light traffic, the sources are not throttled enough when congestion develops. Too many packets are still permitted to be sent by the source, and the queue at the intermediate node is not given a chance to fully dissipate. The problem is exacerbated on high-speed networks where the window size may be very large, and hence the queue length and associated queuing delay may be very large.

Ideally, the window size should be dynamically adjusted according to the level of congestion in the network. This is a difficult control problem due to the relatively large delay before the source receives feedback from the intermediate node. The source would have to adjust its windows based on 'old' information, which may result in unnecessary oscillations in window size. Additionally, changing the window size of one source affects the traffic level of all of the intermediate nodes in its path, which in turn affects the congestion experienced by other sources that send data through these nodes. Therefore, the window sizes of the sources and intermediate nodes should be adjusted collectively; otherwise, the adjustments may counteract each other or result in overcompensation, and the congestion will be shifted rather than lessened. This is difficult to accomplish in practice due to the large scale of the problem.

Furthermore, when queuing delays arise, the window mechanism will often time-out waiting for an ACK. A packet will be interpreted as being lost if an ACK has not been received by a certain time, and will be retransmitted. (The value of W limits the number of different unacknowledged packets there can be, not the number of 'copies' of these packets.) It is likely that several sources will simultaneously experience the delay in feedback, and resend packets unnecessarily. This will worsen the backlog at the intermediate node and cause further queuing delays. This could result in low throughput and high delay.

The window mechanism may also react unwisely when the sources have desired rates of transmission that differ by orders of magnitude (a situation likely to arise on high-speed networks). For example, assume that there are 20 sources having R = 50 Mbit/s and r = 75 Mbit/s. Assume that there is an additional source with R^* = 10 Kbit/s and r^* = 15 Kbit/s. If the packet length and round trip delay are the same for all users, then $W = (5{,}000)(W^*)$. Assume that the sources feed into an intermediate node having an outgoing link of capacity 200 Mbit/s. After congestion develops, the effective rate of the sources will be in proportion to their window sizes:

$$R_{\text{eff}} = 10 \text{ Mbit/s}, \quad R^*_{\text{eff}} = 2 \text{ Kbit/s}$$

Thus, the window mechanism reacts to the congestion by reducing all source rates by the same proportion. Rather than reducing R^* to such a low rate, however, it would be more reasonable to leave R^* at 10 Kbit/s, and reduce the other sources to a rate slightly less than 10 Mbit/s. This would satisfy the needs of the low-rate source at very little expense to the other users.

2.2. Rate-Based Flow Control

In rate-based flow control, there are three parameters that are specified to control the data flow, as compared to the single parameter (i.e., window size) that is used in the window-based scheme. First, each source is assigned an average transmission rate. These rates should be assigned such that the aggregate average arrival rate into any intermediate node is less than the capacity of the outgoing link of the node. The sources are essentially guaranteed of achieving their assigned average data rate.

Each source is also assigned a maximum rate of transmission (which is often simply the capacity of the access link), and a limit on the length of time it can send data at that rate. These parameters control the ability of the source to send a burst of data. In window-based flow control, the burst size is essentially the window size, which is determined by the rate of transmission. In the rate-based method, the maximum burst size can be set independently of the rates, thus allowing more control of the data flow.

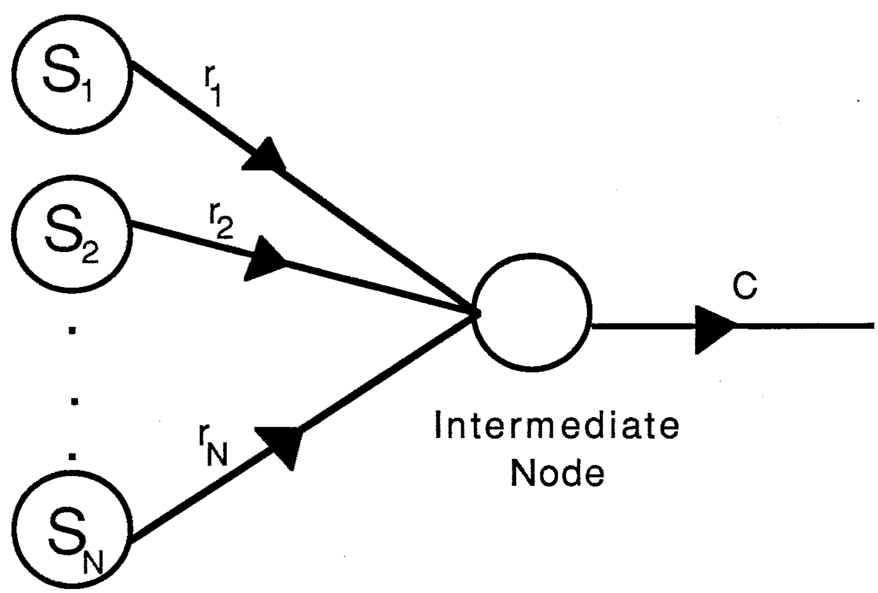
There are a variety of ways in which rate-based flow control can be implemented. In one scheme, a source generates tokens at a fixed rate and can only send a packet when there is a token available. If the source temporarily has no data to send, the tokens are allowed to build up, up to the burst size limit. The source can then send a burst of packets at its maximum rate.

In general, when a call is set up, the source specifies the average traffic rate it expects to have. The network decides whether a call with this average rate can be accepted, or negotiates with the source to determine a more acceptable rate. It then assigns the source a maximum transmission rate and a maximum burst size that are appropriate for the type of data the source is sending. The parameters should be assigned in such a way that there is only a small probability of the instantaneous aggregate arrival rate at any intermediate node being larger than the capacity of its outgoing link. This keeps queuing delay small and minimizes the chance of packets being dropped due to buffer overflow. The expressions derived in Section 4 below provide insight into evaluating how well a given set of rates will meet this goal.

If retransmissions are necessary, they are sent using the same rate-based scheme, and thus they are not really 'extra' traffic to the system as they are in the window-based scheme. This should prevent the system from entering a situation in which high queuing delays and retransmissions feed off of each other. Additionally, fine tuning of the rates can be done through feedback from the intermediate node, but as with windows, there will be a delay before such changes take effect.

3. DEFINITIONS

The rate-based scheme will be analyzed using the following system as a model:



S_i - i^{th} Source

C - capacity of outgoing link from intermediate node

N - number of sources

r_i - assigned maximum rate of i^{th} source

p_i - probability the i^{th} source is in the on-state

$(p_i)(r_i)$- assigned average rate of i^{th} source

$\Lambda = \sum_i (r_i)(s_i)$ $s_i = 1$ if i^{th} source is in the on-state

$s_i = 0$ if i^{th} source is in the off-state

$E[\Lambda] = \sum_i (p_i)(r_i)$ = average aggregate arrival rate

There are *N* mutually independent sources feeding into an intermediate node that has an outgoing link of capacity C. An on/off model will be used to model the burstiness of the sources. Source *i* is assumed to be in the on-state with probability p_i, and the off-state with probability $1-p_i$. When a source is on, it is assumed that it will transmit at its maximum rate, r_i. When it is off, the source will transmit nothing. With this definition, $(p_i)(r_i)$ corresponds to the average transmission rate and r_i corresponds to the maximum rate, as described in the previous section.

The criterion to be used to analyze the model, as explained below, does not depend on the time statistics of the on/off model. Thus, a specific distribution need not be assumed for the on time or the off time, and no limit is placed on the maximum burst size.

Λ represents the instantaneous aggregate arrival rate at the intermediate node. It is assumed that $E[\Lambda]$ is less than C.

4. CRITERION TO EVALUATE RATE ASSIGNMENTS

The criterion that will be used to analyze different allocations of rates is $P(\Lambda>C)$ - the probability that the instantaneous aggregate arrival rate at the intermediate node is greater than C. The rationale for this is that it provides quantitative insight into the possible queuing delay that might develop and the probability of the buffers at the intermediate node being overrun. This criterion is made simpler by the fact that it does not involve analysis of time; i.e., how long the system is in a 'bad' state. Additionally, if the rates of the sources and the outgoing link were scaled up by the same factor, while the buffer size were held constant, then the probability of Λ being greater than C approaches the probability of the buffer being overrun as the scale factor approaches $\propto$.

An asymptotic expression for $P(\Lambda>C)$ will be used, which is based on a derivation in appendix 5A of Reference 3 and is summarized below. When $E[\Lambda]$ is relatively close to C, this becomes a simple Gaussian approximation to $P(\Lambda>C)$.

4.1 Asymptotic Expression for P(Λ>C)

Let R_i be the random variable representing the rate of the i^{th} source. The R_i's are assumed to be independent. The probability density for R_i is:

$$P(R_i = r_i) = p_i\ ,\ \ P(R_i=0) = 1-p_i$$

There are *N* sources, where *N* is assumed to be a large number.

The random variable Λ equals $\sum_i R_i$ and $E[\Lambda] = \sum_i (p_i)(r_i)$.

The semi-invariant moment-generating function of R_i is defined as:

$$\mu_i(s) \ = \ \ln E[\exp(sR_i)] \ = \ \ln [p_i \exp(sr_i) + (1-p_i)]$$

In this expression and in the expressions below, *s* is an arbitrary parameter, and it is assumed that it can only take on non-negative values. The first and second derivatives of $\mu_i(s)$ are:

$$\mu_i'(s) = \frac{p_i r_i \exp(sr_i)}{E[\exp(sR_i)]} \qquad \mu_i''(s) = \frac{p_i r_i^2 \exp(sr_i)}{E[\exp(sR_i)]} - [\mu_i(s)]^2$$

The semi-invariant moment-generating function of Λ is:

$$\mu_\Lambda(s) = \ln E[\exp(s\Lambda)] = \sum_i \mu_i(s) \qquad (1)$$

Define a set of tilted random variables, $\tilde{R}_i$ and $\tilde{\Lambda}$, where:

$$P(\tilde{R}_i=r_i) = \frac{\exp(sr_i)}{E[\exp(sR_i)]} P(R_i=r_i)\ , \quad P(\tilde{R}_i=0) = \frac{1}{E[\exp(sR_i)]} P(R_i=0)\ , \quad \tilde{\Lambda} = \sum_i \tilde{R}_i$$

The means and variances of the tilted variables are:

$$E[\tilde{R}_i] = \mu_i'(s) \qquad E[\tilde{\Lambda}] = \sum_i \mu_i'(s) = \mu_\Lambda'(s)$$

$$\mathrm{var}[\tilde{R}_i] = \mu_i''(s) \qquad \mathrm{var}[\tilde{\Lambda}] = \sum_i \mu_i''(s) = \mu_\Lambda''(s) \qquad (2)$$

The probability that $\tilde{\Lambda}$ equals any value λ is:

$$P(\tilde{\Lambda}=\lambda) = \sum_{r_1..r_N} \left(\prod_i P(\tilde{R}_i=r_i) \right)$$

(The sum is taken over all r_1 ... r_N, such that r_1 +..+ $r_N = \lambda$)

$$= \sum_{r_1..r_N} \left(\prod_i \frac{\exp(sr_i)\, P(R_i=r_i)}{E[\exp(sR_i)]} \right)$$

$$= \frac{\exp(s\lambda)}{\prod_i E[\exp(sR_i)]} \sum_{r_1..r_N} \left(\prod_i P(R_i=r_i) \right)$$

$$= \frac{\exp(s\lambda)}{\prod_i E[\exp(sR_i)]} P(\Lambda=\lambda)$$

Using equation (1), $\prod_i E[\exp(sR_i)] = E[\exp(s\Lambda)] = \exp(\mu_\Lambda(s))$

Thus, $P(\tilde{\Lambda}=\lambda) = \exp(s\lambda-\mu_\Lambda(s))\ P(\Lambda=\lambda)$

and $P(\Lambda=\lambda) = \exp(\mu_\Lambda(s)-s\lambda))\ P(\tilde{\Lambda}=\lambda)$

We are interested in determining P(Λ>C), where:

$$P(\Lambda>C) = \sum_{\lambda>C} P(\Lambda=\lambda)$$

$$= \sum_{\lambda>C} \exp(\mu_\Lambda(s)-s\lambda)\ P(\tilde{\Lambda}=\lambda)$$

$$= \exp(\mu_\Lambda(s)) \sum_{\lambda>C} \exp(-s\lambda)\, P(\tilde{\Lambda}=\lambda) \quad . \qquad (3)$$

$\tilde{\Lambda}$ is the sum of a large number of independent, but not identically distributed, random variables. It is reasonable to assume that for any values of C and N, no single r_i would be larger than about (10 C/N); i.e., no one source dominates $\tilde{\Lambda}$. With this assumption, the variance of $\tilde{\Lambda}$ grows toward $\propto$ as N and C become large, while any individual var[$\tilde{R}_i$] is bounded. Thus, the Lindeberg condition is satisfied (see page 256 of Reference 7) and the distribution of $\tilde{\Lambda}$ can be approximated by the Gaussian distribution:

$$P(\lambda < \tilde{\Lambda} < \lambda + \Delta\lambda) \approx \frac{\Delta\lambda}{\sqrt{2\pi\,\mathrm{var}[\tilde{\Lambda}]}} \exp \frac{-(\lambda - E[\tilde{\Lambda}])^2}{2\,\mathrm{var}[\tilde{\Lambda}]} \qquad (4)$$

(for very small $\Delta\lambda$)

Using this approximation and letting the sum go to an integral, Equation (3) becomes:

$$P(\Lambda > C) \approx \exp(\mu_\Lambda(s)) \int_C^\infty \exp(-s\lambda) \frac{1}{\sqrt{2\pi\,\mathrm{var}[\tilde{\Lambda}]}} \exp \frac{-(\lambda - E[\tilde{\Lambda}])^2}{2\,\mathrm{var}[\tilde{\Lambda}]}\, d\lambda \qquad (5)$$

The Gaussian approximation in equation (4) is best when λ is close to $E[\tilde{\Lambda}]$. $E[\tilde{\Lambda}]$ is a function of s and can take on values between $\sum_i (p_i)(r_i)$ and $\sum_i (r_i)$. Assuming C is between these values, s can be chosen such that $E[\tilde{\Lambda}]$ equals C. With this value of s, which will be denoted by s^*, the approximation in equation (5) should be very good. The exponential term in equation (5) is decreasing in λ so that there is a very small contribution to the integral when λ is much greater than C.

Thus, with $s = s^*$:

$$P(\Lambda > C) \approx \exp(\mu_\Lambda(s^*)) \int_C^\infty \exp(-s^*\lambda) \frac{1}{\sqrt{2\pi\,\mathrm{var}[\tilde{\Lambda}]}} \exp \frac{-(\lambda - C)^2}{2\,\mathrm{var}[\tilde{\Lambda}]}\, d\lambda \qquad (5a)$$

$$= \exp(\mu_\Lambda(s^*) - s^*C) \int_C^\infty \exp(-s^*(\lambda - C)) \frac{1}{\sqrt{2\pi\,\mathrm{var}[\tilde{\Lambda}]}} \exp \frac{-(\lambda - C)^2}{2\,\mathrm{var}[\tilde{\Lambda}]}\, d\lambda$$

$$= \exp(\mu_\Lambda(s^*) - s^*C) \int_0^\infty \frac{1}{\sqrt{2\pi\,\mathrm{var}[\tilde{\Lambda}]}} \exp -\left(s^*\lambda + \frac{\lambda^2}{2\,\mathrm{var}[\tilde{\Lambda}]} \right) d\lambda \qquad (6)$$

$$= \exp(\mu_\Lambda(s^*) - s^*C) \exp\left(\frac{s^{*2}\,\mathrm{var}[\tilde{\Lambda}]}{2} \right) \int_0^\infty \frac{1}{\sqrt{2\pi\,\mathrm{var}[\tilde{\Lambda}]}} \exp \frac{-(\lambda + s^*\mathrm{var}[\tilde{\Lambda}])^2}{2\,\mathrm{var}[\tilde{\Lambda}]}\, d\lambda$$

$$= \exp(\mu_\Lambda(s^*) - s^*C) \exp\left(\frac{s^{*2}\,\mathrm{var}[\tilde{\Lambda}]}{2} \right) \Phi(-s^* \sqrt{\mathrm{var}[\tilde{\Lambda}]}) \qquad (7)$$

where $\Phi(x)$ represents the unit normal Gaussian cumulative distribution function (CDF).

As a test of how good this approximation is, consider N identical sources, with parameters $p = 0.8$ and $r = 0.15$, feeding into a node with an outgoing link of $C = 129.975$. (The units on r and C do not matter as long as they are the same.) The actual CDF for the associated Λ has steps in it at intervals of 0.15, while the Gaussian CDF is a smooth function. Therefore, C was chosen to be in the middle of an interval, where the Gaussian CDF would be the best approximation to the actual CDF. If the values for r had been arbitrary and non-lattice (i.e., if they could not all be expressed as an integer multiple of some constant), then the intervals between steps in the actual CDF would be very small, and the Gaussian CDF should be a good approximation for any C.

Since the N sources are identical, the exact value of P(Λ>C) can be calculated using the sum of binomial terms. With N equal to 1,000:

$$P(\Lambda>C) = \sum_{n=867}^{1000} \frac{N!}{n!(N-n)!} (.8^{n}) (.2^{N-n}) = 1.8 \times 10^{-8}$$

The approximation in equation (7) yields the same result.

Equation (7) can be difficult to manipulate due to the Φ(x) term. In equation (6), the term $\dfrac{\lambda^2}{2\,\mathrm{var}[\tilde{\Lambda}]}$ is always greater than 0, and thus an upper bound for this approximation to P(Λ>C) is:

$$\exp(\mu_\Lambda(s^*) - s^*C) \int_0^\infty \frac{1}{\sqrt{2\pi\,\mathrm{var}[\tilde{\Lambda}]}} \exp -(s^*\lambda)\, d\lambda$$

$$= \exp(\mu_\Lambda(s^*) - s^*C)\, \frac{1}{s^*\sqrt{2\pi\,\mathrm{var}[\tilde{\Lambda}]}} . \qquad (8)$$

The term s*λ is also always greater than 0, and thus a second upper bound for this approximation to P(Λ>C) is:

$$\exp(\mu_\Lambda(s^*) - s^*C) \int_0^\infty \frac{1}{\sqrt{2\pi\,\mathrm{var}[\tilde{\Lambda}]}} \exp\frac{-\lambda^2}{2\mathrm{var}[\tilde{\Lambda}]}\, d\lambda$$

$$= .5 \exp(\mu_\Lambda(s^*) - s^*C) . \qquad (9)$$

When N and s^* are large enough (s^* becomes larger the smaller E[Λ] is compared to C), the bound of equation (8) is tighter than the bound of equation (9), and in fact is a good approximation to P(Λ>C). Thus, when E[Λ] is much less than C, it will be assumed that:

$$P(\Lambda>C) \approx \exp(\mu_\Lambda(s^*) - s^*C)\,\frac{1}{s^*\sqrt{2\pi\,\mathrm{var}[\tilde{\Lambda}]}}$$

$$= \exp(\mu_\Lambda(s^*) - s^*C)\,\frac{1}{s^*\sqrt{2\pi\,\mu_\Lambda''(s^*)}} \qquad (10)$$

Note that the exponential term is identical to the Chernoff bound. (In the Chernoff bound, the same s^* is chosen to make the bound the tightest.) Thus, equation (10) is the Chernoff bound multiplied by a coefficient term. This approximation will be referred to as the asymptotic expression for P(Λ>C). Substituting values for $\mu_\Lambda(s^*)$ and $\mu_\Lambda''(s^*)$, equation (10) can also be expressed as:

$$\left(\frac{\exp(-s^*C)}{s^*\sqrt{2\pi\sum_i \frac{(p_i r_i^2 \exp(s^* r_i))(1-p_i)}{[p_i \exp(s^* r_i)+(1-p_i)]^2}}}\right)\prod_i \left(p_i \exp(s^* r_i)+(1-p_i)\right)$$

4.2 Gaussian Approximation to P(Λ>C)

When E[Λ] is close to C, the value of s^* is very close to 0. Using the Gaussian approximation as before and letting $s = 0$, equation (5) becomes:

$$P(\Lambda>C) \approx \int_C^\infty \frac{1}{\sqrt{2\pi\,\mathrm{var}[\tilde{\Lambda}]}}\exp\frac{-(\lambda-E[\tilde{\Lambda}])^2}{2\,\mathrm{var}[\tilde{\Lambda}]}\,d\lambda \quad . \qquad (11)$$

This approximation is in general not as good as equation (5a), and therefore not as good as equation (7), because $s = 0$ is being used rather than the proper value of s^*. With $s = 0$, C is greater than E[$\tilde{\Lambda}$], so that the integral in equation (11) is taken over a region beginning to the right of E[$\tilde{\Lambda}$], and the Gaussian approximation is not as good as λ moves away from the mean.

With s = 0, the variables are no longer tilted and equation (11) is equivalent to:

$$P(\Lambda>C) \approx \int_{C}^{\infty} \frac{1}{\sqrt{2\pi \, var[\Lambda]}} \exp \frac{-(\lambda - E[\Lambda])^2}{2 \, var[\Lambda]} \, d\lambda$$

$$= \Phi\left(\frac{-C + E[\Lambda]}{\sqrt{var[\Lambda]}}\right) \qquad (12)$$

This is just the Gaussian approximation to the distribution of the original untilted variables. This approximation to P(Λ>C) will be used whenever E[Λ] is close to C.

Graph 1 plots the equation (7) approximation to P(Λ>C) for various values of C and arbitrary fixed values of p_i and r_i. The p_i's used are uniformly distributed between 0 and 1; the r_i's are uniformly distributed between 0 and about 1% of C. C is in the range of 600 Mbit/s to 750 Mbit/s. The number of sources used is 1,000. (These same values are also used for all of the simulations below.)

Graph 2 compares the asymptotic expression and the Gaussian expression to the approximation of equation (7). As can be seen from this graph, the Gaussian expression is best used when C is within about two standard deviations of E[Λ]. When C is larger than this, the asymptotic expression should be a better approximation to P(Λ>C).

Two different types of approximations were made in deriving the asymptotic and Gaussian expressions from equation (5). The asymptotic expression makes use of the proper value of s^*, yielding equations (5a) and (6). In evaluating the integral of equation (6), however, a term which is always less than unity is ignored. Thus, as Graph 2 shows, the asymptotic expression is always greater than the approximation of equation (7). In the Gaussian expression, $s = 0$ was used rather than the proper s^*. This yielded equation (11) rather than equation (5a). However, the integral in equation (11) is evaluated without making any approximations. For the parameters used in Graph 2, the Gaussian expression yields an underestimate, as compared to the approximation of equation (7).

5. ANALYSIS

5.1 Asymptotic Expression

The two components of the asymptotic expression for P(Λ>C), the Chernoff bound term and the coefficient term, are analyzed separately in this section (refer to equation (10)).

The Chernoff bound term is: $\exp(\mu_\Lambda(s) - sC)$. A graphical representation of this term is shown in Graph 3. The curve is a sample plot of $\mu_\Lambda(s)$, where

$$\mu_\Lambda(s) = \sum_i \mu_i(s) = \sum_{i=1} \ln[p_i \exp(sr_i) + (1-p_i)]$$

The value of s that is used in equation (10) and that also yields the tightest Chernoff bound is s^*, where $\mu_\Lambda'(s^*) = C$. Graphically, this can be found where a line of slope C is tangent to the curve of $\mu_\Lambda(s)$. The intercept of this line with the y-axis yields the term $(\mu_\Lambda(s^*) - s^*C)$. This is the exponent in the Chernoff bound term.

First, we will examine how this term changes as the value of C changes (refer to Graph 4). (The effect of modifying C is important because when a particular source is on, the effective C for the remaining sources appears smaller. This will be used later on to analyze the effect on $P(\Lambda>C)$ of just one source.) Let the new value of C be represented by $\overline{C}$, and let the new optimal value of s be $\overline{s^*}$. Thus, $\mu_\Lambda'(\overline{s^*}) = \overline{C}$. Additionally, let $\Delta C = \overline{C} - C$, and let $\Delta s = \overline{s^*} - s^*$. As can be seen from Graph 4, if ΔC is less than zero, then Δs is also less than zero.

If only second order terms and lower are used to represent $\mu_\Lambda(s^*)$, then:

$$\mu_\Lambda(\overline{s^*}) = \mu_\Lambda(s^*) + \Delta_s\, \mu_\Lambda'(s^*) + (\Delta_s{}^2/2)\, \mu_\Lambda''(s^*) \qquad (13)$$

$$\mu_\Lambda'(\overline{s^*}) = \mu_\Lambda'(s^*) + \Delta_s \mu_\Lambda''(s^*) = \overline{C} \qquad (14)$$

Combining Equation (14) with $\mu_\Lambda'(s^*) = C$ yields:

$$\Delta_s = \frac{\Delta_C}{\mu_\Lambda''(s^*)} \quad \text{and thus} \quad \overline{s^*} = s^* + \frac{\Delta_C}{\mu_\Lambda''(s^*)} \quad . \qquad (15)$$

The new exponent in the bound term will be $(\mu_\Lambda(\overline{s^*}) - \overline{s^*}\,\overline{C})$. Using Equations (13) and (15), and representing $\overline{C}$ by $(\mu_\Lambda'(s^*) + \Delta C)$ yields:

$$\begin{aligned}
\text{new exponent} &= \mu_\Lambda(s^*) + \Delta_s\, \mu_\Lambda'(s^*) + (\Delta_s{}^2/2)\, \mu_\Lambda''(s^*) - \overline{s^*}\,\overline{C} \\
&= \mu_\Lambda(s^*) + \frac{\Delta_C C}{\mu_\Lambda''(s^*)} + \frac{1}{2}\left(\frac{\Delta_C}{\mu_\Lambda''(s^*)}\right)^2 \mu_\Lambda''(s^*) \\
&\qquad - \left(s^* + \frac{\Delta_C}{\mu_\Lambda''(s^*)}\right)(\mu_\Lambda'(s^*) + \Delta_C) \\
&= \mu_\Lambda(s^*) - s^*C - s^*\Delta_C - \frac{\Delta_C{}^2}{2\mu_\Lambda''(s^*)} \\
&= (\text{original exponent}) - s^*\Delta_C - \frac{\Delta_C{}^2}{2\mu_\Lambda''(s^*)}
\end{aligned}$$

Thus, the new Chernoff bound term, as a function of ΔC, can be approximated by:

$$\text{new Chernoff bound} = (\text{original Chernoff bound})\cdot\exp\left(-s^*\Delta_C - \frac{\Delta_C^2}{2\mu_\Lambda''(s^*)}\right) \qquad (16)$$

Graph 5 shows how the Chernoff bound changes when C is varied, for fixed, arbitrary values of p_i and r_i. Graph 6 compares the approximation of equation (16) with the actual calculation of the Chernoff bound term for an arbitrary choice of C, for a range of ΔC within about ±5% of C.

The coefficient term of the asymptotic approximation to P(Λ>C) is:

$$\frac{1}{s\sqrt{2\pi\mu_\Lambda''(s)}}$$

The same value of *s* which optimizes the Chernoff bound term is used here.
Again, it will be assumed that $\mu_\Lambda''(s^*)$ remains constant even as s* changes. Therefore, using Equation (15) for $\overline{s^*}$, if C changes by an amount ΔC, then:

$$\text{new coefficient} = (\text{original coefficient})\left(\frac{s^*}{s^* + \dfrac{\Delta_C}{\mu_\Lambda''(s^*)}}\right). \qquad (17)$$

Graph 7 is a plot of the coefficient term for various values of C and fixed, arbitrary values of p_i and r_i. Graph 8 compares the approximation of equation (17) to the actual calculated coefficient for an arbitrary C and ΔC in the range ± 5% of C.

Combining equations (16) and (17) yields:

$$P(\Lambda > C+\Delta_C) \approx P(\Lambda > C)\left(\frac{s^*}{s^* + \dfrac{\Delta_C}{\mu_\Lambda''(s^*)}}\right)\exp\left(-s^*\Delta_C - \frac{\Delta_C^2}{2\mu_\Lambda''(s^*)}\right) \qquad (18)$$

To simplify the equations below, let Q(ΔC) represent the term:

$$\left(\frac{s^*}{s^* + \dfrac{\Delta_C}{\mu_\Lambda''(s^*)}}\right)\exp\left(-s^*\Delta_C - \frac{\Delta_C^2}{2\mu_\Lambda''(s^*)}\right)$$

for given values of s* and $\mu_\Lambda''(s^*)$. Then Equation (18) can be written: P(Λ > C+ΔC) = P(Λ > C) Q(ΔC).
For negative ΔC, the approximation of equation (18) is only valid if:

$$\frac{|\Delta C|}{\mu_\Lambda''(s^*)} < s^*$$

Otherwise it yields a probability less than 0. With this restriction for negative ΔC, the exponent term is always greater than 0 if ΔC is less than 0. Thus, as expected, Q(ΔC) is greater than 1 if ΔC is negative; it is less than 1 if ΔC is positive.

Graph 9 shows how P(Λ>C) changes with C, as calculated using the asymptotic expression. Graph 10 gives an example of how closely the approximation of equation (18) follows the value of P(Λ > C+ΔC) calculated using the asymptotic expression.

Equation (18) provides insight into the effect of changing the capacity of the intermediate node on the probability of overrunning this node. For the simulations which have been run (with N = 1,000, C = 750 Mbit/s, and all of the r_i's less than 1 % of C), s* is in the range of 0.05 to 0.5, and $\mu_\Lambda''(s^*)$ is approximately 1,000. If ΔC is very small then: P(Λ > C+ΔC) = P(Λ>C) exp(-s*ΔC). As ΔC becomes larger, the term $\frac{-\Delta C^2}{2\mu_\Lambda''(s^*)}$ has a greater effect: if ΔC is positive, the decrease in P(Λ>C) is greater than exponential; if ΔC is negative the increase is less than exponential.

5.1.1 Effect of Changing Rate and Burstiness of One Source

The approximation of equation (18) can be used to analyze the effect on P(Λ>C) of changing the rate and burstiness of one source. Let the rate and on-mode probability of this one source be r and p, respectively; let the instantaneous aggregate arrival rate of all of the sources be represented by Λ, and let the instantaneous aggregate arrival rate of all sources excluding this one source be represented by Λ'. Consider keeping the average rate of this source the same but changing the relative values of p and r. Thus, $\bar{p} = p/\gamma$ and $\bar{r} = r \bullet \gamma$ for some positive constant γ. $\bar{\Lambda}$ is the new instantaneous aggregate arrival rate of all sources. Then:

$$P(\Lambda>C) = (1-p)\, P(\Lambda' > C) + p\, P(\Lambda' > C-r)$$

$$P(\bar{\Lambda}>C) = (1-p/\gamma)\, P(\Lambda' > C) + (p/\gamma)\, P(\Lambda' > C-r\gamma)$$

$$\begin{aligned}\Delta P &= P(\bar{\Lambda}>C) - P(\Lambda>C)\\ &= (p - p/\gamma)\, P(\Lambda' > C) + (p/\gamma)\, P(\Lambda' > C-r\gamma) - p\, P(\Lambda' > C-r) \qquad (19)\\ &\approx (p - p/\gamma)\, P(\Lambda' > C) + (p/\gamma)\, P(\Lambda' > C)\, Q(-r\gamma) - p\, P(\Lambda' > C)\, Q(-r)\\ &= [(1 - 1/\gamma) + (1/\gamma)\, Q(-r\gamma) - Q(-r)]\, p\, P(\Lambda' > C)\\ &= [(1/\gamma)\, (Q(-r\gamma) -1) - (Q(-r) -1)]\, p\, P(\Lambda' > C)\end{aligned}$$

(The Q(•) term was defined earlier to simplify Equation 18.) The sign of the term in brackets determines the sign of ΔP. When γ equals 1, the term is obviously 0. The term $(Q(-r\gamma) - 1)$ increases super-linearly in γ. Thus, when γ is greater than 1, ΔP is positive; when γ is less than 1, ΔP is negative.

Therefore, for a given average transmission rate, the burstier a source is (i.e., the smaller p is), the greater the probability that the intermediate node will be overrun. By forcing the source to be less bursty, the network can decrease $P(\Lambda>C)$ without having to decrease the average rate of the source. The change is linear in p, so for the same r and γ, the magnitude of ΔP is larger for a larger p. If r is very small, the term in brackets is close to 0. Thus, for very small r, the relative values of p and r are not important, just the average value (p)(r).

5.1.2 Effect of Adding A New Source

The approximation in equation (18) can also be used to examine the effect a new call has on $P(\Lambda>C)$. Assume there are initially *N* sources whose instantaneous aggregate arrival rate is Λ'. An additional call is set up such that the new instantaneous aggregate arrival rate is Λ. The new call has statistics p and r. Then:

$$P(\Lambda>C) = (1-p)\, P(\Lambda' > C) + (p)\, P(\Lambda' > C-r)$$

$$\Delta P = P(\Lambda>C) - P(\Lambda' > C) = (p)\, [P(\Lambda' > C-r) - P(\Lambda' > C)] \qquad (20)$$

Using equation (18) with $\Delta C = -r$ yields:

$$\Delta P \approx (p)\, [P(\Lambda' > C)] \left(\left(\frac{s^*}{s^* - \frac{r}{\mu_\Lambda''(s^*)}} \right) \left(\exp\left(s^* r - \frac{r^2}{2\mu_\Lambda''(s^*)} \right) \right) - 1 \right)$$

It can be seen that the increase is linear in p, and super-linear in r. Thus, as was shown in the previous section, for the same average rate, a more bursty source will cause a larger increase in $P(\Lambda > C)$. For a very small r, $\Delta P \approx (p)\, [P(\Lambda' > C)]\, (\exp(s^* r) - 1)$. Using the approximation $\exp(x) = 1 + x$ for small x, yields $\Delta P \approx (p)\, [P(\Lambda' > C)]\, (s^* r)$. Thus, for very small r, the increase in the overrun probability is linear in both p and r; only the average rate is important. It is also true that the larger $P(\Lambda' > C)$ is, the larger ΔP will be for given p and r values.

The effect of adding a second source with the same p and r values is not as straightforward to approximate since the values of s* and $\mu_\Lambda''(s^*)$ may change significantly after the first source is added. However, bounds can be obtained in terms of the effect of adding a single new source. Let:

$P_2 = P(\Lambda>C)$ after two new sources with statistics p and r have been added

$P_a = P(\Lambda>C)$ after one new source with statistics p^2 and 2r have been added

$P_b = P(\Lambda>C)$ after one new source with statistics p and 2r have been added

Λ' = instantaneous aggregate arrival rate before any sources have been added

Then:

$$P_2 = p^2\, P(\Lambda' > C-2r) + 2p(1-p)\, P(\Lambda'>C-r) + (1-p)^2\, P(\Lambda'>C)$$

$$P_a = p^2\, P(\Lambda' > C-2r) + (1-p^2)\, P(\Lambda'>C)$$

$$P_b = p\, P(\Lambda' > C-2r) + (1-p)\, P(\Lambda'>C)$$

First, express P_2 in terms of P_a:

$$\begin{aligned} P_2 &= P_a + 2p(1-p)\, P(\Lambda'>C-r) + (2p^2 - 2p)\, P(\Lambda'>C) \\ &= P_a + 2p(1-p)\, [P(\Lambda'>C-r) - P(\Lambda'>C)] \qquad (21) \end{aligned}$$

Since $P(\Lambda' > C-r)$ is greater than $P(\Lambda'>C)$, the term in brackets is greater than 0. Therefore, P_2 is greater than P_a. Now, express P_2 in terms of P_b:

$$\begin{aligned} P_2 &= P_b + (p^2-p)\, P(\Lambda'>C-2r) + 2p(1-p)\, P(\Lambda'>C-r) + (p^2-p)\, P(\Lambda'>C) \\ &= P_b + (p-p^2)\, [2\, P(\Lambda'>C-r) - P(\Lambda'>C-2r) - P(\Lambda'>C)] \\ &= P_b + (p-p^2)\, [\{P(\Lambda'>C-r) - P(\Lambda'>C)\} - \{P(\Lambda'>C-2r) - P(\Lambda'>C-r)\}] \qquad (22) \\ &= P_b + (p-p^2)\, P(\Lambda'>C)\, [\{Q(-r) - Q(0)\} - \{Q(-2r) - Q(-r)\}] \end{aligned}$$

Since $(Q(-x) - 1)$ increases super-linearly with x, the difference between $Q(-2r)$ and $Q(-r)$ is greater than the difference between $Q(-r)$ and $Q(0)$. Therefore, the term in brackets is less than 0, and thus, P_2 is less than P_b. Thus, P_2 can be bounded in terms of P_a and P_b: $P_a < P_2 < P_b$

$$P_b - P_a = (p - p^2)\, P(\Lambda>C-2r) + (p^2 - p)\, P(\Lambda>C) = (p - p^2)\, P(\Lambda>C)\, (Q(-2r) - 1)$$

Thus, the bound is very tight when p is close to 0 or 1. When p equals 1, the bounds coincide; this is expected because two sources that are always transmitting at rate r are equivalent to one source always transmitting at rate 2r. The bound is also tight when r is small. Whether P_a or P_b is the tighter bound depends on the value of r.

5.2 Gaussian Approximation

When $E[\Lambda]$ is close to C, $P(\Lambda>C)$ can be approximated by a Gaussian expression as in equation (12):

$$P(\Lambda>C) \approx \int_{C}^{\infty} \frac{1}{\sqrt{2\pi \, var[\Lambda]}} \exp\left(\frac{-(\lambda - E[\Lambda])^2}{2\, var[\Lambda]}\right) d\lambda$$

If C increases by ΔC, then:

$$P(\Lambda>C+\Delta_C) \approx \int_{C+\Delta_C}^{\infty} \frac{1}{\sqrt{2\pi \, var[\Lambda]}} \exp\left(\frac{-(\lambda - E[\Lambda])^2}{2\, var[\Lambda]}\right) d\lambda$$

Graph 11 provides an example of how P(Λ>C) increases when ΔC is negative. The increase in P(Λ>C) can be approximated by the area of a rectangle having a width of a C and a height of:

$$\frac{1}{\sqrt{2\pi \, var[\Lambda]}} \exp\left(\frac{-(C + \Delta_C/2 - E[\Lambda])^2}{2\, var[\Lambda]}\right)$$

$$\text{Thus, } P(\Lambda>C+\Delta_C) \approx P(\Lambda>C) - \frac{\Delta_C}{\sqrt{2\pi \, var[\Lambda]}} \exp\left(\frac{-(C + \Delta_C/2 - E[\Lambda])^2}{2\, var[\Lambda]}\right). \qquad (23)$$

For simplicity, this will be expressed as: P(Λ > C+ΔC) ≈ P(Λ>C) - ΔC Z(ΔC), where Z(ΔC) represents the term:

$$\frac{1}{\sqrt{2\pi \, var[\Lambda]}} \exp\left(\frac{-(C + \Delta_C/2 - E[\Lambda])^2}{2\, var[\Lambda]}\right)$$

for fixed values of C, E[Λ], and var[Λ]. Obviously, if ΔC is positive, P(Λ>C) decreases, and if ΔC is negative, P(Λ>C) increases. It is assumed that when ΔC is negative, we are only interested in the case where |ΔC| < (C – E[Λ]) . With this restriction, the more negative ΔC is, the larger Z(ΔC) is.

Graph 12 plots P(Λ>C) for various values of C, as calculated using the Gaussian expression. Graph 13 provides an example of how closely the approximation of equation (23) follows the value of P(Λ>C+ΔC) calculated using the Gaussian expression. (In this example, for positive ΔC, the approximation switches from an underestimate to an overestimate when the Gaussian curve switches from concave to convex.)

If C-E[Λ] is small and ΔC < 0, then the exponential term in equation (23) is approximately equal to 1, and:

$$P(\Lambda>C+\Delta_C) \approx P(\Lambda>C) - \frac{\Delta_C}{\sqrt{2\pi \, var[\Lambda]}}$$

The increase of P(Λ>C) is linear in ΔC, and the increase is smaller as var[Λ] becomes larger.

Equation (23) can also be written as:

$$P(\Lambda>C+\Delta C) \approx P(\Lambda>C) - \frac{\Delta C}{\sqrt{2\pi\, var[\Lambda]}} \exp\frac{-(C-E[\Lambda])^2}{2var[\Lambda]} \exp\left[\frac{-(C-E[\Lambda])}{2var[\Lambda]}\Delta C - \frac{\Delta C^2}{8\, var[\Lambda]}\right]$$

so that the exponential term involving ΔC is in a form similar to the exponential term in the asymptotic expression of equation (18). The exponent in equation (18) is about four times greater in magnitude than the exponent in brackets in the above equation for the same ΔC and with reasonable values for the other parameters. Thus, the exponential effect is smaller in the Gaussian approximation, and the changes in P(Λ>C) are closer to linear. This can also be seen from comparing Graph 12 (Gaussian approximation) to Graph 9 (asymptotic expression).

5.2.1 Effect of Changing Rate and Burstiness of One Source

The effect of changing p and r of one source, while keeping (p)(r) fixed can be analyzed as was done in Section 5.1.1. With $\overline{p}$ = p/γ and r = r•γ, and using equation (19):

$$\begin{aligned}\Delta P &= (p - p/\gamma)\, P(\Lambda' > C) + (p/\gamma)\, P(\Lambda' > C-r\gamma) - p\, P(\Lambda' > C-r)\\ &\approx (p - p/\gamma)\, P(\Lambda' > C) + (p/\gamma)\, [P(\Lambda' > C) + r\gamma\, Z(-r\gamma)]\\ &\qquad - p[P(\Lambda' > C) + r\, Z(-r)]\\ &= p\, r\, [Z(-r\gamma) - Z(-r)]\end{aligned}$$

If γ > 1, then Z(-r γ) > Z(-r), and the increase in P(Λ>C) is positive. As was shown with the asymptotic expression, the more bursty a source is, the larger P(Λ>C) is. The change in P(Λ>C) is again linear in p, and close to 0 for very small r and γ.

5.2.2 Effect of Adding A New Source

Equation (23) can also be used to analyze the impact on P(Λ>C) of adding a new source to the system, as was done with the asymptotic expression in section 5.1.2. From equation (20):

$$\begin{aligned}\Delta P &= P(\Lambda>C) - P(\Lambda' >C) = (p)\, [P(\Lambda' > C-r) - P(\Lambda' > C)]\\ &= \frac{p\, r}{\sqrt{2\pi\, var[\Lambda]}} \exp\left(\frac{-(C-r/2-E[\Lambda])^2}{2\, var[\Lambda]}\right)\end{aligned}$$

If C-E[Λ] is small, then:

$$\Delta P \approx \frac{p\, r}{\sqrt{2\pi\, var[\Lambda]}},$$

which is linear in both p and r, and is smaller as var[Λ] becomes larger.

When adding two new sources with statistics p and r, the same bounds hold as in section 5.1.2. From equation (21):

$$P_2 = P_a + 2p(1-p)\ [P(\Lambda'>C-r) - P(\Lambda'>C)]$$

$$= P_a + 2p(1-p)\ [r\ Z(-r)]$$

The term in brackets is always positive, thus: $P_2 > P_a$.

From equation (22):

$$P_2 = P_b + (p-p^2)\ [\{P(\Lambda'>C-r)- P(\Lambda'>C)\} - \{P(\Lambda'>C-2r) - P(\Lambda'>C-r)\}]$$

$$= P_b + (p-p^2)\ [\{P(\Lambda'>C) + r\ Z(-r) - P(\Lambda'>C)\} - \{P(\Lambda'>C) + 2r\ Z(-2r) - P(\Lambda'>C) - r\ Z(-r)\}]$$

$$= P_b + (p-p^2)\ [2r\ Z(-r) - 2r\ Z(-2r)]$$

$$= P_b + 2r\,(p-p^2)\ [Z(-r) - Z(-2r)]$$

The term in brackets is always negative; thus, $P_2 < P_b$, which leads to: $P_a < P_2 < P_b$.

6. CONCLUSION

A criterion for evaluating how well a given set of source rates will perform in preventing congestion has been developed. It only evaluates the data flow at one intermediate node, rather than the network as a whole. Furthermore, it only considers the instantaneous data flow, not the traffic over a period of time. Nevertheless, it provides insight into the possible queuing delay which might develop and the probability of a buffer overrun.

The equations as presented do not necessarily suggest the initial rates that should be assigned to the sources. They are more suited to analyzing a set of rates that has already been determined, or examining the effect of changing one of the rates. In general, it was shown that for the same average transmission rate, the more bursty a source is, the more likely it is to cause congestion.

REFERENCES

[1] D. Clark, M. Lambert, and L. Zhang, "NETBLT: A High Throughput Transport Protocol," *ACM*, 1988.

[2] D. Cheriton and C. Williamson, "VMTP as the Transport Layer for High-Performance Distributed Systems," *IEEE Communications Magazine*, June, 1989.

[3] R. Gallager, *Information Theory and Reliable Communication*, J. Wiley, 1968.

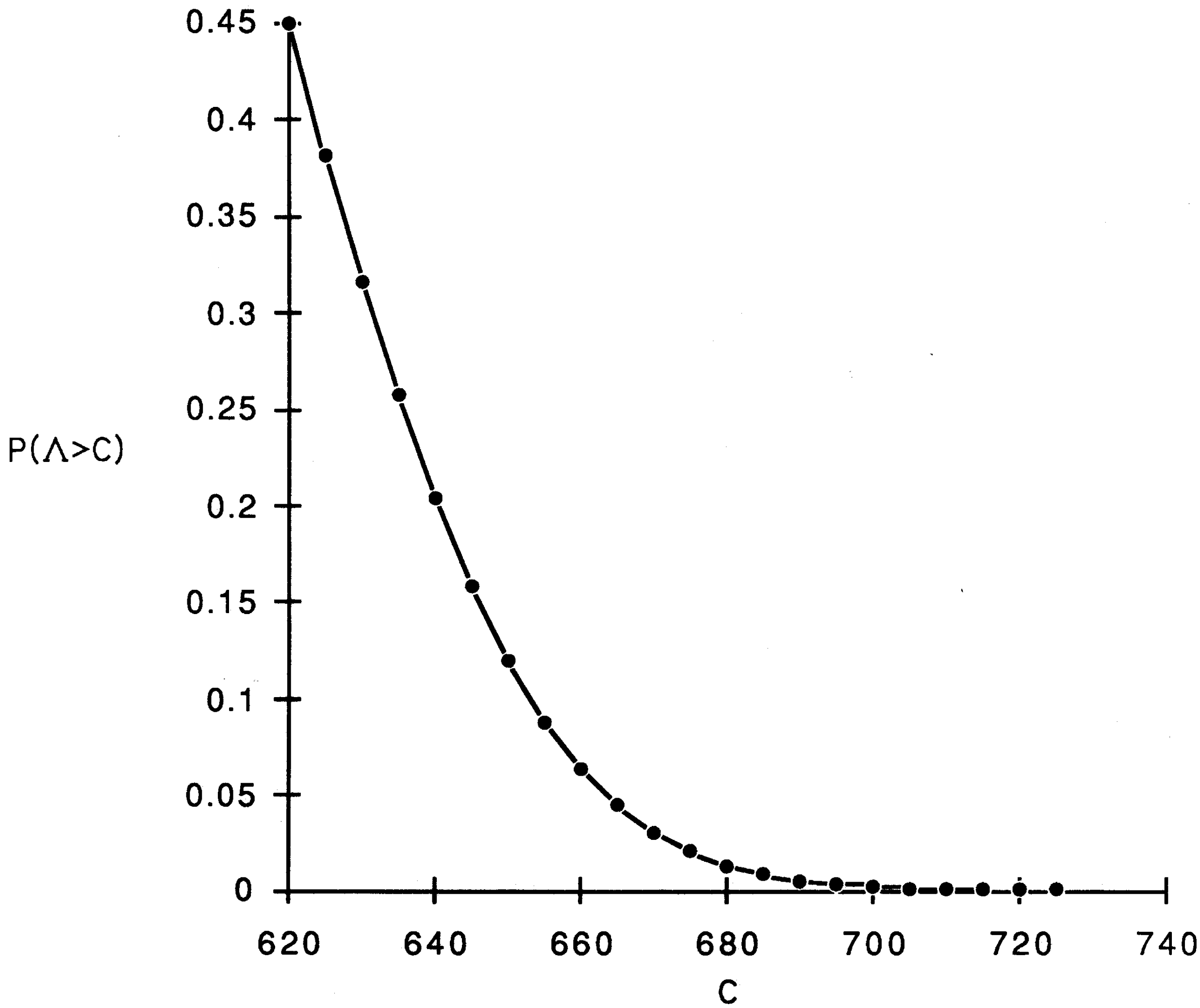
[4] D. Bertsekas and R. Gallager, *Data Networks*, Prentice Hall, 1987.

[5] I. Cidon and I. Gopal, "PARIS: An Approach to Integrated High-Speed Private Networks," *International Journal of Digital and Analog Cabled Sustems*, 1988.

[6] Van Jacobson, "Congestion Avoidance and Control," *ACM*, 1988.

[7] W. Feller, *An Introduction to Probability Theory and Its Applications*, Volume II, 1966.

GRAPH 1



Approximation to P(Λ>C) using:

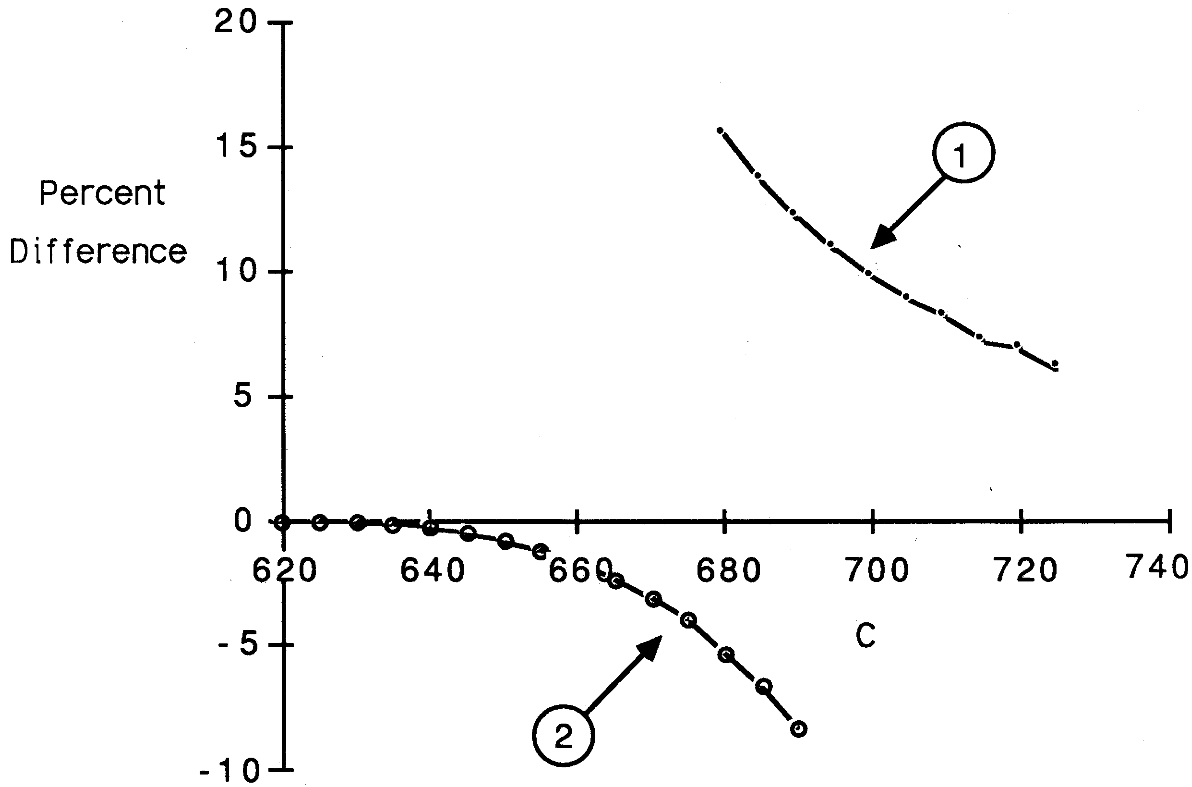
$$\exp(\mu_{\Lambda}(s^*) - s^*C)\,\exp\left(\frac{s^{*2}\,\mathrm{var}[\tilde{\Lambda}]}{2}\right)\Phi(-s^*\sqrt{\mathrm{var}[\tilde{\Lambda}]})$$

In this example, $E[\Lambda] \approx 616$

$\mathrm{var}[\Lambda] \approx 805$

$\mathrm{std\ dev}\,[\Lambda] \approx 28$

GRAPH 2



Comparisons of Different Approximations to $P(\Lambda > C)$

Series 1: $\dfrac{(Y-X)}{X} * 100$ Series 2: $\dfrac{(Z-X)}{X} * 100$

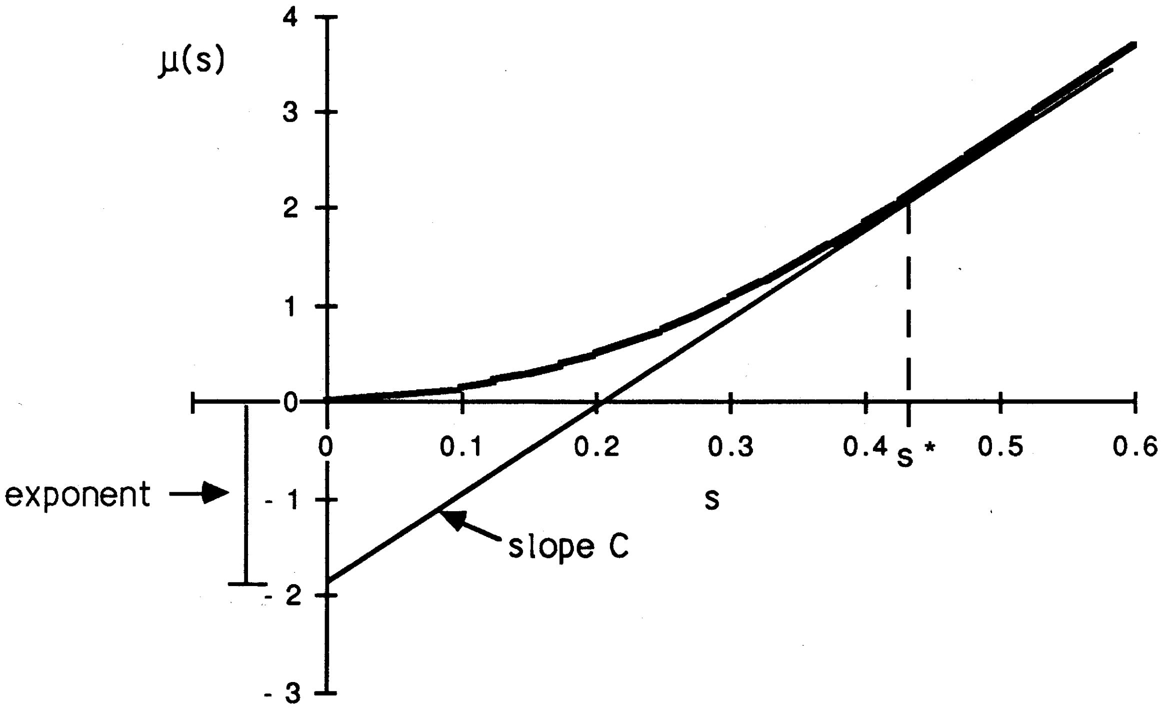
X is approximation using: $\exp(\mu_\Lambda(s^*) - s^*C)\exp\left(\dfrac{s^{*2}\,\mathrm{var}[\tilde{\Lambda}]}{2}\right)\Phi(-s^*\sqrt{\mathrm{var}[\tilde{\Lambda}]})$

Y is approximation using: $\exp(\mu_\Lambda(s^*) - s^*C)\,\dfrac{1}{s^*\sqrt{2\pi\mathrm{var}[\tilde{\Lambda}]}}$ (Asymptotic Expression)

Z is approximation using: $\Phi\left(\dfrac{-C+E[\Lambda]}{\sqrt{\mathrm{var}[\Lambda]}}\right)$ (Gaussian Approximation)

In this example, $E[\Lambda] \approx 616$
$\mathrm{var}[\Lambda] \approx 805$
$\text{std dev }[\Lambda] \approx 28$

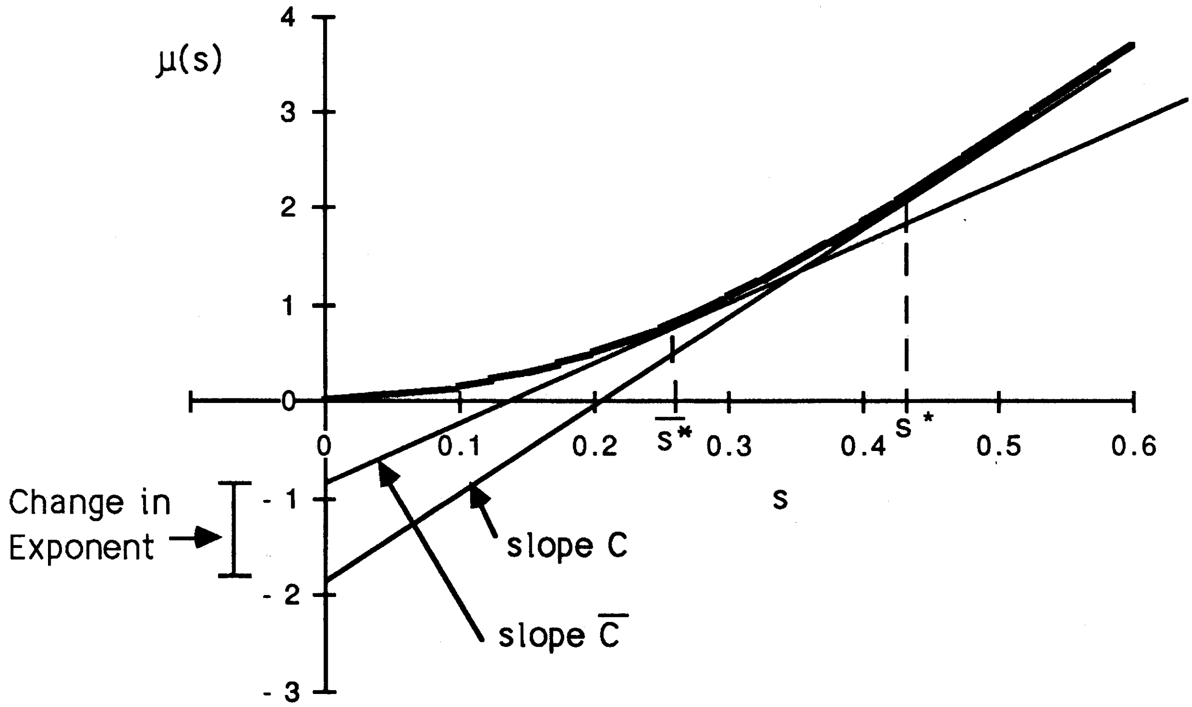
GRAPH 3



Graphical Representation of the Chernoff Bound Exponent

Exponent = $\mu(s^*) - s^*C$

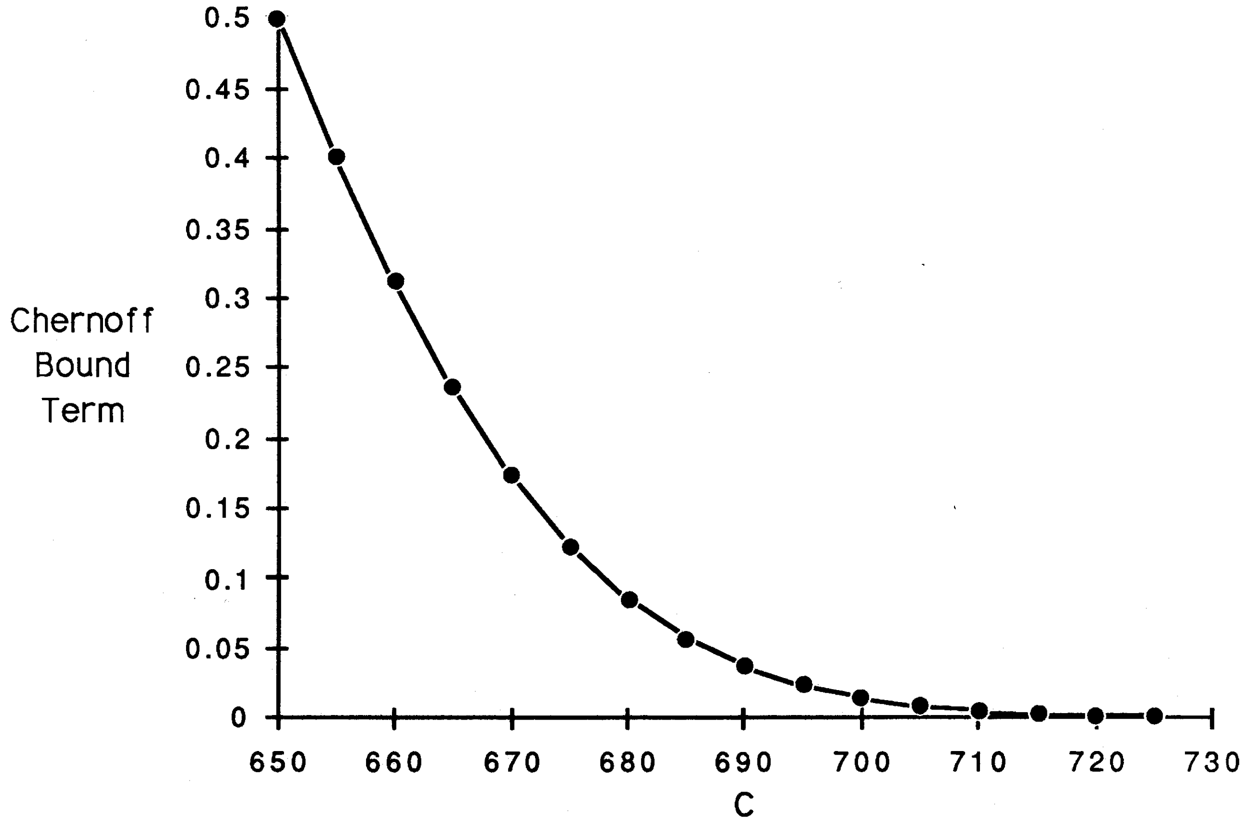
GRAPH 4



Effect of Changing C on the Chernoff Bound Exponent

If $\overline{C} < C$, then $\overline{s}^* < s$ and the magnitude of the exponent is smaller

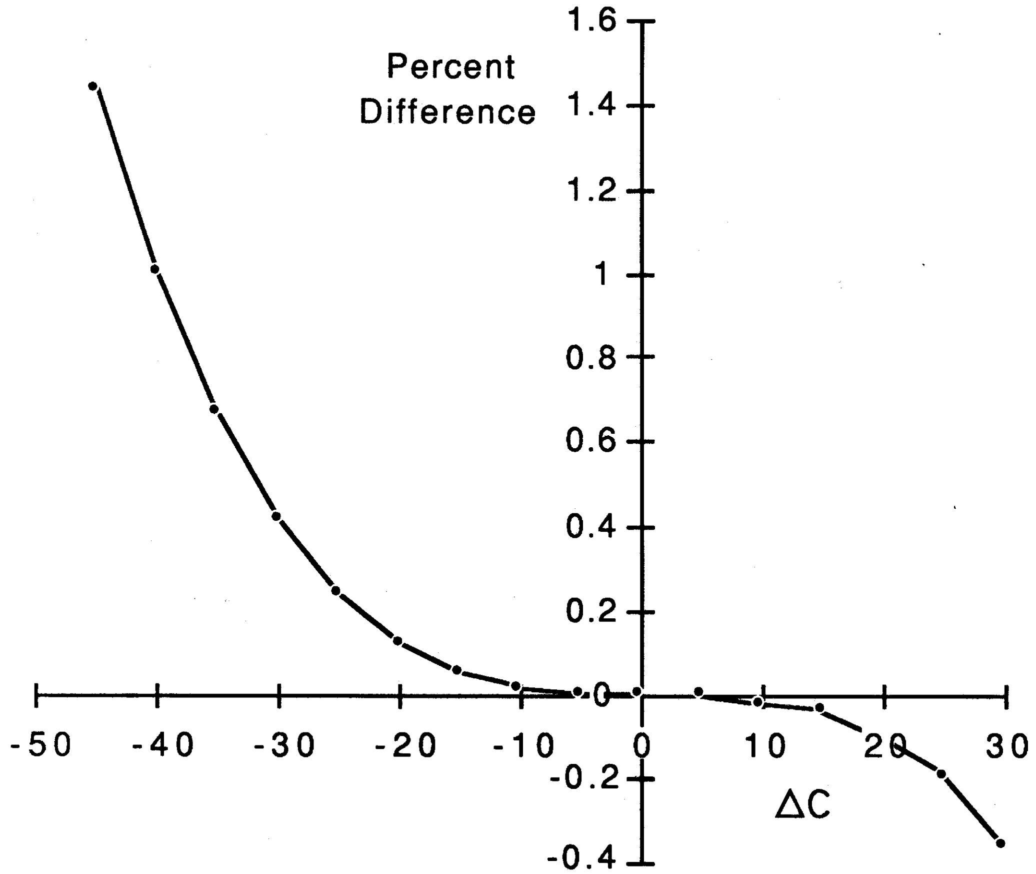
GRAPH 5



Chernoff Bound Term as C Changes

Calculated using: $\mu_{\Lambda}(s^*) - s^* C$; s^* such that $\mu_{\Lambda}'(s^*) = C$

GRAPH 6



Approximation to Chernoff Bound Term as C Changes

Plot of $\frac{(Y-X)}{X} * 100$ where:

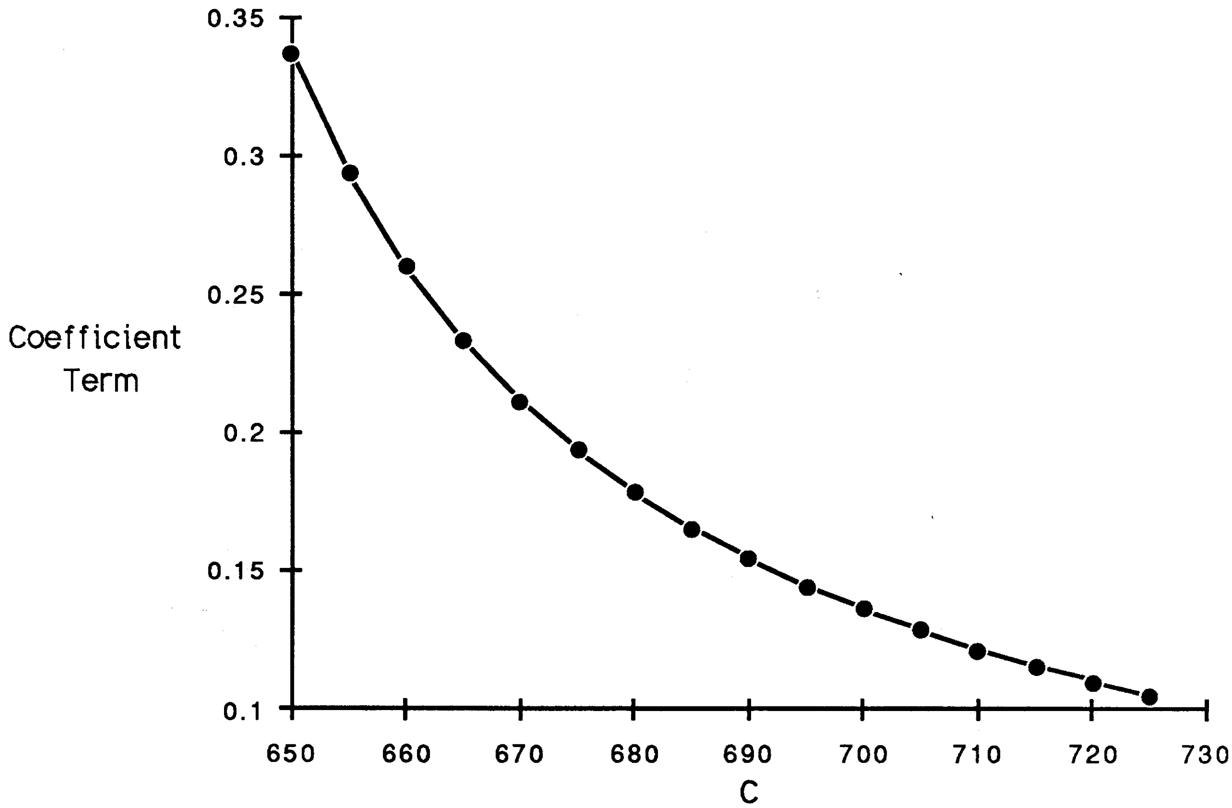
X is Chernoff bound term: $\mu_\Lambda(s^*) - s^*(C + \Delta C)$; s^* such that $\mu_\Lambda'(s^*) = C + \Delta C$

Y is Approximation to Chernoff bound term:

$$[\mu_\Lambda(s) - sC] \exp\left(-s\Delta_C - \frac{\Delta_C^2}{2\mu_\Lambda''(s)}\right) ; \text{ s fixed such that } \mu_\Lambda'(s) = C$$

In this example C = 695

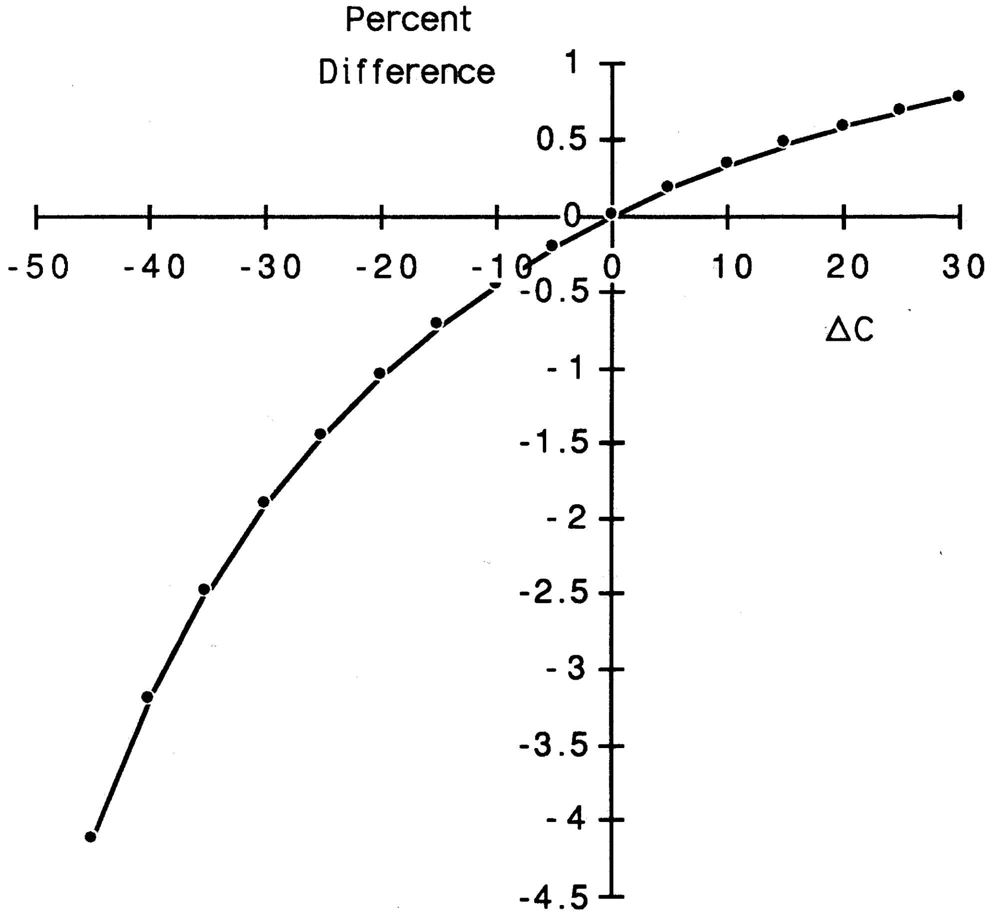
GRAPH 7



Coefficient Term as C Changes

Calculated using: $\dfrac{1}{s^*\sqrt{2\pi\mu_\Lambda''(s^*)}}$; s^* such that $\mu_\Lambda'(s^*) = C$

GRAPH 8



Approximation to Coefficient Term as C Changes

Plot of $\frac{(Y-X)}{X} * 100$ where:

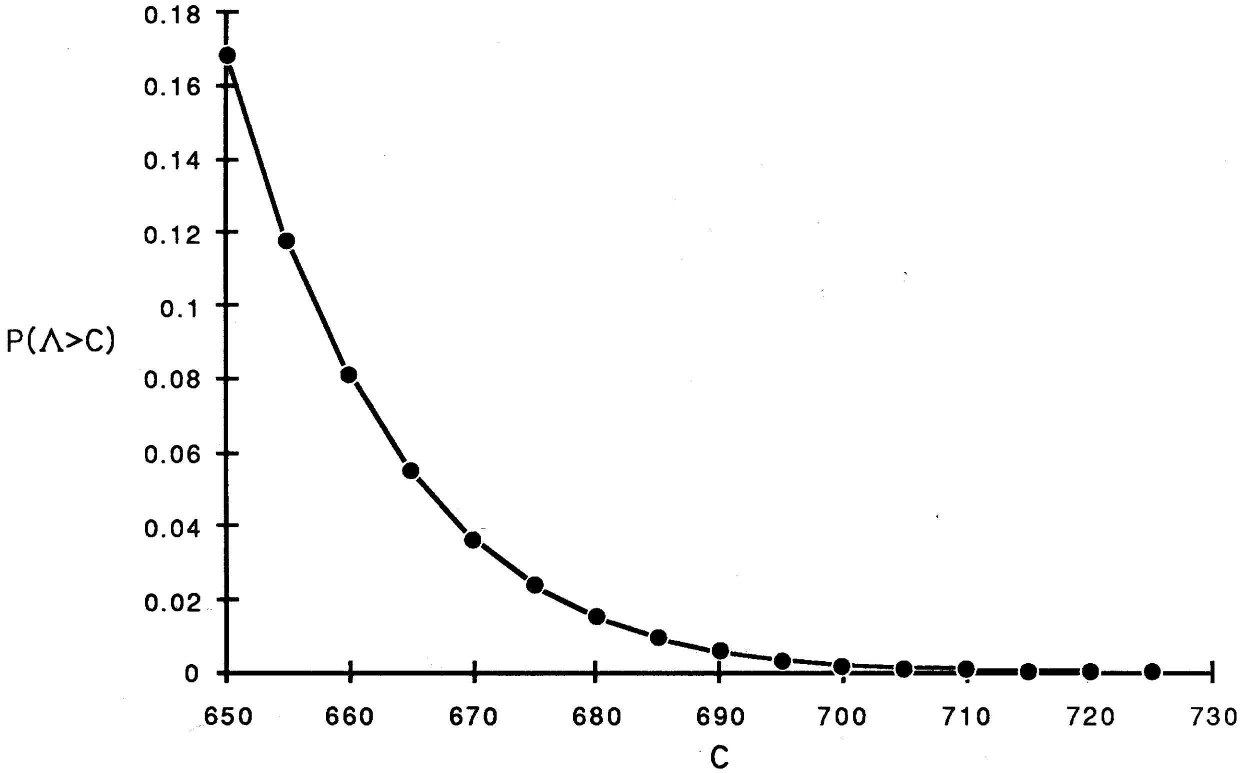
X is coefficient term: $\frac{1}{s^* \sqrt{2\pi \mu_\Lambda''(s^*)}}$; s^* such that $\mu_\Lambda'(s^*) = C + \Delta C$

Y is approximation to coefficient term:

$$\left(\frac{1}{s\sqrt{2\pi\mu_\Lambda''(s)}}\right)\left(\frac{s}{s+\frac{\Delta c}{\mu_\Lambda''(s)}}\right);\quad s \text{ fixed such that } \mu_\Lambda'(s) = C$$

In this example C = 695

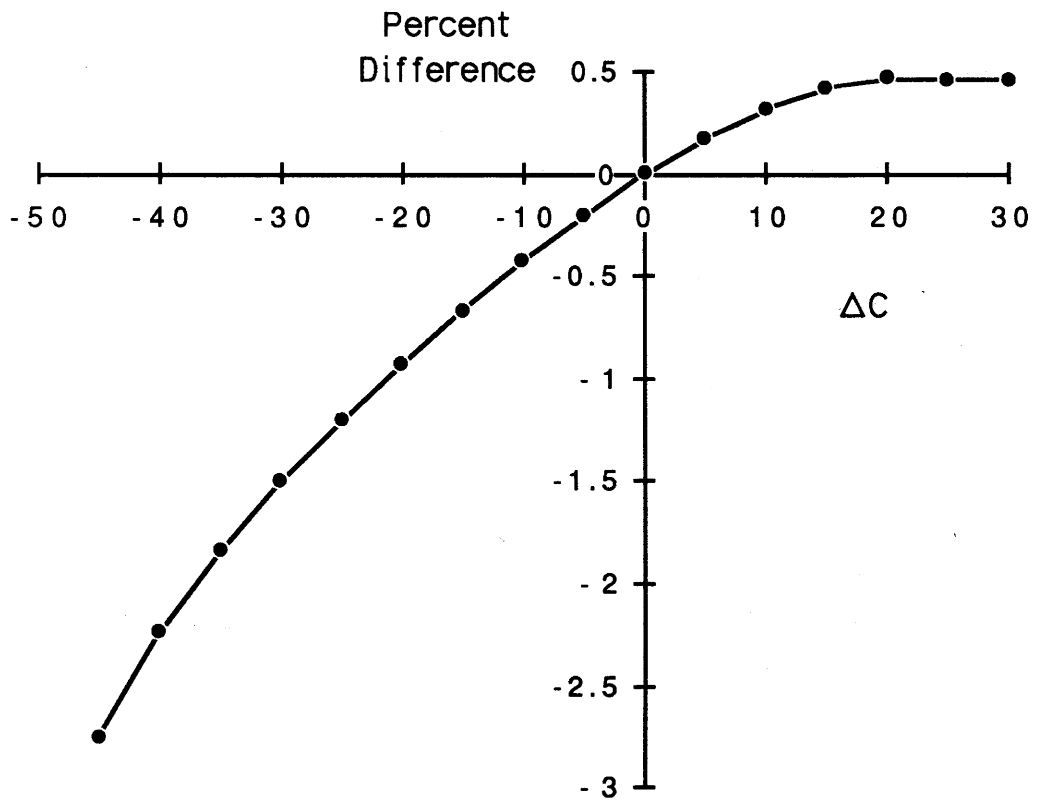
GRAPH 9



Asymptotic Expression for $P(\Lambda > C)$ as C Changes

Calculated using: $\exp(\mu_\Lambda(s^*) - s^* C) \dfrac{1}{s^* \sqrt{2\pi \mu_\Lambda''(s^*)}}$; s^* such that $\mu_\Lambda'(s^*) = C$

GRAPH 10



Approximation to Asymptotic Expression for $P(\Lambda>C)$ as C Changes

Plot of $\frac{(Y-X)}{X} * 100$ where:

X is asymptotic expression for $P(\Lambda>C)$:

$$\exp(\mu_\Lambda(s^*) - s^*(C+\Delta C)) \frac{1}{s^*\sqrt{2\pi\mu_\Lambda''(s^*)}} \quad ; \quad s^* \text{ such that } \mu_\Lambda'(s^*) = C + \Delta C$$

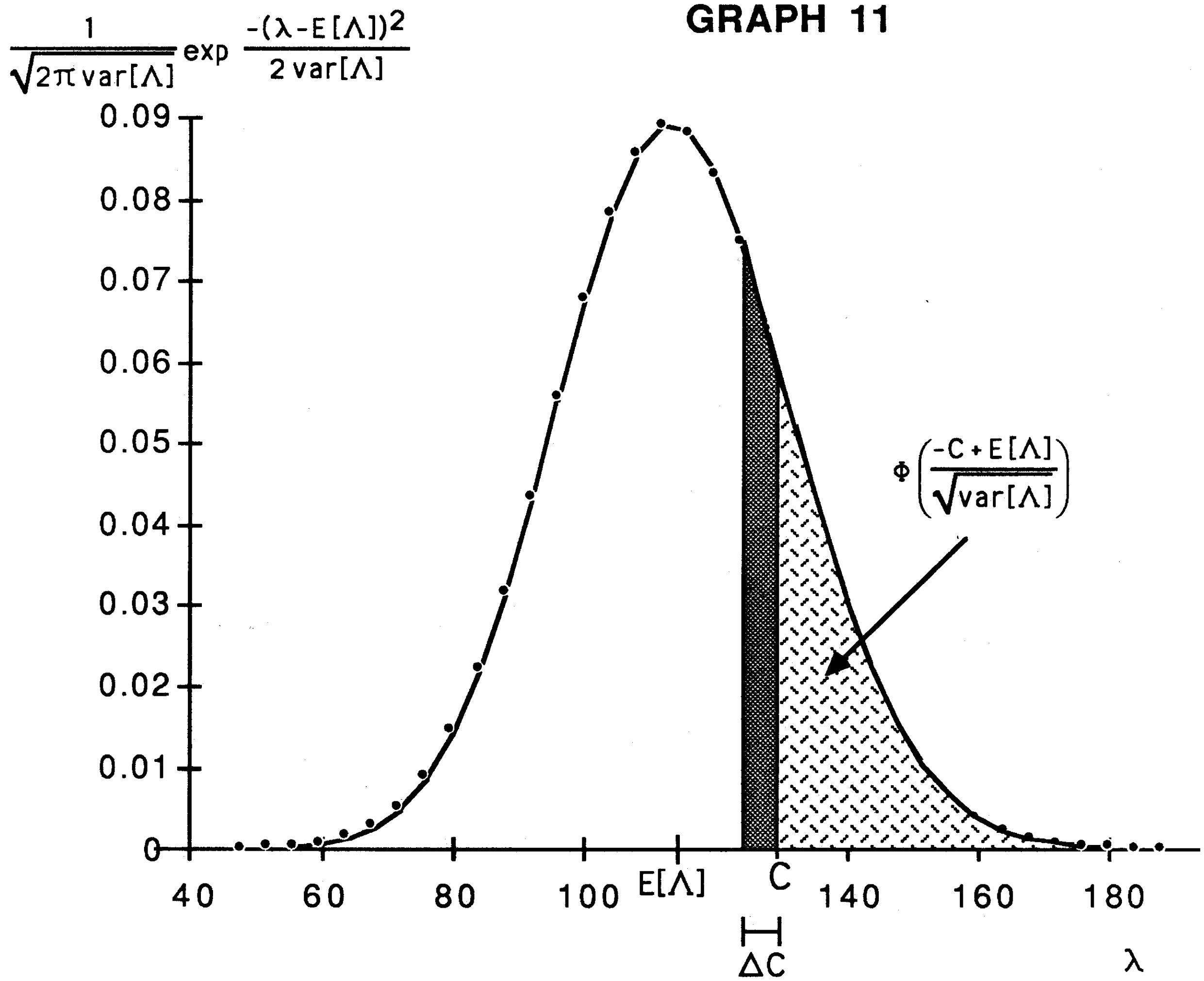
Y is approximation to asymptotic expression:

$$\exp(\mu_\Lambda(s) - sC)\frac{1}{s\sqrt{2\pi\mu_\Lambda''(s)}}\exp\left(-s\Delta_C - \frac{\Delta_C^2}{2\mu_\Lambda''(s)}\right)\left(\frac{s}{s+\frac{\Delta_C}{\mu_\Lambda''(s)}}\right)$$

s fixed such that $\mu_\Lambda'(s) = C$

In this example C = 695

GRAPH 11

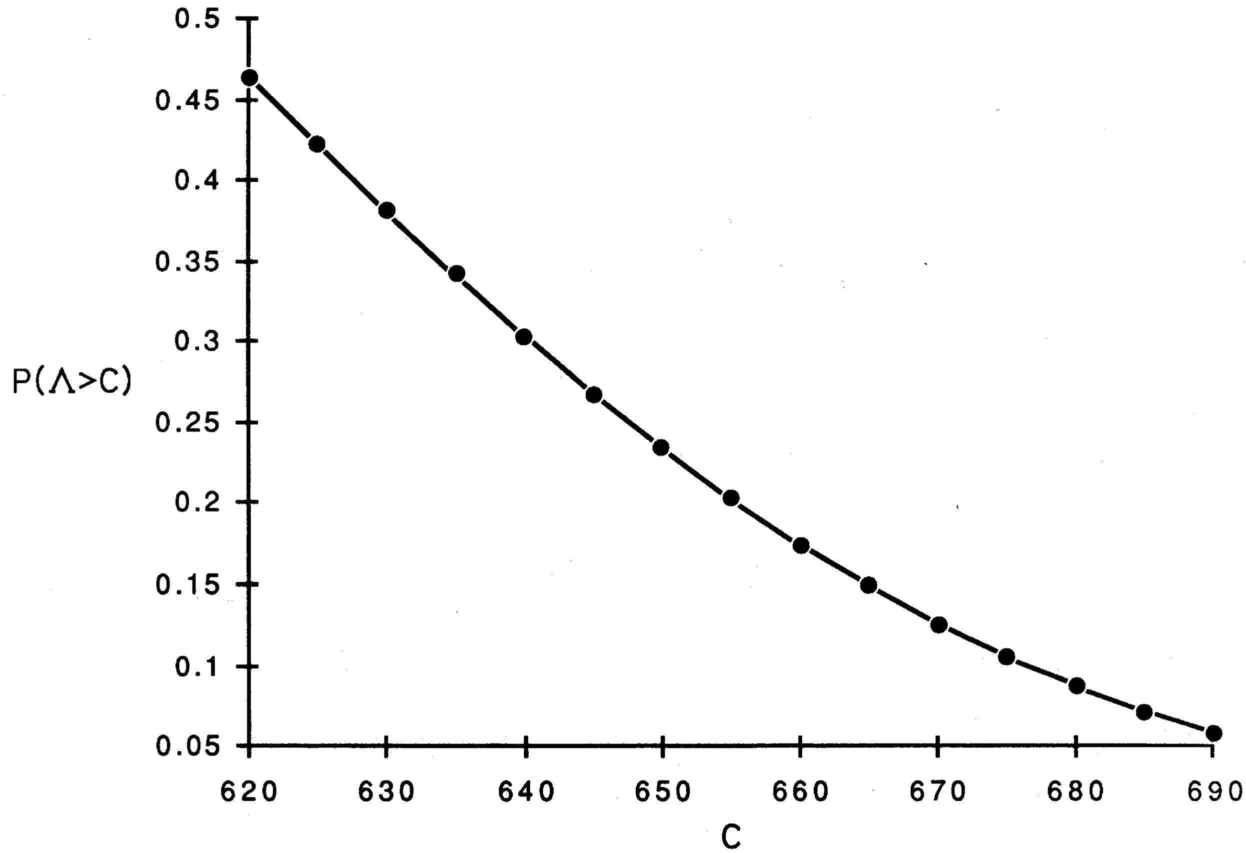


The shaded region represents the increase in $\Phi\left(\frac{-C+E[\Lambda]}{\sqrt{\mathrm{var}[\Lambda]}}\right)$ for $\Delta C < 0$

The area of this region can be approximated by the area of a rectangle:

$$\frac{\Delta C}{\sqrt{2\pi\,\mathrm{var}[\Lambda]}}\exp\left(\frac{-(C+\Delta C/2-E[\Lambda])^2}{2\,\mathrm{var}[\Lambda]}\right)$$

GRAPH 12



Gaussian Expression for P(Λ>C) as C Changes

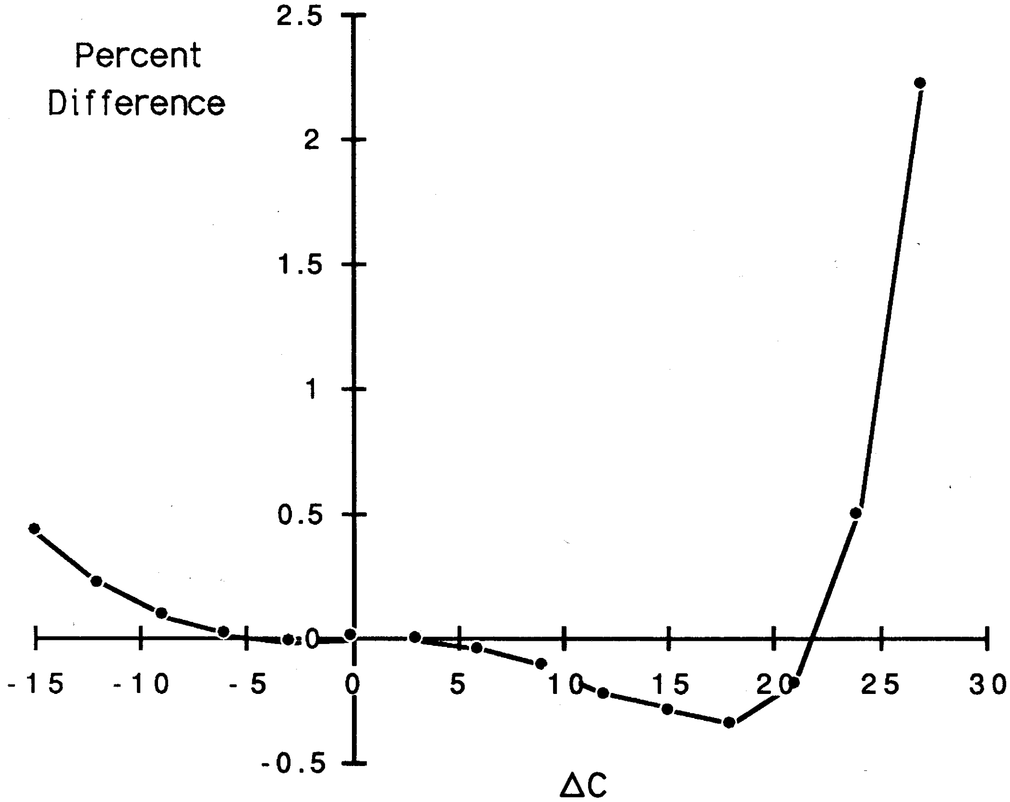
Calculated using: $\Phi\left(\frac{-C + E[\Lambda]}{\sqrt{var[\Lambda]}}\right)$

In this example, $E[\Lambda] \approx 616$

$var[\Lambda] \approx 805$

$\text{std dev } [\Lambda] \approx 28$

GRAPH 13



Approximation to Gaussian Expression for P(Λ>C) as C Changes

Plot of $\frac{(Y-X)}{X} * 100$ where:

X is Gaussian expression for P(Λ>C): $\Phi\left(\frac{-(C+\Delta C)+E[\Lambda]}{\sqrt{\mathrm{var}[\Lambda]}}\right)$

Y is approximation to Gaussian expression:

$$\Phi\left(\frac{-C+E[\Lambda]}{\sqrt{\mathrm{var}[\Lambda]}}\right) - \frac{\Delta C}{\sqrt{2\pi\,\mathrm{var}[\Lambda]}}\exp\left(\frac{-(C+\Delta C/2-E[\Lambda])^2}{2\,\mathrm{var}[\Lambda]}\right).$$

In this example, $E[\Lambda] \approx 616$
$\mathrm{var}[\Lambda] \approx 805$
std dev $[\Lambda] \approx 28$
$C = 634$