STUDY OF CLUSTERED ARRIVAL PROCESSES AND SIGNALING LINK DELAYS

Ronald A. SKOOG

AT&T Bell Laboratories
Holmdel, New Jersey 07733 USA

This paper models the message arrival process for signaling links in common channel signaling networks, approximates the message arrival process using two simple clustered arrival processes, shows when an M/G/1 approximation can be used to characterize signaling link queueing delays, and examines other approximation methods for heavy traffic conditions.

1. INTRODUCTION

The motivation for this work comes from investigating queueing delays on signaling links in Common Channel Signaling (CCS) networks. Signaling traffic consists of messages sent between switching systems to set up and take down connections and messages sent between switching systems and network databases that are used to provide intelligent network services. The aspect of signaling traffic we are concerned with is that even if the call arrival process is Poisson the message arrival process to signaling links will not be Poisson! This is because the message arrival times for a particular call are correlated. For example, looking at the signaling messages leaving a switch that is originating a call, the Initial Address Message (IAM) is sent out first, and then there are subsequent messages that are sent at times that are dependent on when the IAM was sent. For services requiring database transactions (e.g., 800 or Freephone service) there will be a database query followed by trunk set-up and take down message sequences.

In examining the nature of signaling traffic, it is seen that the message arrival process is what has been called a clustered\(^1\) or branching\(^2\) process. In these processes there are primary events separated by random interarrival times, and associated with each primary event is a cluster of secondary events. There are a random number of secondary events in a cluster, and they are separated by random interarrival times. If the arrival process for the primary events is Poisson, the process is called a clustered Poisson process.

It has been found in other studies of packet networks (e.g., see References\(^3\) \(^4\) \(^5\) ) that interarrival time dependencies can have a significant effect on queueing delays. We are interested in determining when the interarrival time dependencies in signaling traffic can be ignored, and thus determine when the signaling link queueing delays can be approximated by an M/G/1 model when the call arrival process is Poisson. One would intuitively expect that if the time separation between dependent message arrivals is large enough, then the correlations between interarrival times would not have an appreciable effect, and the M/G/1 queue would be an accurate approximation. The notion of a large time separation has meaning in reference to the length of the busy periods of the queueing system; when the time separation is large compared to the length of most busy periods, a correlated arrival will come, with high probability, after a busy period and the system has lost all memory of the previous correlated event. Therefore, the system should behave as though the arrivals were not correlated.

To gain some understanding of what time separation is large enough between correlated arrivals that their dependency can be ignored, we will examine two simple types of clustered Poisson process. We then study a model of the actual signaling traffic and show that the M/G/1 model is accurate for a wide range of parameter values. The results from the analysis of the simple clustered Poisson processes provide an understanding of why this is the case. Finally, it is observed that the M/G/1 model becomes inaccurate under extremely heavy load (i.e., link utilization near 1), and we discuss some approximation methods to use under these heavy traffic conditions.

The signaling link queueing delays that are considered here are for the signaling link transmit buffer at the end of a link connected to a switching office. Thus the signaling message stream consists of the messages the switch is sending out to the network, and the arrival time of a message is when the switch sends the message. Other links in a signaling network (e.g., those connected to a Signal Transfer Point) will have a more complicated message arrival process due to messages passing through other links before arriving at the link being considered. Results for these more complicated situations will be reported in a subsequent paper.
2. ANALYSIS OF SOME SIMPLE CLUSTERED POISSON PROCESSES

We examine two simple forms of a clustered Poisson process, called Type 1 and Type 2. In both Type 1 and Type 2 arrival processes there is a single secondary event after the primary event. In a Type 1 process the time separation between the primary and secondary event is a random time, \( \Delta \). In a Type 2 process, the secondary event occurs at a random time, \( \Delta \), after the service completion of the primary event. For both of these processes we will consider the case when \( \Delta \) is deterministic, and we examine how the mean queueing delay performance changes for the primary and secondary messages as a function of \( \Delta \).

2.1 Results for Type 1 Clustered Poisson Processes

For the two extreme cases of \( \Delta = 0 \) and \( \Delta \to \infty \), expressions for mean delays are easily obtained. For \( \Delta = 0 \) the batch arrival M/G/1 model[6] can be used. Let \( Q_p(\Delta) \) and \( Q_s(\Delta) \) denote the mean queueing delay for the primary and secondary messages, respectively, and assume the primary and secondary messages have the same service-time distribution with first and second moments \( \bar{s} \) and \( s^2 \), respectively. The service-time coefficient of variation (the variance divided by the square of the mean) is denoted by \( c_2^2 \). Let \( \rho \) denote the link utilization. Then \( Q_p(0) \) and \( Q_s(0) \) are given by:

\[
Q_p(0) = \frac{\rho \bar{s}}{1 - \rho} \left( 1 + \frac{c_2^2}{2} \right)
\]

\[
Q_s(0) = Q_p(0) + \bar{s} = \frac{s}{1 - \rho} \left( 1 + \frac{c_2^2}{2} \right)
\]

For \( \Delta \to \infty \) it is clear that the M/G/1 model describes the queueing delay for both primary and secondary messages, and we denote this delay by \( Q_{\infty} \). \( Q_{\infty} \) is given by

\[
Q_{\infty} = \frac{\rho \bar{s}}{2(1 - \rho)} \left( 1 + c_2^2 \right)
\]

The ratios \( Q_p(0)/Q_{\infty} \) and \( Q_s(0)/Q_{\infty} \) give the maximum error that can result from using an M/G/1 approximation to determine mean queueing delays. These ratios are equal to \( (2 + c_2^2)/(1 + c_2^2) \) and \( (2/\rho + c_2^2)/(1 + c_2^2) \), respectively, and they are greatest when the service-time is deterministic \( (c_2^2 = 0) \). We consider deterministic service-times and examine how large \( \Delta \) must be for the mean queueing delays \( Q_p(\Delta) \) and \( Q_s(\Delta) \) to be closely approximated (say within 10\%) by the M/G/1 model.

Let \( \Delta_{\text{conv}} \) denote the value of \( \Delta \) at which the mean queueing delay converges to 10\% of the asymptotic value, \( Q_{\infty} \). As discussed above, it is meaningful to relate \( \Delta \) to the length of busy periods. For this purpose, we will use the mean busy period for the M/G/1 queue, which will be denoted by \( \tau_{bp} \), and we determine \( \Delta_{\text{conv}}/\tau_{bp} \). For the M/G/1 queue, \( \tau_{bp} = \bar{s}/(1 - \rho) \).

Simulations were run for signaling links having a transmission rate of 56 Kbps, message arrival times given by a Type 1 clustered Poisson process with deterministic \( \Delta \), and a message length of 30 octets for both primary and secondary messages. The link speed and message length were chosen to be representative of signaling systems. The results are actually independent of the message service-time in the sense that changing \( \bar{s} \) simply changes the time scale. All results were obtained by simulating 8 minutes of signaling link operation, starting from an idle system, and computing means after a 1 minute "warm-up" period to allow the system to reach steady state.

The simulation results are shown in Table 1. As might be expected, the convergence time for primary messages is shorter than for secondary messages. More significant, however, is that the convergence time \( \Delta_{\text{conv}}/\tau_{bp} \) increases with increasing utilization. Since the mean busy period \( \tau_{bp} \) increases with increasing utilization, this shows that the convergence time increases dramatically with increasing utilization.

2.2 Results for Type 2 Clustered Poisson Processes

A signaling link with a Type 2 message arrival process can be modeled as a single server queue with feedback through a delay \( \Delta \) (this is an infinite server queue with service time \( \Delta \)). Primary messages arrive to the system

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation Results for the Type 1 Clustered Poisson Process</td>
</tr>
<tr>
<td>(4.3 ms Primary and Secondary Message Service-Times)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>( \rho )</td>
</tr>
<tr>
<td>0.25</td>
</tr>
<tr>
<td>0.5</td>
</tr>
<tr>
<td>0.8</td>
</tr>
</tbody>
</table>
according to a Poisson arrival process. Primary messages leaving the single server queue are sent with probability one to the delay and change class from primary to secondary. After going through the delay Δ, the secondary messages join the single server queue, and when they complete service they leave the system.

As for Type 1 arrival processes, we consider a deterministic delay Δ and examine how the mean queueing delays change as a function of Δ. For this model the service-time distribution can have a profound effect on the system performance. In particular, if the service-times are exponentially distributed, the system has a product form solution[7] and therefore, for any value of Δ, the mean queueing delay is the same as in an M/M/1 queue!

Simulation results were obtained for the case of a deterministic service-time, and the results are shown in Table 2. The simulation time and warm-up period were 8 minutes and 1 minute, respectively; the same as those used for the Type 1 simulations. If there were no dependencies in arrival times, the system would behave as a M/D/1 queue, and therefore we normalize queueing delays to the M/D/1 mean queueing delay $Q_{M/D/1} = S/2(1 - \rho)$. When Δ = 0 the primary and secondary normalized delays are greater than one, but not as high as the normalized delays with Type 1 message arrivals. More interesting is the behavior as Δ → ∞. For Type 2 arrivals the mean delays become less than $Q_{M/D/1}$. This implies that having secondary messages delayed after the service of the primary message smooths the arrival process of secondary messages. This smoothing reduces the effect of dependencies when Δ is small, and causes mean delays to be reduced below $Q_{M/D/1}$ when Δ gets large. As Δ → ∞ the normalized delay for secondary messages are considerably less than for primary messages. However, as the utilization increases the difference between the two asymptotic normalized delays decreases.

For the Type 2 arrivals, $\Delta_{conv}/\tau_{bp}$ is less than what was found for Type 1 arrivals. For the primary messages it is considerably less, and for secondary messages it is slightly less.

### 3. ANALYSIS OF SIGNALING TRAFFIC MESSAGE STREAMS

The signaling message stream model considered is for call set-up using CCITT Signaling System No. 7 ISDN User Part (ISUP).[8] We do not consider call irregularities, and so the calls considered complete to ringing or busy. It is assumed that 75% of all calls complete to ringing and are answered, and for half of those calls the caller hangs up first. It is assumed that 13% of all calls are placed to busy lines, and 12% of all calls complete to ringing with no answer.

The ISUP message stream coming out of an office consists of the superposition of two independent streams, one for messages traveling from the calling end to the called end and one for messages traveling from the called end to the calling end. We will assume call arrivals to the network are uniformly distributed, so half of the calls at an office contribute to the outgoing calling-to-called message stream and half contribute to the outgoing called-to-calling message stream.

#### 3.1 The Calling-to-Called Message Stream

For an individual call, this stream consists of an Initial Address Message (IAM) followed by a Continuity check message (COT) followed by a Release message (REL) if the caller hangs up first and a Release Complete message (RLC) if the called party hangs up first. The time interval between sending the IAM and the COT message is the sum of the time it takes the IAM to reach the far-end office (i.e., the office at the other end of the trunk being seized), the time it takes the far-end office to put a loop on the trunk, and the time it takes the originating office to measure the return signal and determine that it is satisfactory (the originating office puts tone on the outgoing side of the trunk and listens to the incoming side). We assume the time interval between the transmission of the IAM (i.e., when the last bit of the IAM is sent on the signaling data link) and the sending of the COT message (i.e., the last bit of the COT message is placed in the outgoing signaling link transmit buffer) is

### Table 2

Simulation Results for the Type 2 Clustered Poisson Process

<table>
<thead>
<tr>
<th>ρ</th>
<th>$Q_{M/D/1}$ (ms)</th>
<th>$\tau_{bp}$ (ms)</th>
<th>$Q_{p}(0)$</th>
<th>$Q_{p}(\infty)$</th>
<th>$\Delta_{conv}$</th>
<th>$Q_{s}(0)$</th>
<th>$Q_{s}(\infty)$</th>
<th>$\Delta_{conv}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.7</td>
<td>5.7</td>
<td>1.08</td>
<td>0.92</td>
<td>0.6</td>
<td>1.75</td>
<td>0.57</td>
<td>2.8</td>
</tr>
<tr>
<td>0.5</td>
<td>2.1</td>
<td>8.6</td>
<td>1.23</td>
<td>0.90</td>
<td>1.9</td>
<td>1.64</td>
<td>0.68</td>
<td>4.5</td>
</tr>
<tr>
<td>0.8</td>
<td>8.6</td>
<td>21.4</td>
<td>1.25</td>
<td>0.85</td>
<td>3.7</td>
<td>1.41</td>
<td>0.74</td>
<td>10.6</td>
</tr>
</tbody>
</table>

[6] We do not consider call irregularities, and so the calls considered complete to ringing or busy. It is assumed that 75% of all calls complete to ringing and are answered, and for half of those calls the caller hangs up first. It is assumed that 13% of all calls are placed to busy lines, and 12% of all calls complete to ringing with no answer.

[8] We use the notation: IAM, COT, RLC, REL.
uniformly distributed between 0.25 and 0.75 seconds. In the context of the simple clustered arrival processes discussed above, the IAM/COT pair is a Type 2 stream.

The last message to be sent in the caller-to-called stream is the REL or RLC used for taking the call down. The time the REL or RLC message is sent will be a random time, $\tau_{REL}$ or $\tau_{RLC}$, after the IAM is transmitted. To specify the distribution for $\tau_{REL}$, we consider three possibilities: the called line is busy (13% of all calls), the call completes to ringing but there is no answer (12% of all calls), and the call is answered and the caller hangs up first (37.5% of all calls). If the called line is busy, it is assumed $\tau_{REL}$ is uniformly distributed between 3 and 6 seconds. If there is no answer, it is assumed $\tau_{REL}$ is uniformly distributed between 15 and 30 seconds. If the call is answered, $\tau_{REL}$ will be dominated by the call holding time, which is assumed to be exponential. For this case $\tau_{REL}$ will be assumed to be exponentially distributed and different means are studied.

The RLC message is sent when the call is answered and the called party hangs up first (37.5% of all calls). The distribution for $\tau_{RLC}$ is therefore assumed to be the same as for $\tau_{REL}$ when the call is answered and the caller hangs up first; namely, the distribution is exponential and different means are studied.

The REL and RLC messages sent in the Caller-to-Called Stream correspond to Type 2 clustered processes discussed above, since the time they are sent is a random time after the IAM is transmitted.

### 3.2 The Called-to-Calling Message Stream

For an individual call this stream consists of an Address Complete Message (ACM) followed by an Answer Message (ANM) if the call is answered, followed by a RLC or REL message depending on whether the caller or called party hangs up first, respectively. The timing of all these messages is related to when the switching system completes processing the IAM and either begins ringing or sends an IAM to another network. At that time the ACM is sent, and it is assumed that if answer occurs (which is 75% of the time) the ANM message is sent at a time uniformly distributed between 3 and 6 seconds after the time the ACM is sent. The conditions when RLC and REL messages are sent in the Called-to-Caller Stream correspond to the conditions when REL and RLC, respectively, are sent in the Caller-to-Called Stream. Let $\tau_{RLC}$ and $\tau_{REL}$ be the time after ACM is sent that the RLC or REL message, respectively, is sent. Then, corresponding to the busy condition (occurring in 13% of the calls), $\tau_{RLC}$ will be uniformly distributed between 3 and 6 seconds; for ringing with no answer (occurring in 12% of the calls) $\tau_{RLC}$ is uniformly distributed between 15 and 30 seconds; and when there is answer and the caller hangs up first (occurring in 37.5% of calls) $\tau_{RLC}$ is exponentially distributed with a mean that depends on call holding time. The REL message occurs in 37.5% of all calls, and $\tau_{REL}$ is exponentially distributed with mean determined by the call holding time.

For the Called-to-Caller Stream it is clear that the message arrival process is a clustered process with the ADC messages being the primary messages and the ANM, REL and RLC messages being the secondary messages. Since the secondary messages are sent at times that are random intervals from the time the ACM is placed in the outgoing signaling link transmit buffer, the primary/secondary pairs are Type 1.

### 3.3 The Composite Message Stream

Assuming half of the call arrivals are incoming and half are outgoing, the outgoing message stream will consist of the superposition of Caller-to-Called and Called-to-Caller Streams, with equal arrival rates for each stream type. Table 3 gives the corresponding number of outgoing messages per call for each message type. Also shown are the message lengths for the different message types. All messages except the IAM have fixed lengths. The IAM has various optional fields, and the length given for the IAM is a typical value for its total length.

### Table 3

<table>
<thead>
<tr>
<th>Message Type</th>
<th>Message Length (octets)</th>
<th>Messages per Call</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAM</td>
<td>67</td>
<td>0.5</td>
</tr>
<tr>
<td>ACM</td>
<td>20</td>
<td>0.5</td>
</tr>
<tr>
<td>COT</td>
<td>18</td>
<td>0.5</td>
</tr>
<tr>
<td>ANM</td>
<td>18</td>
<td>0.375</td>
</tr>
<tr>
<td>REL</td>
<td>22</td>
<td>0.5</td>
</tr>
<tr>
<td>RLC</td>
<td>17</td>
<td>0.5</td>
</tr>
<tr>
<td>MEAN</td>
<td>27.4</td>
<td>2.875</td>
</tr>
</tbody>
</table>

### 3.4 Simulation Results for ISUP Message Streams

The mean queueing delays were obtained for the above ISUP message stream model using simulation. A time duration of 23 minutes was simulated with a 6 minute "warm-up" period used before computing means. The results are shown in Table 4 for a link utilization $p = 0.8$ and a 56 Kbps link speed. The lengths of the different messages are shown in Table 3, and the corresponding service-time moments are: $S = 3.9$ ms and $S^2 = 22.1$ ms$^2$. The corresponding M/G/1 mean queueing delay is 11.3 ms.

The results show that the M/G/1 model gives a very good approximation, even for very short (e.g., 5 second) call holding times. The results of Section 2 provide some insight as to why this should be the case. From Tables 1 and 2 it follows that for $p = 0.8$, $\Delta_{comp} < 12 \tau_{bp}$ for both Type 1 and Type 2 clustered arrivals. For the ISUP message stream and $p = 0.8$ the mean M/G/1 busy period, $\tau_{bp}$, is 19.6 ms. Therefore, if the time intervals between primary and secondary events exceeds 235 ms (12×19.6), there should be little effect of arrival time.
dependence. In the ISUP message stream model developed above, the shortest time interval between a primary and secondary message was between the IAM and COT messages, and this interval was between 250 and 750 ms. All other intervals between primary and secondary messages are much longer. Therefore, it follows from the results of Section 2 that the intervals between primary and secondary messages are large enough that there should be no significant impact of dependent arrival times at utilizations below 0.8. As the utilization increases toward unity, however, the busy period lengths will increase and the effects of the dependencies will become important. This is discussed further in the next section.

4. APPROXIMATIONS FOR HEAVY TRAFFIC CONDITIONS

From the above results it is clear that the M/G/1 approximation works well for utilizations that are not close to 1. However, as ρ→1 the length of the busy periods will increase to the point that we can expect the interarrival time dependencies will have an effect on the queueing delay. In this section we look at some approximation techniques for characterizing signaling link queueing delays with ISUP message streams in heavy traffic situations (i.e., when ρ→1). The approximations we consider are based on the asymptotic approximation of the type developed by Whitt. The basic idea behind this method is to approximate the input arrival process with a renewal process that has the same mean arrival rate and an appropriate variability in the interarrival times, measured by the squared coefficient of variation, $c^2_d$, of the interarrival times of the approximating renewal process.

Let $N(t)$ denote the number of arrivals in the interval $(0,t)$ in the arrival process to be approximated. It has been found (e.g., see References[4][5]) that an important measure in considering approximations of arrival streams is the index of dispersion for counts (IDC), denoted by $I(t)$, and defined by

$$I(t) = \frac{\text{Var}[N(t)]}{E[N(t)]}, \quad t > 0.$$  (4)

If a process is a renewal process, $I(t)$ asymptotically approaches the squared coefficient of variation of the interarrival time distribution as $t \to \infty$ (see pg. 72, Reference[2]). The asymptotic approximation matches the behavior of $I(t)$ of the approximating renewal process and the actual process as $t \to \infty$. Therefore, for the asymptotic approximation

$$c^2_d = \lim_{t \to \infty} I(t).$$  (5)

The importance of the asymptotic approximation is that as $t \to \infty$, $I(t)$ measures the cumulative covariances among all of the interarrival times, and this cumulative covariance seems to be a major factor in determining queueing delays under heavy loads.[5] In addition, it can be proven to be asymptotically correct as ρ→1 in a number of situations (e.g., $\Sigma G_i / G/1$ queues).[9]

To apply the asymptotic approximation to ISUP traffic streams, we first estimate $E[N(t)]$ and $\text{Var}[N(t)]$ as $t \to \infty$. Let $C_i$ denote the number of call arrivals in the time interval $(0,t)$ and $m$ denote the random variable for the number of outgoing messages for a call. The number of messages that will be sent due to the calls that arrive in time interval $(0,t)$ is denoted by $M_t$ and given by

$$M_t = \sum_{i=1}^{C_t} c_i,$$  (6)

where $m_i$ is the number of messages for the $i^{th}$ call. We will use $M_t$ to approximate the behavior of $N_t$ as $t \to \infty$ as follows:

$$E[N(t)] = E[M_t] = C_t \bar{m},$$  (7)

$$\text{Var}[N(t)] \approx \text{Var}[M_t] = C_t \text{Var}[m] + \bar{m}^2 \text{Var}[C_t].$$  (8)

If we assume the call arrival process is Poisson, then $C_t = \text{Var}[C_t] = \lambda t$, where $\lambda$ is the mean call arrival rate. Using (7) and (8) it follows that:

$$I(t) \approx \frac{\bar{m}^2}{\bar{m}}$$  as $t \to \infty$.  (9)

For a wide class of clustered Poisson processes (9) can be shown to be exact (see pg. 190, Reference[2]).

One approach to approximate the signaling link queueing delays is to choose an analytically tractable model having an approximating arrival process with mean message

<table>
<thead>
<tr>
<th>Message Type</th>
<th>Call Holding Time (sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAM</td>
<td>11.6, 11.6, 11.7, 11.7</td>
</tr>
<tr>
<td>ACM</td>
<td>11.5, 11.5, 11.5, 11.6</td>
</tr>
<tr>
<td>COT</td>
<td>11.6, 11.6, 11.6, 11.7</td>
</tr>
<tr>
<td>ANM</td>
<td>11.4, 11.3, 11.3, 11.4</td>
</tr>
<tr>
<td>REL</td>
<td>10.9, 11.2, 11.4, 11.5</td>
</tr>
<tr>
<td>RLC</td>
<td>10.9, 11.2, 11.3, 11.6</td>
</tr>
<tr>
<td><strong>MEAN</strong></td>
<td><strong>11.3, 11.4, 11.5, 11.6</strong></td>
</tr>
</tbody>
</table>
arrival rate $\lambda \bar{m}$ and $c_2^2$ given by (9). This is the approach suggested by Delbrouck and Appenzeller\[11]\ where they used an interrupted poisson process\[12]\ as the approximating arrival process. Since the asymptotic approximation is only being considered for heavy traffic conditions, we will use the simpler approach of invoking the QNA approximations.\[10]\ Specifically, the mean queueing delay is approximated by $SP(c_2^2 + c_2^2)/(1 - \rho)$. Substituting (9) for $c_2^2$ in this expression gives the approximate mean queueing delay, $Q_{\text{Approx}}$, as:

$$Q_{\text{Approx}} = \frac{p \bar{m}^2}{2(1 - \rho)} \left( \frac{\bar{m}}{\bar{m}} + c_2^2 \right).$$

(10)

It is interesting to observe that the delay given by (10) is the same as the mean batch queueing delay (i.e., the queueing delay for the first message in a batch) of an M/G/1 batch arrival system with first and second moments of the batch size given by $\bar{m}$ and $\bar{m}^2$, respectively.\[6]\ This suggests a possibly more accurate approximation for clustered Poisson processes in heavy traffic would be the M/G/1 batch arrival system. That is, for heavy traffic conditions it would be assumed that all messages in a cluster (for ISUP streams this is all messages related to a call) would arrive to the queue (in our case at signaling link transmit buffer) in a batch. This would clearly give an upper bound on the queueing delays.

Applying the asymptotic approximation to the ISUP message stream, it is easily shown that $\bar{m} = 2.875$ and $\bar{m}^2 = 8.375$. Therefore, $c_2^2 = 2.91$, and the approximate mean queueing delays would approach $6.56\rho/(1 - \rho)$ as $\rho \to 1$. The M/G/1 approximation gives a mean queueing delay of $2.83\rho/(1 - \rho)$ for the ISUP traffic model given in Section 3. Thus the asymptotic approximation gives delays about 2.3 times higher than the M/G/1 model. Further studies are needed to establish the accuracy of the asymptotic approximation. Preliminary results suggest that it considerably over estimates the mean delays, and that this is caused by the Type II secondary arrivals being "smoothed." The asymptotic approximations do not capture this smoothing effect of Type II processes.

5. CONCLUSIONS

It has been shown that when call arrivals are Poisson, the M/G/1 model gives a good approximation for signaling link queueing delays out to very large utilizations ($\rho < 0.8$) even though the arrival times are dependent. Some simple clustered Poisson processes have been studied to provide some insight into how these types of systems behave and why the signaling links behave as they do. We have also shown how to apply asymptotic approximations for heavy traffic conditions, and for heavy traffic it appears that the bulk arrival M/G/1 model would give a good approximation for Type I clustered arrivals. For Type II arrivals processes the secondary arrivals are smoothed, causing lower queueing delays which the asymptotic approximation does not capture. This is an area needing further study.

REFERENCES


