SICS Technical Report
T2000:03

ISRN:SICS-T–2000/03-SE
ISSN 1100-3154

Capacity Study of Statistical Multiplexing for IP Telephony [1]

Anders Andersson

Swedish Institute of Computer Science
Box 1263, SE-164 29 Kista, Sweden

January, 2000

Abstract

Transmitting telephone calls over the Internet causes problems not present in current telephone technology such as packet loss and delay due to queueing in routers. In this undergraduate thesis we study how a Markov modulated Poisson process is applied as an arrival process to a multiplexer and we study the performance in terms of loss probability. The input consists of the superposition of independent voice sources. The predictions of the model is compared with results obtained with simulations of the multiplexer made with a network simulator. The buffer occupancy distribution is also studied and we see how this distribution changes as the load increases.

[1] This work was supported by Telia AB

Contents

1 Introduction

The growth of the Internet has been exceptional during recent years and is expected to continue in the future. Currently the Internet is mainly used for applications without real-time constraints such as file transfer, E-mail and WWW traffic. However in the future as technology develops more real-time applications are expected to appear on the Internet. One such application is Internet telephony. The callers voice is sampled, packetized and sent to its destination over the Internet. An overall perspective is shown below in Figure 1.

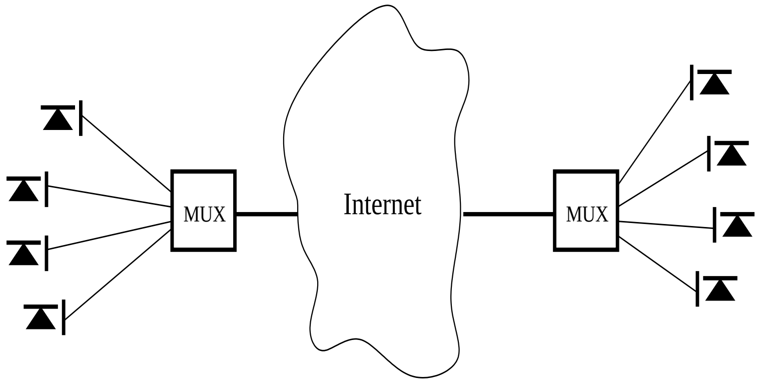

Figure 1: The overall perspective

There are several advantages with Internet telephony. It is much cheaper to send data over the Internet than over a telephone line. The equipment to send data packets costs less and is also easier to maintain. The connection-less nature of the Internet makes it possible to send packets only when the caller is talking thus leaving the silent periods free for other calls. This in contrast to current telephone technology where a certain amount of bandwidth is allocated for each caller during the call no matter if the caller is speaking or not. It also introduces some new problems. First packets may be lost in the network primarily due to limited buffer space in routers along the path towards the destination. For applications like Internet telephony, some losses are acceptable if the quality of the sound is not degraded in a noticeable way. Second, a packet may be delayed due to queuing in routers, propagation delay or audio coding. The delay is not such a big problem, however the variance to delay rate. If the delay varies a lot this will cause problems when the packets are played in the receiving end.

In this thesis we study a multiplexer which is multiplexing a number of connected calls over a common outgoing line. The multiplexer is equipped with a buffer to store incoming traffic before it is sent to the outgoing link. Calls connected to the multiplexer are assumed to be *independent*. Packets that arrive to a full buffer are considered as lost.

Several approaches have been suggested in the literature to solve this problem. In [1] is a multiplexer with infinite buffer studied with a *Stochastic Fluid Flow* model but it is shown in [8] that this model only work for a multiplexer under heavy load. A multiplexer with finite buffer is studied in [7] using the fluid flow model but it does not work well for small buffers. A two-state *Markov Modulated Poisson process* (MMPP) is used quite successfully in [5] to estimate the delay in a multiplexer with infinite buffer an it is suggested that the same approach for calculating the parameters of the MMPP can be used for a multiplexer with finite buffer, but in [6] it is shown that this does not work in the finite buffer case. Instead is a different method for finding the parameters of the MMPP developed. Also in [3] is the arrival process approximated with a two-state MMPP and a method called *asymptotic matching* is suggested for the calculation of the parameters of the MMPP and it is this approach we will follow in this thesis.

The outline of the rest of the thesis is as follows. In section 2 some basic concepts of computer networking are explained, and section 3 introduces some of the mathematics necessary to understand the rest of the thesis. In section 4 we discuss the modeling of the arrival process, and in the next section is of the approximation of the arrival process with an MMPP discussed. Section 19 introduces the mathematical model for packet loss analysis of the MMPP/D/1/K queue. A comparison between the model developed in section 19 and simulations made with a network simulator is made in section 7. And we study how the queue length varies for different loads in section 8.

2 Preliminaries

Some basic concepts of computer networking are explained in this section.

2.1 Internet, Internet protocol and IP telephony

The Internet is the largest computer network in the world. It consists of interconnected networks of different type. Originally it was designed for military use. The US military wanted a communication infrastructure that was robust and could find new paths for data by itself without human intervention if nodes were taken out of action by a large scale nuclear attack. Soon the academic world realized the potential of the Internet and many universities got connected to the Internet. During recent years have many companies found their way to the Internet. However many commercial applications such as telephony have little in common with the requirements the military and the academic world put on the Internet. Support for billing and *Quality of service*(QoS) is very poor but this is slowly beginning to change. The Internet is held together by the *Internet Protocol* (IP). IP is a *connection-less* protocol that allows data to be interpreted consistently as they travel across the network. Every computer connected to the Internet has a 32-bit IP-address. This address provides a uniform way of identifying the destination in the network. *IP telephony* uses IP together with UDP and RTP to transmit telephone calls over the Internet. The *User Datagram Protocol* (UDP) is a connection-less end-to-end protocol. UDP is used instead of the *Transmission Control Protocol* (TCP) which is a connection oriented end-to-end protocol, since there is no time to ask for retransmission of lost or garbled packets. The *Real-time Transport Protocol* (RTP) provides end-to-end delivery services for data with real-time characteristics.

2.2 Router

A *router* handles the movement of packets towards its destination on the Internet. It has by definition interfaces on more than one network. Every router has a *routing table* with at least two fields: a network number and the interface on which to send packets with that network number. When a packet arrives to the router it reads the destination address from the packet header and looks up the network number and interface in the routing table, then it forwards the packet on the appropriate network.

2.3 Multiplexing, statistical multiplexing

Multiplexing means that a resource is shared among a number of users. The device which handles the multiplexing is called a *multiplexer* (MUX). When a data communication link is multiplexed, two types of multiplexing are possible. Firstly we have *Time Division Multiplexing* (TDM) where periodically one user at a time gain control of the full capacity of the link for a short instance of time. Secondly we have *Frequency Division Multiplexing* (FDM) where each user get an exclusive share of

the links capacity.
If the sum of the peak rates P_i, is not allowed to exceed the output link rate, i.e., if

$$\sum P_i \leq C$$

then we say that the multiplexer is working under *peak rate allocation*. The advantages of peak rate allocation multiplexing are, no packet loss due to buffer overflow as well as minimal packet delay. The disadvantage is that bandwidth is wasted when the input links are sending at a lower rate than their peak rate P_i. This motivates the argument for *statistical multiplexing* where the sum of the connection peak rates is allowed to exceed the link capacity. The *statistical multiplexer* uses statistical knowledge about the users and the system in order to guarantee quality of service.

2.4 Buffer

The *buffer* is the physical entity in the router or switch where packets are stored before they are sent on the output links. The buffer is known as *waiting room* in queuing theory.

3 Mathematical preparations

In this section we give a short introduction to the theory of stochastic processes. It is assumed that the reader is familiar with the basic concepts of theory of probability such as expectation value, variance and covariance taught at the introductory course in theory of probability given at universities. Theorems are given without proofs. The interested reader should check the references for further details and proofs of the theorems. Reference [13] contains a thorough presentation of Markov processes, reference [12] treats renewal processes and [11] gives a detailed presentation of Poisson processes.

3.1 Stochastic Processes

A stochastic process is a random function which varies in time for instance. Its future values can not be precisely predicted, only with a certain amount of probability. This does not mean that the process behaves in a completely unpredictable manner, it's behaviour is governed by a random mechanism. Figure 2 shows five realizations of a stochastic process. As can be seen from the figure, the behaviour of the processes are a little different but a clear pattern of the behaviour can be seen. We are now ready for the definition of a stochastic process.

Definition 3.1 *A stochastic process with parameter space T is a family $\{X(t), t \in T\}$ of stochastic variables defined in the same sample space Ω. If T is an interval of real numbers, the process is said to have continuous time, if T is a sequence of integers, the process is said to have discrete time.*

The state space of a stochastic process may be continuous or it may be discrete. In this thesis will we only meet stochastic processes with a discrete state space. Also, in this thesis will only continuous time processes be considered, so stochastic process refers to continuous time stochastic process.

3.2 Markov Processes

An important class of stochastic processes are Markov processes. This class of processes have some special properties that make them manageable to treat mathematically. A Markov process is governed by the *Markov property* which states that the future behaviour of the process given its path only depends on its present state. Before we proceed into what a Markov process is we need to introduce the concept of *intensity*.

3.2.1 The Intensity concept

The concept of intensity comes from the question "If we know that an event at time T has not yet occurred, what is the probability that it does in just a moment?". It is possible to summarize this in one formula:

$$P(T \leq t + \Delta t \mid T > t) \tag{1}$$

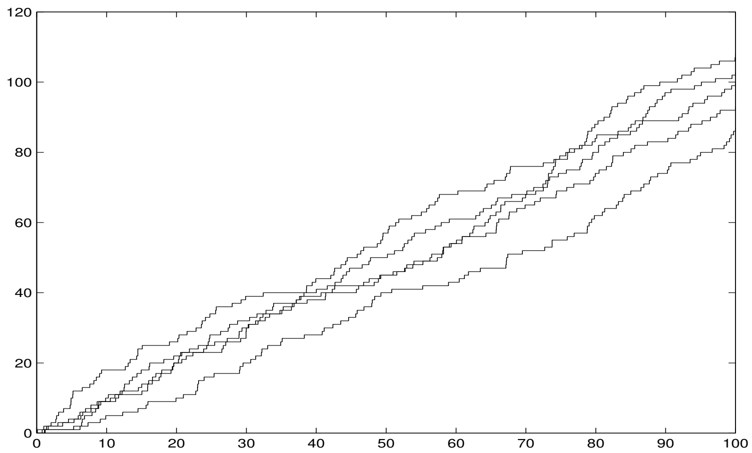

Figure 2: Five sample paths of a stochastic process

where Δt is small. By using the definition of conditional probability we obtain:

$$\begin{aligned} P(T \leq t + \Delta t \mid T > t) &= \frac{P(T \leq t + \Delta t, T > t)}{P(T > t)} \\ &= \frac{P(t < T \leq t + \Delta t)}{P(T > t)} \\ &= \frac{F(t + \Delta t) - F(t)}{1 - F(t)}, \end{aligned}$$

where F is T's distribution function. Assuming that F is differentiable, i.e T has density function f, we have

$$F(t + \Delta t) - F(t) = f(t)\Delta t + o(\Delta t).$$

Hence the probability in equation 1 is essentially equal to $\lambda(t) \cdot \Delta t$, where the function

$$\lambda(t) = f(t)/\{1 - F(t)\} \tag{2}$$

is called the *intensity function* for T. Intuitively, if T does not occur at t, is the probability that it is going to occur in the interval $(t, t + \Delta t]$ for small Δt approximately proportional to Δt, and the proportionality constant is the intensity $\lambda(t)$.

3.2.2 The Markov property

Definition 3.2 *Let $\{X(t), t \geq 0\}$ be a time continuous stochastic process which assumes non-negative integer values. The process is called a discrete Markov process if for every $n \geq 0$, time points $0 \leq t_0 < t_1 < \cdots < t_n < t_{n+1}$ and states $i_0, i_1, ..., i_{n+1}$ it holds that*

$$\begin{aligned} P(X(t_{n+1}) = i_{n+1} \mid X(t_n) = i_n, X(t_{n-1}) = i_{n-1}, ..., X(t_0) = i_0) \\ = P(X(t_{n+1}) = i_{n+1} \mid X(t_n = i_n)). \end{aligned}$$

The definition states that only the present state gives any information of the future behaviour of the process. Knowledge of the history of the process does not add any new information.

In this thesis we only concern ourselves with processes who have time-homogeneous properties. In other words the intensity of leaving a state is constant in time. So it is natural to make the following definition which states that the transition probabilities only depends on which state the process is in and not on the time.

Definition 3.3 *Let $\{X(t), t \geq 0\}$ be a discrete Markov process. If the conditional probabilities $P(X(s+t) = j \mid X(s) = i)$, for $s, t \geq 0$, do not depend on s, the process is said to be time homogeneous. Then we define the transition probability functions $p_{ij}(t) = P(X(t) = j \mid X(0) = i)$ and the transition matrix $\mathbf{P}(t)$, whose element with index (i,j) is $p_{ij}(t)$.*

Note that $p_{ii}(0) = 1$ and $p_{ij}(0) = 0$ for $i \neq j$, so that $\mathbf{P}(0) = I$. In many cases it is necessary to study the time between occurrences and therefore we can make the following definition.

Definition 3.4 *Let $X(t), t \geq 0$ be a Markov process. We call the times $0 \leq T_1 < T_2 < T_3 < \ldots$ such that at T_i the process makes a transition from one state to another the occurrence times. We also introduce the duration $Y_n = T_n - T_{n-1}$, where $T_0 = 0$.*

Figure 3 illustrates the durations of a stochastic process. At times $T_1, T_2, \ldots$ are there state transitions and the durations $Y_1, Y_2, \ldots$ are the time spent in a particular state before the next state transition.

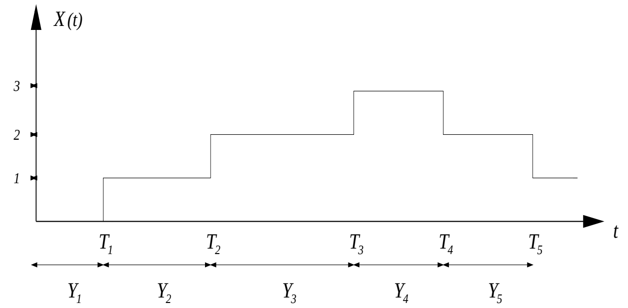

Figure 3: The duration time

3.2.3 The Intensity matrix

It is desirable to characterize the behaviour of a Markov process in one matrix. It turns out that the derivative of $\mathbf{P}(t)$ yields such an form.

Let $q_{ij} = p'_{ij}$, for $i \neq j$ we call q_{ij} the transition intensity from state i to state j.

Definition 3.5 *Let $X(t), t \geq 0$ be a discrete Markov process. Assume that $q_{ij} = p'_{ij}(0) \geq 0$ for $i \neq j$ and $q_{ii} \leq 0$ such that*

$$P(X(t+h) = j \mid X(t) = i) = q_{ij}h + o(h)$$

and

$$P(X(t+h) \neq i \mid X(t) = i) = -q_{ii}h + o(h) = q_i h + o(h),$$

where $q_i = -q_{ii}$, and

$$q_i = \sum_{j \neq i} q_{ij}.$$

The indices i and j shall run through the whole state space of the process. We will name $q_{ij}, i \neq j$, the transition intensity from state i to state j. And we define the intensity matrix $\mathbf{Q}$, whose elements with index (i,j) are q_{ij}.

Every Markov process in this thesis will meet the assumption made in the definition above. Another name for the intensity matrix is the *generator matrix*. The summation condition $q_i = \sum_{i \neq j} q_{ij}$ is quite natural since q_{ij} is the transition intensity from state i to state j, and if these intensities are summed over $i \neq j$ we should receive the total intensity out of state i, which is given by q_i. This also means that the rows of $\mathbf{Q}$ sum up to zero.

3.2.4 Birth-Death processes

A useful class of Markov processes when analyzing queueing systems are *birth-death* processes. The only possible state transitions in this kind of processes are from i to $i-1$ or from i to $i+1$. The transition intensity from state i to $i+1$ is designated $\lambda_i \geq 0$ for $i \geq 0$ and the transition intensity from state i to state $i-1$ is designated $\mu_i \geq 0$ for $i \geq 1$.

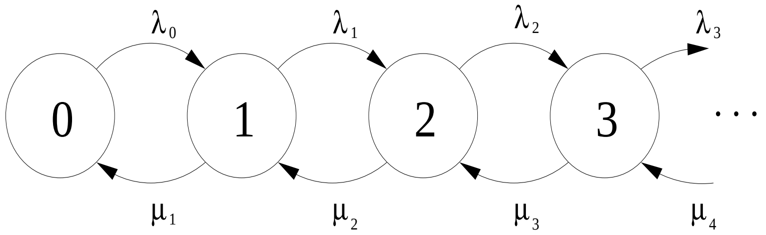

Figure 4: Model graph for a Birth-death process

The state space of the birth-death process is $\{0, 1, 2, 3, ...\}$. The intensity matrix will be of tri-diagonal type since there are only two ways of leaving a state. Hence, we have the intensity matrix

$$\mathbf{Q} = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0 & \ldots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & 0 & \ldots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & 0 & \ldots \\ \vdots & \vdots & & & & \ddots \end{pmatrix}$$

As mentioned earlier, certain types of queuing systems are suitably modeled by birth-death processes. The numbers $\{\lambda_i\}$ and $\{\mu_i\}$ are interpreted as the arrival rate of the queue and service rate of the server, respectively.

3.3 The Poisson process

This section gives a short introduction to Poisson processes. A more complete description of this type of processes can be found in [11]. Poisson processes are one of the most important classes of stochastic processes, and find applications in diverse areas of science such as physics, teletraffic modeling and biology.
We introduce a counter which count the number of occurrences from a starting point, and set

$$X(t) = number\ of\ occurrences\ in\ the\ interval\ (0, t]$$

Thus $X(t)$ will increase by one for every occurrence. These increases of one will be called *increments* in the sequel. In many applications is it realistic to assume that occurrences in disjunct intervals are independent of each other, i.e the process $X(t), t \in T$ is said to have independent increments. If the distribution of the increments does not change in time, the process $X(t)$ is said to have stationary increments. If the number of occurrences after time t follows the probability function

$$p_{X(t)}(x) = P(X(t) = x) = e^{-\lambda t}\frac{(\lambda t)^x}{x!}\ for\ x = 0, 1, 2, ...$$

where λ is the intensity of the occurrences, we say that $X(t)$ is a Poisson process. Let us now summarize these requirements in a definition.

Definition 3.6 *A Poisson process with intensity* $\lambda > 0$ *is a stochastic process* $X(t), t \geq 0$ *such that*

$$(i)\ X(t)\ is\ integervalued,\ increasing\ and\ X(0) = 0,$$
$$(ii)\ X(t)\ has\ independent\ and\ stationary\ increments,$$
$$(iii)\ X(t) \in Po(\lambda t).$$

This is one of many possible definitions of the Poisson process. In [11] there are three different definitions given, all of them equivalent, however proofs of some properties may be considerably simplified with the right choice of definition. In the next section are some important properties of Poisson processes stated.

3.3.1 Properties

Since $X(t) \in Po(\lambda t)$ we know that

$$E[X(t)] = \lambda t\ and\ Var[X(t)] = \lambda t$$

It can easily be shown that if $X(t)$ is Poisson distributed then the increments from s to a later point $s + t$ is also Poisson distributed. We conclude this in a theorem.

Theorem 3.1 *Let $X(t)$ be a Poisson process and $s, t \geq 0$, then the following is true*

$$X(s+t) - X(t) \in Po(\lambda t) \tag{3}$$

The theorem states that if you move the starting time to s and observe what occurs after s you simply get a new Poisson process. It can be shown that the Poisson process can only increase one unit at a time. The next theorem states the distribution of the duration Y_n defined as in definition 3.4.

Theorem 3.2 *Let $X(t), t \geq 0$ be a Poisson process with intensity λ and let $Y_1, Y_2, Y_3, ...$ be the durations. Then $Y_n \in Exp(\frac{1}{\lambda})$ for $n = 1, 2, 3, ...$*

It can also be shown that the Y_n:s are independent of each other. Another important property of Poisson processes is the so called *lack of memory* property.

Theorem 3.3 *If $T \in Exp(\frac{1}{\lambda})$ then we have*

$$P(T > t + s \mid T > s) = e^{-\lambda t} = P(T > t) \tag{4}$$

The theorem states that if at time s, we know that there has been no occurrence, then the residual waiting time until an occurrence is $Exp(\frac{1}{\lambda})$-distributed, i.e . the residual time has the same distribution and expectation value as T itself. This is the reason why it is sometimes said that the exponential distribution is memoryless, the lack of memory property.

3.4 Renewal processes

A large class of stochastic processes are *renewal processes.* This class of processes are used to model independent identically distributed occurrences.

Definition 3.7 *Let $Y_1, Y_2, Y_3, ...$ be independent identically distributed and positive stochastic variables, and set $T_n = Y_1 + \ ... \ + Y_n$. Then the process $X(t), t \geq 0$, where*

$$X(t) = max\{n : T_n \leq t\},$$

is called a renewal process.

The name renewal process is motivated by the fact that every time there is an occurrence the process "starts all over again", it renews itself. Also note that since Y_i and Y_j are independent for $i \neq j$, we have $Cov(Y_i, Y_j) = 0$. The Poisson process is a special case of a renewal process where the time between occurrences is exponentially distributed. Renewal processes usually do not possess the exact properties Poisson processes possess ($X(t) \in Po(\lambda t)$ or $E[X(t)] = \lambda t$), however it is possible to prove some important asymptotic results for renewal processes when $t \to \infty$. We state three important propositions below taken from [12], readers not familiar with the concept of *almost sure* convergence can consult [11] page 149.

Theorem 3.4 *Let* $\{X(t), t \geq 0\}$ *be a renewal process with durations* Y_n *such that* $\mu = E[Y_1]$ *and* $\sigma^2 = Var[Y_1]$ *, then*

$$
\begin{array}{ll}
(i) & \dfrac{X(t)}{t} \to \dfrac{1}{\mu}, \ \textit{as } t \to \infty, \textit{almost surely.} \\
(ii) & \dfrac{E[X(t)]}{t} \to \dfrac{1}{\mu}, \ \textit{as } t \to \infty. \\
(iii) & \dfrac{Var[X(t)]}{t} \to \dfrac{Var[Y_1]}{E[Y_1]^3}, \ \textit{as } t \to \infty.
\end{array}
$$

The simple definition of renewal processes indicate that many types of stochastic processes can be described as renewal processes. It is often the case that a complex stochastic model has one or more embedded renewal processes, this fact can be of great help for the analysis of such processes and is basic to the idea of *regeneration*, which allows a process to be decomposed into independent identically distributed blocks of random length.

3.5 Superposition of Poisson processes

In previous sections we have seen examples of different types of stochastic processes. In this section we will see what happens when we take the superposition of Poisson processes, i.e add them together. We will see that the superposition of Poisson processes is also a Poisson process. This is concluded in the following theorem.

Theorem 3.5 *Let* $\{X_i(t), t \geq 0, i = 1, 2, ..., m\}$ *denote independent Poisson processes with intensities* $\lambda_1, \lambda_2, ..., \lambda_m$ *respectively. Then the superposition*

$$Z(t) = \sum_{i=1}^{m} X_i(t)$$

is a Poisson process with intensity $\lambda = \lambda_1 + \lambda_2 + ... + \lambda_m$.

This theorem will be important to us later on when we approximate the arrival process at the multiplexer with a Markov modulated Poisson process. An example of a superposition of three Poisson processes is shown in figure 5.

However nothing can be said about the superposition of renewal processes. There is nothing that ensures that the renewal property is preserved when it is superposed.

3.6 Index of dispersion for intervals (IDI)

To describe the dependence between successive arrivals of an arrival process an index of dispersion of intervals (IDI) is often used. In the literature it is used as a measure of the burstiness of a signal. Let $\{Y_k, k \geq 1\}$ be the durations of a stochastic process. We assume that $\{Y_k, k \geq 1\}$ is stationary, by which we mean that the joint distribution of $(Y_k, Y_{k+1}, \cdots, Y_{k+m})$ is independent of k for all m. The sum of k consecutive inter-arrival times is denoted $S_k = Y_1 + Y_2 + \cdots + Y_k$. The IDI or the k-interval squared coefficient of variation is defined as follows.

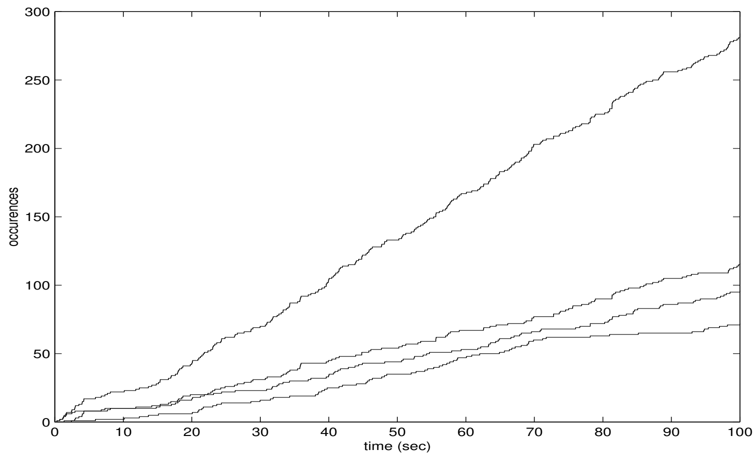

Figure 5: Superposition of Poisson processes

Definition 3.8 *The index of dispersion for intervals (IDI) is defined as*

$$c_k^2 = \frac{kVar(S_k)}{[E(S_k)]^2} \tag{5}$$

Together with the fact that the sum $S_k = Y_1+Y_2+\cdots+Y_k$ of k consecutive durations of a Poisson process with intensity λ is $\Gamma(k, \frac{1}{\lambda})$ distributed (see for instance [11] page 202), one can easily verify that the IDI of a Poisson process will be equal to one for all k. This is stated in the next theorem.

Theorem 3.6 *The index of dispersion for intervals (IDI) of a Poisson process*

$$c_k^2 = \frac{kVar(S_k)}{[E(S_k)]^2} = 1$$

for all k.

This fact suggests that it can be used to measure the deviation from a Poisson process. For $k > 1$, c_k^2 measures the cumulative covariance (normalized by the square of the mean) among k consecutive inter-arrival times. In order to see this we note that the variance of S_k can be rewritten as

$$\begin{aligned} Var(S_k) &= Var(Y_1 + Y_2 + \cdots + Y_k) \\ &= kVar(Y_1) + 2\sum_{j=1}^{k-1}(k-j)Cov(Y_1, Y_{1+j}) \end{aligned} \tag{6}$$

We can now rewrite the expression for c_k^2 in the following manner

$$c_k^2 = \frac{k\ Var(S_k)}{[E(S_k)]^2} = \frac{k\ Var(S_k)}{k\ [E(Y_1)]^2}$$

$$
\begin{aligned}
&= \frac{k\ Var(Y_1) + 2\sum_{j=1}^{k-1}(k-j)Cov(Y_1, Y_{1+j})}{k\ [E(Y_1)]^2} \\
&= c_1^2 + \frac{2\sum_{j=1}^{k-1}(k-j)Cov(Y_1, Y_{1+j})}{k\ [E(Y_1)]^2} \qquad (7)
\end{aligned}
$$

From equation 7 it is clear that c_k^2 measures the cumulative covariance (normalized by the square of the mean) among k consecutive inter-arrival times. Inspection of the equations above and using the observation made about the covariance between durations in section 3.4, we may derive a similar proposition about renewal processes as in theorem 3.6.

Theorem 3.7 *For a renewal process the following holds for all $k \geq 1$,*

$$c_k^2 = \frac{Var[Y_1]}{E[Y_1]^2} = c_1^2.$$

Deviations from the renewal property can be detected by searching for fluctuations in the IDI sequence $\{c_k^2, k \geq 1\}$.
The *index of dispersion of counts* uses the counting process instead of the durations to characterize an arrival process.

Definition 3.9 *Let $X(t)$ denote the counting process associated with an arrival process. The index of dispersion for counts (IDC) is defined as*

$$I(t) = \frac{Var[X(t)]}{E[X(t)]},\ t > 0.$$

Since $I(t) = 1$ when $X(t)$ is a Poisson process, the Poisson process serves as a reference process for the IDC. With the help of the propositions made in theorem 3.4 we are able to conclude what happens with $I(t)$ for a renewal process when $t \to \infty$. First we rewrite the expression for the definition of $I(t)$.

$$I(t) = \frac{Var[X(t)]}{E[X(t)]} = \frac{Var[X(t)]/t}{E[X(t)]/t}$$

Now we let $t \to \infty$ and use theorem 3.4.

$$\lim_{t\to\infty} I(t) = \frac{Var[Y_1]/E(Y_1)^3}{1/E[Y_1]} = \frac{Var[Y_1]}{E[Y_1]^2} = c_1^2. \qquad (8)$$

This expression links the two measures together when renewal processes are being studied. It is interesting to see what happens with $I(t)$ if the superposition of a number of renewal processes is taken into consideration. We let $X(t)$ denote a renewal processes, and let $I_n(t)$ denote the IDC of a superposition of n identical independent renewal processes.

$$I_n(t) = \frac{Var[\sum_{i=1}^{n} X(t)]}{nE[X(t)]} = \frac{Var[X(t)]}{E[X(t)]} = I(t) \qquad (9)$$

Equation 9 shows that $I_n(t)$ is independent of n when identical renewal processes are superposed, and we are able to use equation 8 to draw conclusions about $I_n(t)$. Note that making a similar calculation for IDI is an inherently more difficult problem since we must study the durations of the superposed process which in many cases are very difficult to calculate.

4 Modeling the arrival process

This section describes modeling the arrival process. We start with a single source and proceed into the superposition of independent identically distributed sources.

4.1 Properties of a single source

A single source is a telephone conversation where packets from the caller are transmitted. It does not describe the packets sent by the receiver. In this section we introduce the reader to two descriptions of the behaviour of a single source. The first description with fixed inter-arrival time between packets should be considered as the exact description, whereas the second description with exponentially distributed inter-arrivals should be viewed as an approximation of the exact description.

4.1.1 Deterministic inter-arrivals

The source is characterized by a stream of packets with fixed inter-arrival times T during talkspurts (ON-period) and no arrivals during silences (OFF-period). All packet are assumed to be of equal size. The number of packets in a talkspurt denoted with the stochastic variable N_b is assumed to be geometrically distributed on the positive integers with mean n. This means that we can never have zero packets in a talkspurt. This variant of the geometric distribution is sometimes called *first success* distribution (see for instance [Gut] page 258), and has the probability function

$$P(N_b = k) = qp^{k-1}, k = 1, 2, 3, ... \tag{10}$$

where q represents the probability that a packet is the last one in a talkspurt. This means that $p = \frac{n-1}{n}$. This fact implies that the ON-periods have expectation value $\alpha = nT$, where n is the expected value of the number of packets in a talkspurt. Figure 6 illustrates the behaviour of a single source with deterministic inter-arrivals.

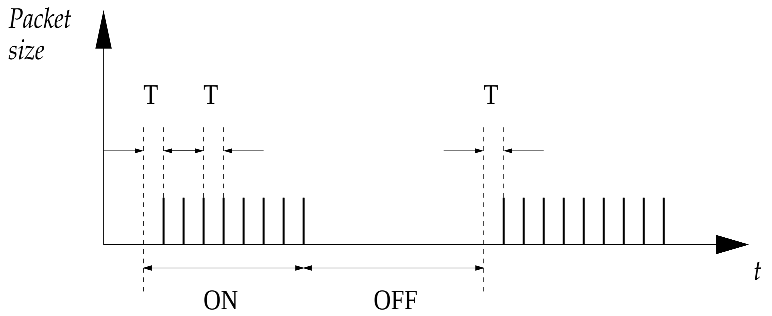

Figure 6: Characteristic of a single source. Deterministic inter-arrivals

We assume that the OFF-periods are exponentially distributed with mean β, which is well documented and discussed in [9]. A voice source may be viewed as a two state birth-death process with birth rate β and death rate α. One state represents the idle periods and the other represents the talkspurts. While in talkspurts packets are generated with a rate of $\frac{1}{T}$packets per second. Figure 7 shows the model graph of the two state birth-death process.

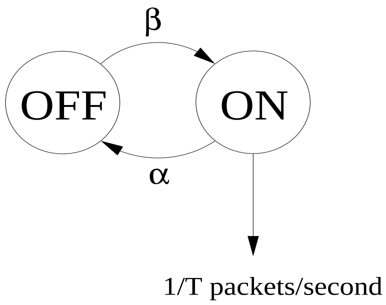

Figure 7: Model graph of a single source. Deterministic inter-arrivals

4.1.2 Exponentially distributed inter-arrivals

In this description the inter-arrival times between packets of equal size during a talkspurt are exponentially distributed with mean T. We let $\tau \in Exp(\frac{1}{T})$ denote the stochastic variable which describes the inter-arrivals during talkspurts, and N_b be the geometrically distributed stochastic variable with the probability function stated in equation 10 with mean n describing the number of packets in a talkspurt. Moreover τ and N_b are assumed to be independent. We want to show that the ON-periods denoted U are exponentially distributed. If we condition on the number of packets in a talkspurt, then $U|N_b = k \in \Gamma(k, \frac{1}{T})$ according to [11] page 202, and the density function is

$$f_{U|N_b=k}(t) = \frac{1}{\Gamma(k)} \frac{t^{k-1}}{T^k} e^{-\frac{t}{T}}.$$

We are now able to calculate the density function of U with the help of the law of total probability.

$$\begin{aligned} f_U(t) &= \sum_{k=1}^{\infty} f_{U|N_b=k}(t) p_{N_b}(k) = \sum_{k=1}^{\infty} \frac{1}{\Gamma(k)} \frac{t^{k-1}}{T^k} e^{-\frac{t}{T}} (1-p) p^{k-1} \\ &= \frac{1-p}{T} e^{-\frac{t}{T}} \sum_{k=1}^{\infty} \frac{(tp/T)^{k-1}}{(k-1)!} = \frac{1-p}{T} e^{-\frac{t}{T}} \sum_{k=0}^{\infty} \frac{(tp/T)^k}{k!} \\ &= \frac{1-p}{T} e^{-\frac{t}{T}} e^{\frac{p}{T}t} = \frac{1-p}{T} e^{-\frac{(1-p)}{T}t} \end{aligned} \quad (11)$$

We see that the length of a talkspurt is exponentially distributed and the mean length of a talkspurt is the same as in the deterministic inter-arrival case(nT). Figure 8 illustrates the behaviour of a single source with exponentially distributed inter-arrivals.

As in the previous section the OFF-periods are assumed to be exponentially distributed with mean β. Because of the exponentially distributed inter-arrival times

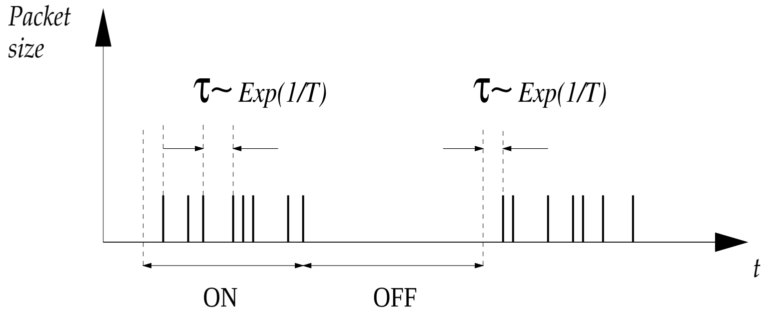

Figure 8: Characteristic of a single source. Exponentially distributed inter-arrivals

during a talkspurt, the emission of packets during an ON-period can be regarded as a Poisson process with intensity T. We can use the two state birth-death process to describe the packet generation with one state representing the idle periods and the other state representing the talkspurts where packets are generated as a Poisson process with intensity T. The model graph of the birth-death process is shown in figure 9.

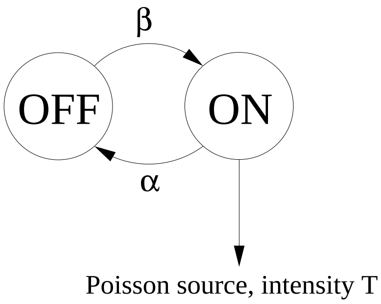

Figure 9: Model graph of a single source. Exponentially distributed inter-arrivals

4.1.3 A single source as a renewal process

To describe the renewal periods we introduce the stochastic variable X which is exponentially distributed with $E(X) = \frac{1}{\beta}$ and describes the silence periods between talkspurts. The inter-arrival time for packets of a single source with deterministic inter-arrival time is T with probability $p = \frac{n-1}{n}$ and of length $X+T$ with probability $1-p = \frac{1}{n}$. We introduce the stochastic variable

$$Y = \begin{cases} T & \text{with probability } p = \frac{n-1}{n}, \\ T+X & \text{with probability } 1-p = \frac{1}{n}. \end{cases}$$

When we have exponentially distributed inter-arrival times we let Y be

$$Y = \begin{cases} \tau & \text{with probability } p = \frac{n-1}{n}, \\ \tau+X & \text{with probability } 1-p = \frac{1}{n}. \end{cases}$$

Considering sequences of independent random variables distributed according to Y, a single source can now be seen as a renewal process both when we have deterministic inter-arrivals and exponentially distributed inter-arrival times during a talkspurt.

4.2 Values of the parameters

In order to perform actual calculations of the model and to perform simulations we need to set the parameters to designated values. In this thesis we use the same values of the parameters as in [9] and [8], i.e.

- The packet size is 64 bytes.
- Inter-arrival time between packets during a talkspurt is 16 ms
- The number of successive packets in one talkspurt is geometrically distributed on the positive integers with a mean of 22.
- The idle time between two successive bursts is exponentially distributed with a mean of 0.65 seconds.

These values will be used in calculations and simulations using a network simulation package Network Simulator(ns) to verify the model described in this thesis.

4.3 The superposition of independent voice sources

The superposition of voice sources can be viewed as a birth-death process where the states represent the number of sources that are currently in the ON-state. Here state i represents, i sources are active(in a talkspurt). We will refer to the birth-death process as the *phase process* $J(t)$. The birth rate is given by the mean of the exponentially distributed idle periods, and we denote the mean as $\frac{1}{\beta}$. The death rate is determined by the mean of duration of the talkspurts and is denoted $\frac{1}{\alpha}$. The probability p_{on} that a source is on is given by

$$p_{on} = \frac{\alpha}{\alpha + \beta}.$$

The limiting state probabilities π_i, that the phase process is in state i is determined by the binomial distribution with parameter p_{on}. And the probability function for π_i $(i = 0, \cdots, N)$ is

$$\pi_i = \binom{N}{i} p_{on}^i (1 - p_{on})^{N-i}. \tag{12}$$

Where N is the number of sources that are superpositioned. Furthermore is the intensity matrix of the phase process

$$
\mathbf{Q} = \begin{pmatrix}
-N\beta & N\beta & 0 & 0 & & \\
\alpha & -[\alpha + (N-1)\beta] & (N-1)\beta & 0 & & \\
0 & 2\alpha & -[2\alpha + (N-2)\beta] & (N-2)\beta & & \\
 & & . & & & \\
 & & . & & & \\
 & & . & & & \\
 & & . & & & \\
 & & . & & & \\
 & & & (N-1)\alpha & -[(N-1)\alpha + \beta] & \beta \\
 & & & 0 & N\alpha & -N\alpha
\end{pmatrix}
$$

So far we have not been talking about arrivals from voice sources. In the following two subsections will we see how the phase process can be used to modulate the arrivals from the voice sources.

4.3.1 Markov modulated rate process

If sources of the type described in section 4.1.1 are superpositioned, the *Markov modulated rate process* (MMRP) provides a good approximation of the new arrival process. When the phase process, as described in the previous section, is in state i, are packets at a rate of $\frac{i}{T}$ packets per second generated. Figure 10 shows the model graph of the arrival process.

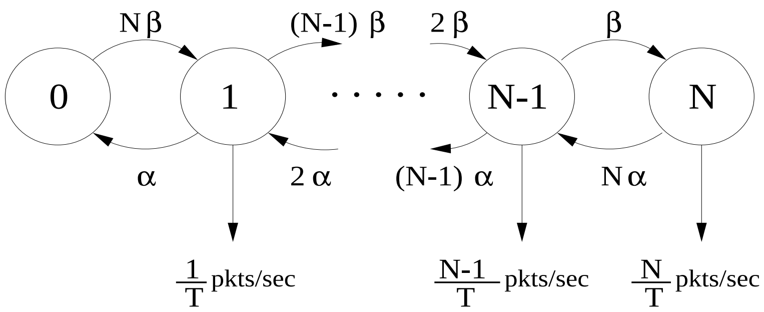

Figure 10: Superposition of N voice sources. Deterministic inter-arrivals

It is difficult to describe the arrival process mathematically when we have deterministic inter-arrival times. This is because the different sources will be unsynchronized and thus is the deterministic interarrival time property not preserved, and this behaviour is difficult to describe mathematically.

4.3.2 Markov modulated Poisson process

The *Markov modulated Poisson process* (MMPP) is a widely used tool for analysis of teletraffic models. It describes the superposition of sources of the type described in section 4.1.2. As for the MMRP model i sources are on when the phase process is in state i. The model graph of the MMPP is shown in figure 11.

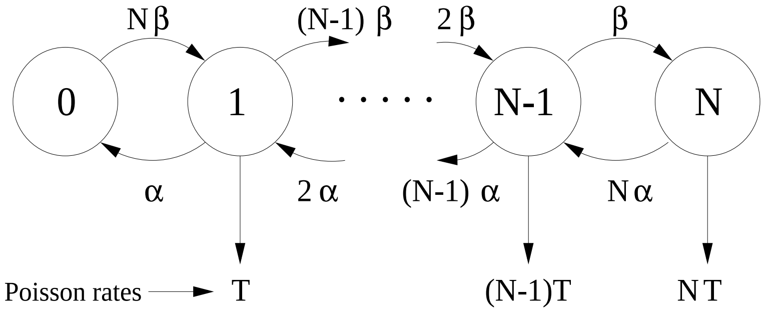

Figure 11: Superposition of N voice sources. Exponentially distributed inter-arrivals

Because of theorem 3.5 are we able to add the intensities of the sources that are currently in a talkspurt and receive a new Poisson process for the superposition, due to this fact the MMPP is much simpler to treat mathematically than the MMRP.

5 Approximating the MMRP with MMPP

As stated in section 4.3.1 is it difficult to find a mathematical description for the superposition of sources of the type introduced in section 4.1.1. But if we superpose sources of the type introduced in section 4.1.2 we are allowed to add the intensities of the sources and get a new Poisson process according to theorem 3.5. What we would like to do is to approximate the Markov modulated rate process with a Markov modulated Poisson process. We motivate this approximation by looking at the k-interval squared coefficient of variation for both the MMRP and MMPP and we will see that they behave in a similar way both as a single source and superpositioned sources.

5.1 The squared coefficient of variation of a single source

In order to calculate the IDI of a single source we make use of the stochastic variables Y and X introduced in section 4.1.3. By using the values of the parameters introduced in section 4.2 we receive $E(X) = \frac{1}{\beta} = 650\ ms$ and $Var(X) = \frac{1}{\beta^2} = 422.5\ ms$. To make the calculations easier to perform we introduce the stochastic variable Z defined as

$$Z = \begin{cases} 1 & \text{with probability } p = \frac{21}{22}, \\ 0 & \text{with probability } 1 - p = \frac{1}{22}. \end{cases}$$

With the help of Z it is now easy to calculate the variance and expectation value of Y. We will do the necessary calculations for both our descriptions of a single voice source and compare the results.

5.1.1 c_1^2 with deterministic inter-arrival times

We start with the calculation of the expectation value of Y.

$$\begin{aligned} E(Y) &= E(TZ + (T+X)(1-Z)) = E(T + X(1-Z)) = T + \frac{1}{\beta}E(1-Z) \\ &= T + \frac{1}{\beta}(1-p) = 16 + \frac{650}{22} = 45.55\ ms \end{aligned} \tag{13}$$

The variance is calculated in a similar way. Note that $E[(1-Z)^2] = E[(1-Z)]$. With this observation in mind are we able to calculate the variance of the inter-arrival times.

$$\begin{aligned} &Var(Y) = V(TZ + (T+X)(1-Z)) = V(T + X(1-Z)) = V(X(1-Z)) \\ &= E[X^2(1-Z)^2] - [E(X)E(1-Z)]^2 = E(X^2)E(1-Z) - [E(X)E(1-Z)]^2 \\ &= E(1-Z)[E(X^2) - E(X)^2E(1-Z)] = (1-p)[E(X^2) - E(X)^2(1-p)] \\ &= (1-p)[V(X) - pE(X)^2] = (1-p)[\frac{1}{\beta^2} + p\frac{1}{\beta^2}] \\ &= \frac{1}{\beta^2}(1-p^2) = 37536.16ms \end{aligned}$$

For simplicity have we used V to denote the variance in the calculations above. Thus, we have all the information we need to calculate c_1^2 of a single source.

$$c_1^2 = \frac{Var[Y]}{[E(Y)]^2} = \frac{1-p^2}{(T\beta+1-p)^2} = \frac{37536.16}{45.55^2} = 18.1 \qquad (14)$$

The high value of c_1^2 reflects the bursty nature of a single source.

5.1.2 c_1^2 with exponentially distributed inter-arrival times

The calculations are similar to the calculations made in the previous section. In fact the calculation of $E(Y)$ is the same as in equation 13. Hence

$$E(Y) = T + \frac{1}{\beta}(1-p) = 45.55\ ms.$$

The calculation of the variance in Y differ slightly from the calculations made in the previous section since the inter-arrival times between the packets in a talkspurt is now exponentially distributed. We let $\tau \in Exp(\frac{1}{T})$ be the inter-arrival time between packets in a talkspurt.

$$\begin{aligned} Var(Y) &= V(\tau Z + (\tau + X)(1-Z)) = V(\tau + X(1-Z)) \\ &= V(\tau) + V(X(1-Z)) = T^2 + \frac{1}{\beta^2}(1-p^2) = 37792.16\ ms \end{aligned}$$

Note that we have assumed that τ and X are independent. And the squared coefficient of variation is

$$c_1^2 = \frac{Var[Y]}{[E(Y)]^2} = \frac{\beta^2 T^2 + 1 - p^2}{(\beta T + 1 - p)^2} = 18.2 \qquad (15)$$

The value of c_1^2 differ only slightly between deterministic and exponential inter-arrival times.

5.2 The squared coefficient of variation of a superposition of voice sources

To calculate the squared coefficient of variation of a superposition of voice sources is a slightly more difficult task. However in [9] is a formula derived for IDI when $k = 1$, where N sources are being superpositioned. The resulting formula is

$$c_{1N}^2 = 1 - \frac{2}{N+1} + (\frac{1-p}{T\beta + 1 - p})^{N+1}(\frac{2}{1-p} - \frac{2N}{N+1}). \qquad (16)$$

Table 1 shows the value of c_{1N}^2 for a number of different values of N. The table shows that the c_{1N}^2 drops rapidly from 18.1 when $N = 1$ towards one as N grows.

Figure 12 shows c_{kN}^2 versus k for k between 1 and 2000 and n equal to $1, 10, 60$ and

N	1	4	10	20	50	100
c_{1N}^2	18.10	5.47	1.18	0.91	0.96	0.98

Table 1: The theoretical value of c_{1N}^2 for different N

130, as a reference have we added the value of c_{kN}^2 for a Poisson process. Data was obtained from simulations using a Matlab program. The solid line shows the c_{kN}^2 for sources with deterministic inter-arrival times between packets during a talkspurt, and the dashed lines show the c_{kN}^2 for sources with exponentially distributed inter-arrival times.

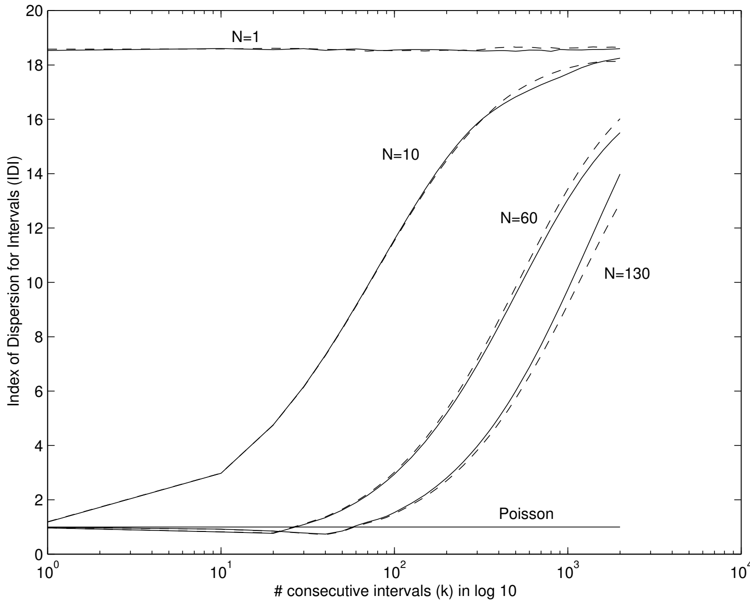

Figure 12: k-interval squared coefficient of variation curves for superposition of N voice sources.

We see in the figure that the two descriptions of a single source behave in a similar way when they are superpositioned. The figure also shows that the superpositioned arrival process behaves as a Poisson process if we look at it for a short instant of time but it is much burstier if we study it over a longer period of time. In other words, the cumulative covariances play an important role in the superpositioned process when we look at the arrival process over a longer time scale. We can also see in the figure and by using theorem 3.7 that a single voice source is a renewal process. But the superpositioned process is far from a renewal process.

5.3 The MMPP approximation

Seen from the IDI perspective the difference between the two types of sources are very small both as a single source and superpositioned sources. The cumulative covariance is still present in the approximation. By approximating the arrival process with a Markov modulated Poisson process do we loose very little in the description of the arrival process and provide mathematical tractability.

6 Loss analysis of the multiplexer queue

So far in this thesis have we only been talking about the arrival process. Now we feed the arrivals into a queueing system. We introduce a server and a buffer of finite size to store incoming packets before they are served by the server.

6.1 Construction of the multiplexer

The multiplexer consists of a buffer of finite size fed by arriving packets from independent voice sources of the type described in section 4.1.1. Packets arriving at a full buffer are considered as lost. Packets in the queue are served by a single server in a FIFO(first in first out) manner. Since packets are of equal size the service time is deterministic and we denote it S_0. The core architecture of the multiplexer is shown in figure 13.

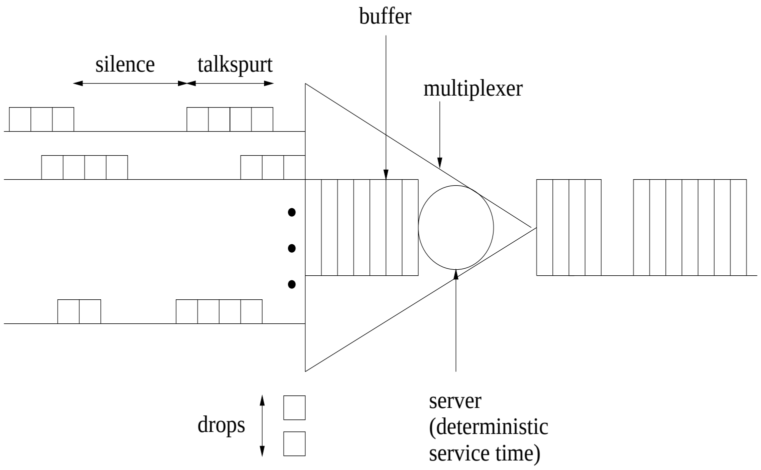

Figure 13: Core architecture of the multiplexer

The number $N$of multiplexed voice sources is considered as fixed. We need to introduce a number of dimensioning parameters for the multiplexer. We denote the *output link capacity* as C. The *peak rate* is defined as

$$h = \frac{B}{T},$$

where we have used B to denote packet size. Assuming peak rate assignment, is the value

$$M = \lfloor C/h \rfloor \tag{17}$$

the maximum number of sources that can be accommodated in the MUX. Furthermore, we assumed that the following stability condition is satisfied

$$\rho = \frac{Nhp_{on}}{C} < 1, \tag{18}$$

where ρ is the traffic intensity(or load factor). And to insure that we are in the statistical multiplexing region, we assume that $Nh > C$.

6.2 Calculating the loss probability of the MMPP/D/1/K queue

From now we assume that the arrival process is described by a Markov modulated Poisson process. In the following sections a model is developed which estimates the loss probability, i.e the probability of losing a packet due to buffer overflow. The description of the model is rather brief, for details and further reference check [2] and [3].

6.2.1 Notation

In this section is the notation of the variables used in the in the calculations presented. The notation here principally follows the notation in [2]. The scalar variables are denoted:

$\Pi(K)$ The loss probability of the MMPP/D/1/K queue.

ρ The mean offered load.

$H(t)$ The distribution function of the service time.

$\tilde{H}(a)$ The *Laplace Stieltjes Transform*(LST) of $H(t)$.

θ The mean service time.

L The packet emission rate of single source while in talkspurt.

The following variables are $(N+1) \times (N+1)$ matrices.

R The intensity matrix of the phase process.

$\mathbf{\Lambda}$ The diagonal matrix, whose element Λ_{ii} is equal to the mean arrival rate while in phase i.

U The matrix given by $(\mathbf{\Lambda} - \mathbf{R})^{-1}\mathbf{\Lambda}$ which accounts for the evolution of the phase process during server's idle periods.

Finally, we introduce some $(N+1)$-dimensional vector variables, denoted as follows:

e The unit column vector.

$\pi_{\mathbf{K}}(\mathbf{i})$ The row vector whose j-th element is the limiting probability at the embedded time instants of having i users in the system and being in the phase j of the MMPP, $i = 0, 1, ..., K$.

q The row vector containing the limiting state probabilities of the phase process. It can be obtained as the unique solution of the system $\mathbf{qR} = 0$ and $\mathbf{qe} = 1$.

6.2.2 Loss analysis of the MMPP/D/1/K queue

According to [2] are we able to calculate the loss probability using the following formula:

$$\Pi(K) = 1 - \frac{1}{\rho[1 + \pi_{\mathbf{K}}(0)\mathbf{U}\mathbf{\Lambda}^{-1}\theta^{-1}\mathbf{e}]}. \tag{19}$$

Given the arrival process characteristics through $\mathbf{R}$ and $\mathbf{\Lambda}$ we calculate the loss probability as follows. Let $x_i(z)$ denote the eigenvalues of the matrix $\mathbf{S}(z) = \mathbf{R} + (z-1)\mathbf{\Lambda}$ and denote the numbers $\eta_i,\ i = 1, ..., N$ and $\zeta_i,\ i = 0, 1, ..., N$ the roots of the equation

$$z = \tilde{H}[-x_i(z)],\ i = 0, 1, ..., N. \tag{20}$$

where η_i are the real roots inside the unit circle and ζ_i are the real roots outside the unit circle. In particular, let ζ_0 denote the least root greater than one. The vector $\pi_{\mathbf{K}}(0)$ can be calculated from the equation

$$\pi_{\mathbf{K}}(0)\mathbf{D}_{\mathbf{K}} = \mathbf{q}, \tag{21}$$

where $\mathbf{D}_{\mathbf{K}}$ is the matrix

$$\mathbf{D}_{\mathbf{K}} = \mathbf{Q}(1) + \mathbf{Q}(\zeta_0)\zeta_0^{-K} + \sum_{i=1}^{N}[\mathbf{Q}(\eta_i)\eta_i^{-K} + \mathbf{Q}(\zeta_i)\zeta_i^{-K}]. \tag{22}$$

The expressions for the matrices $\mathbf{Q}(\cdot)$ are given in appendix A.
Let us now outline an algorithm to be carried out in order to calculate the loss probability. Given the matrices $\mathbf{R}$ and $\mathbf{\Lambda}$ and the distribution of the service time $H(t)$, we need to:

1. Compute all the eigenvalues $x_i(z)$ of the matrix $\mathbf{S}(z) = \mathbf{R} + (z-1)\mathbf{\Lambda}$.

2. Find ζ_0, (i.e. the unique real root of equation 20) and ζ_i and η_i, (i.e. the unique and real roots of equation 20 such that $\eta_i < \zeta_i$,) for $i = 1, ..., N$.

3. Compute the matrices $\mathbf{Q}$ of equation 22 according to the expressions given in appendix A. This in turn requires the computation of the right and left eigenvectors of $\mathbf{S}(\zeta_0), \mathbf{S}(\zeta_i)$ and $\mathbf{S}(\eta_i)$, for $i = 1, ..., N$.

4. Compute the matrix $\mathbf{U} = (\mathbf{\Lambda} - \mathbf{R})^{-1}\mathbf{\Lambda}$.

5. Compute the matrix $\mathbf{D}_{\mathbf{K}}$ according to equation 22, and then $\pi_{\mathbf{K}}(0)$ from equation 21.

6. Compute the loss probability from equation 19.

The computational burden depends mostly on the first three steps of the algorithm, apart from the matrix inversion of $(N+1) \times (N+1)$ matrices in the steps 5 and 6. In the next section we will look at ways of simplifying the calculations.

6.3 Asymptotic matching

When N sources are superposed we need to know the explicit expressions of the eigenvalues of the matrix $\mathbf{S}(z)$ which generally are hard to find. However if the number of states is reduced to two are we able to find expressions for these eigenvalues. It is possible to reduce the number of states by observing that when the number of sources in a talkspurt exceeds M, as defined in equation 17, then the queue builds up because more packets arrive to the multiplexer than the link is able to send. We refer to the states $\{M+1, \cdots, N\}$ as *overload*(OL) states and the remaining $\{0, \cdots, M\}$ states as *underload*(UL) states. Figure 14 illustrates the division of states.

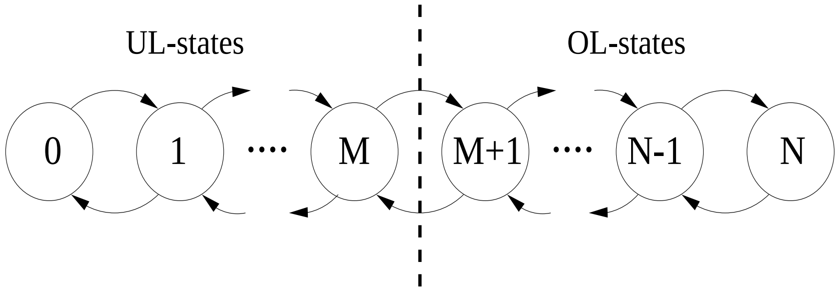

Figure 14: Division of states into OL and UL states

If we group the the UL and OL states together we have two states and we use these states to form a new two state MMPP approximating the original arrival process, one state is called OL state corresponding to the overload states and the other state is called UL state corresponding to the underload states. We need to determine expressions for the four parameters characterizing the MMPP, i.e.

$r_{OL}(r_{UL})$ The mean transition rate out of the OL (UL) state.

$\lambda_{OL}(\lambda_{UL})$ The arrival rate of the Poisson process in the OL (UL) state.

Figure 15 shows the model graph of the two state MMPP. The approximation with

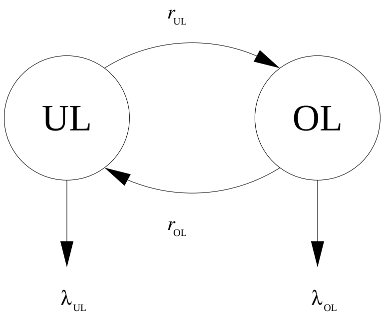

Figure 15: Superposition of Poisson processes

a two state MMPP together with the calculations of the four parameters is called *asymptotic matching* and is briefly presented in the next section. A more detailed presentation can be found in [3].

6.3.1 Calculating the parameters

Let us first consider the OL state. We denote the random variable ν the duration in the overload states in the phase process $J(t)$. Further, let ν^* be a random variable representing the amount of time spent during a visit in the OL state in the two-state MMPP. Now we can find a value of r_{OL} by approximating the distribution of ν with the distribution of ν^*, which is exponential with mean $\frac{1}{r_{OL}}$.

Before we can calculate the value of r_{OL} we have to introduce the concept of absorption. A state in a Markov process is called absorbing if the process is not able to leave the state once it has entered it. This means that the row corresponding to the absorbing state in the intensity matrix defined in section 3.2.3 consists entirely of zeros.

We are now ready to turn to the calculation of r_{OL}. The random variable ν can be identified with the time until absorption in a Markov process $\mathcal{M}$, obtained from the phase process considering only the states $\{M, M+1, \cdots, N\}$ and making the state M absorbing. The model graph of $\mathcal{M}$ is shown in figure 16. According to

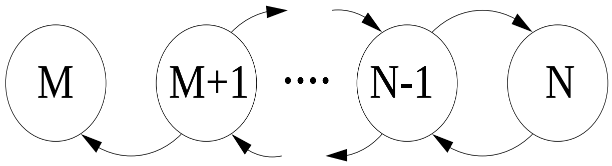

Figure 16: Model graph of $\mathcal{M}$, state M absorbing.

[3] are we able to calculate r_{OL} by calculating the maximal real-part eigenvalue of the matrix $\mathbf{Q_M}$, where $\mathbf{Q_M}$ is the $(N-M)\times(N-M)$ intensity matrix relevant to the states $\{M+1, \cdots, N\}$. The eigenvalue is real and negative. It is denoted by $-\eta,\ \eta > 0$. Now we approximate r_{OL} with η, i.e. we set $r_{OL} = \eta$. The problem of determining the value of r_{OL} is thus reduced to classical numerical analysis problem, the calculation of the maximal real part eigenvalue of $\mathbf{Q_M}$. As for $\lambda_{OL}(\lambda_{UL})$, we can require that it be equal to the mean emission rate in the OL(UL) region of the phase process. Then we obtain:

$$\lambda_{OL} = L \sum_{i=M+1}^{N} i \frac{\pi_i}{\pi_{OL}} \tag{23}$$

$$\lambda_{UL} = L \sum_{i=0}^{M} i \frac{\pi_i}{\pi_{UL}} \tag{24}$$

with $\pi_{OL} = \sum_{i=M+1}^{N} \pi_i$ and $\pi_{UL} = \sum_{i=0}^{M} \pi_i$ and π_i as defined in equation 12. When the values of r_{OL}, λ_{OL} and λ_{UL} have been set, it is necessary to choose the value of r_{UL} properly in order to obtain an overall mean arrival rate λ equal to the mean emission rate of the phase process, i.e. MLp_{on}. On the other hand, we have

$$\lambda = \frac{r_{UL}\lambda_{OL} + r_{OL}\lambda_{UL}}{r_{OL} + r_{UL}}$$

Using the prescription that $\lambda = NLp_{on}$ we obtain:

$$r_{UL} = r_{OL}\frac{NLp_{on} - \lambda_{UL}}{\lambda_{OL} - NLp_{on}} \tag{25}$$

Note that r_{UL} is always positive, since $\lambda_{UL} < Np_{on}L < \lambda_{OL}$.The asymptotic matching can be summarized in the following steps:

1. Compute η and set $r_{OL} = \eta$.
2. Compute λ_{OL} and λ_{UL} using equations 23 and 24.
3. Compute r_{UL} trough equation 25.

The computational burden lies entirely in the computation of the maximal real eigenvalue in the first step.

6.3.2 Computing the LST of the service time distribution

As mentioned earlier in section 6.1 is the service time deterministic, and denoted S_0. Figure 17 shows the graph of the service distribution function.

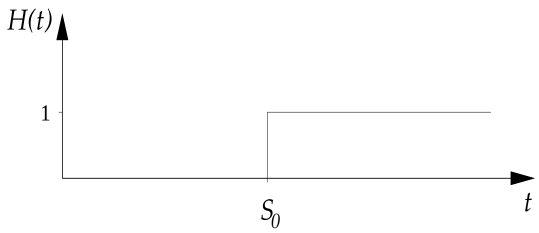

Figure 17: Service time distribution function.

The formal definition of the Laplace-Stieltjes transform is as follows:

$$\tilde{H}(a) = \int_0^\infty e^{-at} dH(t). \tag{26}$$

The fact that the service time is deterministic, i.e. $\theta = S_0$ simplifies the calculations considerably, in fact

$$\tilde{H}(a) = E(e^{-aS_0}) = e^{-aS_0}. \tag{27}$$

6.3.3 Calculating the eigenvalues of the matrix $\mathbf{S}(z)$

Since we have approximated the phase process $J(t)$ with a two state MMPP are the matrices $\mathbf{R}$ and $\boldsymbol{\Lambda}$ 2×2 matrices. Hence the matrices are:

$$\mathbf{R} = \begin{pmatrix} -r_{UL} & r_{UL} \\ r_{OL} & -r_{OL} \end{pmatrix} \quad \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{UL} & 0 \\ 0 & \lambda_{OL} \end{pmatrix}$$

and thus the matrix $\mathbf{S}(z)$ is

$$\mathbf{S}(z) = \begin{pmatrix} (z-1)\lambda_{UL} - r_{UL} & r_{UL} \\ r_{OL} & (z-1)\lambda_{OL} - r_{OL} \end{pmatrix}$$

The calculations of the eigenvalues is now a straight forward task using the characteristic polynomial of $\mathbf{S}(z)$. For convenience we introduce the functions

$$\begin{aligned} f_1(z) &= (z-1)\lambda_{UL} - r_{UL} \\ f_2(z) &= (z-1)\lambda_{OL} - r_{OL} \end{aligned}$$

Using the functions above will simplify the expressions for the eigenvalues. The matrix $\mathbf{S}(z)$ has two eigenvalues denoted $x_1(z)$ and $x_2(z)$, and their values are

$$x_1(z) = \frac{f_1(z) + f_2(z)}{2} - \frac{1}{2}\sqrt{(f_1(z) - f_2(z))^2 + 4r_{OL}r_{UL}} \tag{28}$$

$$x_2(z) = \frac{f_1(z) + f_2(z)}{2} + \frac{1}{2}\sqrt{(f_1(z) - f_2(z))^2 + 4r_{OL}r_{UL}} \tag{29}$$

First we note that $z = 1$ is a root of $x_2(z)$. Now we are ready to calculate the values of η_0, ζ_0 and ζ_1 as the roots of the equation $z = \tilde{H}[-x_i(z)], i = 1, 2$. Explicitly we want to solve the equation 22

$$z = e^{-x_i(z)S_0}, \ i = 1, 2. \tag{30}$$

These roots must be found numerically and we use them in equation 22 to calculate the the matrix $\mathbf{D_K}$. At this point we have all the information we need in order to calculate the loss probability according to equation 19 by carrying out the three last steps of the algorithm in section 6.2.2. Since the matrices involved in the calculations all are 2×2 matrix inversion is elementary.

7 Numerical considerations

In this section we present calculations made in the model and compare results with results obtained from simulations made with a network simulation package.

7.1 The network simulator ns.

To estimate the validity of the model we have used the network simulation package Network Simulator (ns) developed at Lawrence Berkeley National Laboratory, Berkeley. The simulator is written in C++. Commands are given to ns with the Tcl (Tool command language) language. Tcl is a general-purpose scripting language for command giving to applications. For more information on ns or to download code see http://www-mash.cs.berkeley.edu/ns/ns.html.

7.2 Numerical comparisons

We have simulated a multiplexer with an output link capacity of 1.536Mb/s (Megabit per second) and buffer sizes ranging from 2 to 100 packets. With our choice of parameters we introduce a maximum delay of 33 ms in the buffer. According to [10] is a delay 0-150 ms acceptable for telephony. Delay between 150 and 400 ms can also be acceptable, but delay over 400 ms is unacceptable. We feel that a delay of 33 ms is about what we can accept in one hop. The number of sources multiplexed varies from 50 to 130 in steps of 20. The results from the model and simulations are accounted for in tables in Appendix B.

In figures 18 and 19 we see how the calculated values from the model agree with the simulations quite well over a wide range of buffer sizes. Figure 20 shows the loss probability in linear scale and we see that when the buffer is larger than 10 packets we hardly have any packet losses and the model and simulations agree very well. However figure 21 shows the loss probability in logarithmic scale and now the picture is totally different. The model and simulations seem not to agree any more. This deviation can be understood by looking at the size of the loss probabilities predicted by the model. For a buffer size of 100 packets and 90 sources the loss probability is of the size of 10^{-8}, an extremely rare event in other words. We must simulate the multiplexer for a very long time in order to obtain a reasonable measure of the loss probability. For the simulation of 130 sources we found that it was sufficient to simulate the multiplexer for 2000 seconds, when we simulated with 110 sources we had to run the simulation for 50 000 seconds to stabilize the loss probability. And for 90 sources and less we have simulated the multiplexer for 100 000 seconds and we are still not even close of obtaining reasonable data for large buffer sizes.

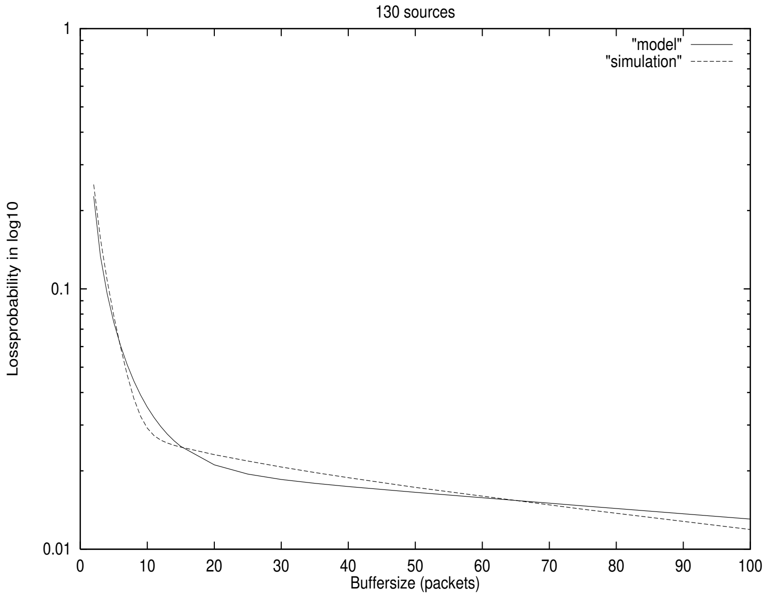

Figure 18: Loss probability in log 10 Vs. buffer size.

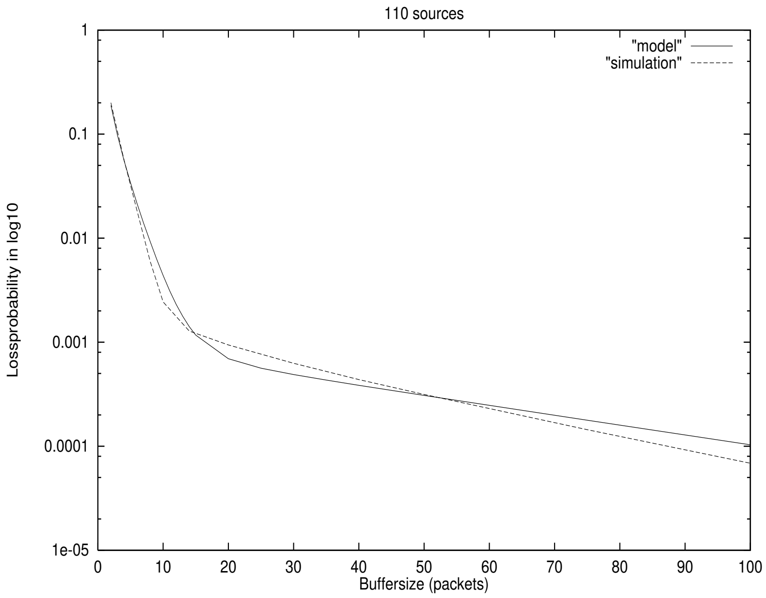

Figure 19: Loss probability in log 10 Vs. buffer size.

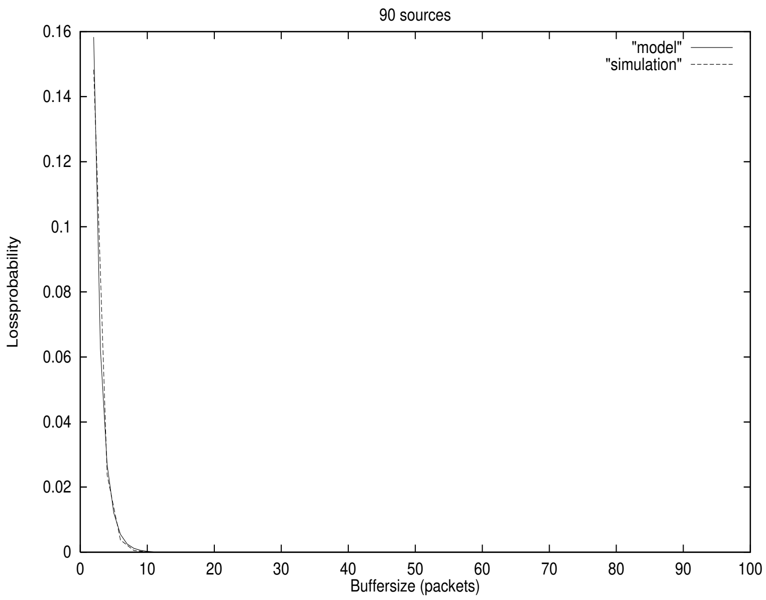

Figure 20: Loss probability Vs. buffer size.

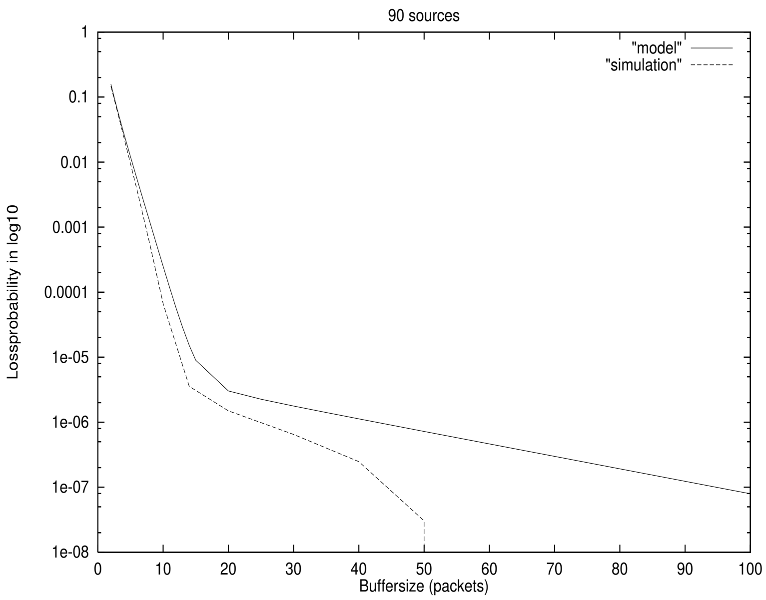

Figure 21: Loss probability in log 10 Vs. buffer size.

8 Buffer occupancy distribution

In this section we take a look at the variance of the delay introduced by the queue and how it varies with load. Data for the plots in this section come from simulations made with ns.

8.1 Queue length distribution

Figure 22 shows the buffer occupancy distribution when $K = 100$, i.e. the probability of having i packets in the buffer($i = 0, 1, 2, \cdots, K$), when 120 sources are multiplexed($\rho = 0.88$). We see that most of the mass is concentrated to the lower regions of the buffer, there are rarely more than 10 packets in the buffer. However when the number of sources is increased to $135(\rho = 0.99)$ something dramatic happens as figure 23 indicates. The buffer occupancy distribution shows a bi-modal behaviour, it is either almost full or almost empty. This means that the variance introduced by the queueing in the buffer increases dramatically when the load factor approaches one.

8.2 Identifying the relevant time scale

To understand this behaviour we use the fact that the limiting state probabilities for the phase process $J(t)$ is binomially distributed. To get a hint about how often the multiplexer is heavily overloaded we approximate the binomial distribution with the normal distribution. This is allowed since $Np_{on}(1 - p_{on}) = 30.75 > 10$ (see any text-book in basic statistics). Calculations show that with probability 0.01 are approximately 60 or more sources in a talkspurt at the same time. Thus, if 60 sources are in a talkspurt we have 12 sources more than the multiplexer can handle. With our choice of parameters it takes on average 128 ms for the buffer to go from empty to full. If we take into account that 60 sources are ON, about 480 packets arrive during this period. From Figure 12 we see that the arrival process deviates substantially from the Poisson process, which means that the queue builds up faster than it would if the traffic was smooth. And if we consider that the average ON-period lasts for 252 ms, about 1320 packets arrive during this period and on this time scale the arrival process is much burstier than the Poisson process. It is quite rare that we have this much overload but as ρ approaches one it happens from time to time, and when it does the queue builds up quickly and we must look at the behaviour of the arrival process over many more arrivals.

Note that when the load i moderate or low it is rare that the multiplexer is heavily overloaded and in the infinite buffer case the M/D/1 queue provides a good description of the queueing behaviour. In the finite buffer case the M/D/1/K queue approximates the queueing behaviour with good precision, however it fails to capture the cumulative correlations that play an important role for larger buffer spaces and the M/D/1/K queue will underestimate the loss probabilities, but these loss probabilities are so small that they are not interesting for the IP-telephony application.

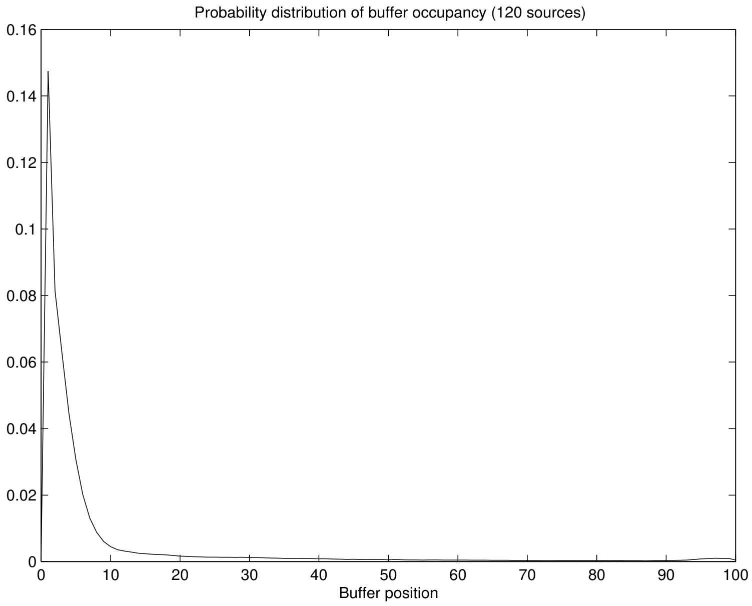

Figure 22: Buffer occupancy distribution, 120 sources.

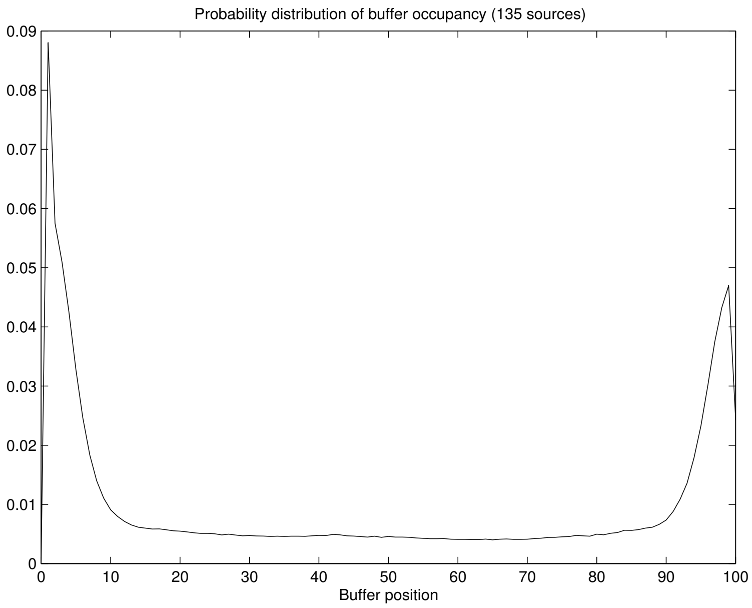

Figure 23: Buffer occupancy distribution, 135 sources.

9 Conclusions and discussion

We have seen how the arrival process can be approximated by an Markov modulated Poisson process in order to calculate the loss probability. We have also seen that when the load is low or moderate the arrival process can be well approximated by a Poisson process and the much more tractable M/D/1/K and M/G/1 queueing systems can be used in the finite and infinite buffer case respectively. When the load is heavy we must look at a longer time scale to understand the congestion that occurs in the buffer. We have also seen that we can connect at least twice as many telephone calls with statistical multiplexing than with peak rate allocation without degrading the quality of the phone call to an unacceptable extent. Note that the approximation of the number of sources that are in a talkspurt with a Markov process is only valid if we assume that the silence and talkspurt periods are exponentially distributed. According to [9] is the match for silence periods reasonable but not perfect, so if another distribution is used for the silence periods the approximation of the arrival process with an Markov modulated Poisson process fail and we must resort to some other model.

Appendix A

In this appendix we give the expressions for the $\mathbf{Q}(\cdot)$ matrices in equation 22. This presentation will be brief, a more complete presentation can be found in [2]. Let $\mathbf{S}(z) = \mathbf{R} + (z-1)\mathbf{\Lambda}$ and let $x_i(z), \mathbf{u}_i(z)$ and $\mathbf{v}_i(z)$ denote the eigenvalues, the right eigenvectors and left eigenvectors of $\mathbf{S}(z)$, respectively $(i + 0, 1, 2, \cdots, N)$. We start with the expression for $\mathbf{Q}(1)$, it reads

$$\mathbf{Q}(1) = \frac{\rho}{1-\rho}\mathbf{U}\mathbf{\Lambda}^{-1}\theta^{-1}\mathbf{eq}. \tag{31}$$

The expressions for the matrices $\mathbf{Q}(\eta_i)$ and $\mathbf{Q}(\zeta_i)$ reads

$$\mathbf{Q}(\eta_i) = \frac{\mathbf{U}\mathbf{\Lambda}\mathbf{u}_i(\eta_i)\mathbf{v}_i(\eta_i)x_i(\eta_i)}{(1-\eta_i)[x_i'(\eta_i)|\tilde{H}'[-x_i(\eta_i)]| - 1]} \tag{32}$$

for $i = 1, \cdots, N$, and

$$\mathbf{Q}(\zeta_i) = \frac{\mathbf{U}\mathbf{\Lambda}\mathbf{u}_i(\zeta_i)\mathbf{v}_i(\zeta_i)x_i(\zeta_i)}{(1-\zeta_i)[x_i'(\zeta_i)|\tilde{H}'[-x_i(\zeta_i)]| - 1]} \tag{33}$$

for $i = 0, 1, \cdots, N$. Since $x_i'(z) = \mathbf{v}_i(z)\mathbf{\Lambda}\mathbf{u}_i(z)$, $i = 0, 1, \cdots, N$. Therefore we need not compute the derivatives of the eigenvalues at η_i and ζ_i explicitly, since knowledge of the eigenvectors is sufficient. Moreover, in case of deterministic service times, $|\tilde{H}[-x_i(\zeta_i)]|$ and $|\tilde{H}[-x_i(\eta_i)]|$ reduce to $\theta\zeta_i$ and $\theta\eta_i$, respectively.

Appendix B: Tables

Buffer size	50 sources		70 sources		90 sources	
	Simulation	Model	Simulation	Model	Simulation	Model
2	0.05400	0.064463	0.09841	0.10897	0.14828	0.15830
4	0.0014412	0.0018736	0.0073770	0.0088109	0.02374	0.027002
6	2.8722e-05	5.6375e-05	0.00046758	0.00078062	0.0037461	0.0055035
8	2.8212e-07	1.69798e-06	2.4498e-05	6.9695e-05	0.00052964	0.0011623
10	2.7302e-08	5.1144e-08	8.5171e-07	6.2270e-06	6.9033e-05	0.00025028
14	0	4.6399e-11	0	4.9802e-08	4.5336e-06	1.5190e-05
20	0	1.1102e-15	0	9.0026e-11	2.2584e-06	3.0229e-06
30	0	0	0	2.5013e-11	1.0280e-06	1.7742e-06
40	0	0	0	1.1712e-11	3.3707e-07	1.1248e-06
50	0	0	0	5.5044e-12	3.3707e-08	7.2084e-07
75	0	0	0	8.3488e-13	0	2.3896e-07
100	0	0	0	1.2656e-13	0	7.9326e-08

Table 2: Loss probability for 50, 70 and 90 sources.

Buffer size	110 sources		130 sources	
	Simulation	Model	Simulation	Model
2	0.20030	0.18875	0.25165	0.22675
4	0.056657	0.055941	0.10806	0.095944
6	0.018410	0.021603	0.060010	0.060740
8	0.0060095	0.0092994	0.037914	0.044270
10	0.0024403	0.0043814	0.029160	0.035138
14	0.0012803	0.0014123	0.025069	0.026108
20	0.000941	0.00069344	0.02308	0.021088
30	0.00062677	0.00048879	0.020681	0.018529
40	0.00043750	0.00038489	0.01881	0.017404
50	0.00031453	0.00030767	0.017269	0.016537
75	0.00014417	0.00017803	0.014233	0.014665
100	6.8689e-05	0.00010328	0.011898	0.013055

Table 3: Loss probability for 110 and 130 sources.

Acknowledgments

I would like to take the opportunity to thank my examiner Ingemar Kaj at the department of Mathematics, Uppsala University for his guidance and advice during my work with this thesis. Also, I would like to thank my advisor Ian Marsh at Swedish Institute of Computer Science for his help with ns and valuable comments. I am also grateful to Bengt Ahlgren at Swedish Institute of Computer Science for his suggestions and for offering me the opportunity to write my thesis at the Swedish Institute of Computer Science.

References

[1] S. Anick, D. Mitra, M.M. Sondhi, "Stochastic Theory of a Data-Handling System with Multiple Sources" *Bell System Tech. J.*, Vol. 61, NO 8, pp 1871-1894, October 1982.

[2] Andrea Baiocchi, Nichola Blefari Melazzi, Aldo Roveri, "Buffer dimensioning criteria for an ATM multiplexer loaded with homogeneous ON-OFF sources" *ITC 13*, Copenhagen, Denmark, June 1991.

[3] Andrea Baiocchi, Nichola Blefari Melazzi, Marco Listanti, Aldo Roveri, Roberto Winkler, "Loss Performance Analysis of an ATM Multiplexer Loaded with High-Speed ON-OFF Sources" *IEEE J. Selec. Areas Commun.* Vol 9. NO. 3, pp. 388-393, April 1991.

[4] A Baiocchi, N Blefari Melazzi, A Roveri, F Salvatore, "Stochastic Fluid Analysis of an ATM Multiplexer Loaded with Heterogeneous ON-OFF Sources: an Effective Computational Approach" *INFOCOM 92*, Florence, Italy, May 4-8 1992.

[5] Harry Heffes, David M. Lucatoni, "A Markov Modulated Characterization of Packetized Voice and Data Traffic and related Statistical Multiplexer Performance" *IEEE J. Selec. Areas Commun.* Vol SAC-4 NO. 6, pp 856-868, September 1986.

[6] Ramesh Nagarajan, James F. Kurose, Don Towsley, "Approximation Techniques for Computing Packet Loss in Finite- Buffered Voice Multiplexers" *IEEE J. Selec. Areas Commun.* Vol 9. NO 3, pp 368-377, April 1991.

[7] Roger C. F. Tucker, "Accurate Method for Analysis of a Packet-Speech Multiplexer with Limited Delay" *IEEE Transactions on Communications* Vol 36. NO 4, pp 479-483, April 1988.

[8] Zheng Sun "*Capacity Study of Statistical Multiplexing for IP Telephony*" Technical report. Department of Mathematics, Linköping University, Sweden. LiTH-MAT-EX-98-12.

[9] Kotikalapudi Sriram, Ward Whitt, "Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data" *IEEE J. Selec. Areas Commun.* Vol. SAC-4, NO. 6, September 1986.

[10] Recommendation G.114, "*General recommendations on the transmission quality for an entire international telephone connection*", 1993.

[11] Allan Gut, "*An Intermediate Course in Probability.*", Springer-Verlag New York 1995.

[12] Howard M. Taylor, Samuel Karlin, "*An Introduction To Stochastic Modeling 3rd edition*", Academic Press London 1998.

[13] Sidney I. Resnick "*Adventures in Stochastic Processes*", Birkhäuser Boston 1992.