Survey of Source Modeling Techniques for ATM Networks

Sponsor: Sprint

Yong-Qing Lu
David W. Petr
Victor S. Frost

Technical Report TISL-10230-1

Telecommunications and Information Sciences Laboratory
Department of Electrical Engineering and Computer Science
University of Kansas

September 1993

Abstract

Analytical modeling of traffic sources in a packet-switched communications network is the basis of performance analysis and network design. Along with the development of high speed networks, more types of traffic are being applied to the network, such as real time video traffic and multimedia traffic with voice, data, image and video. To support quality service for various types of traffic, it is very important to characterize the properties of the traffic through accurate modeling. For decades, traffic modeling has been an active research area.

This paper is a review of the traffic studies that have been presented in the literature as well as an introduction to the recent studies that either reveal new traffic phenomenon that has not been seen from the earlier models or imply new directions in traffic modeling. The relationships between different models are also explored.

1. Introduction

The objective of traffic modeling is to approximate the arrival process of either single source or multiplexed sources, so that the traffic models can be used for network performance analysis. Networks are in turn designed based on the correct knowledge provided by the performance analysis. Only with a thorough understanding of the traffic and with appropriate models, is the study of network performance and effective design of networks possible. Many different approaches to traffic modeling have been proposed in the literature[1][7][9][11][16][19][24]. Since the majority of real traffic possesses correlated interarrival sequences which have a major effect on queue performance, the focus has been to build up models which best capture the correlation characteristics of different traffic.

General traffic modeling involves the following procedure:

- Set up a mathematical stochastic process as an approximation model for the target traffic based on its properties observed from the empirical data.
- Determine the parameters of the model based on the traffic properties. This step usually involves many analytical techniques. How to choose the parameters is also critical to the accuracy of the model.
- Use the model to analyze the queue performance (queue length, queue delay, loss probability etc.) analytically or by simulation.
- If possible test the model by comparing the analytical solution to the simulation result and find out the limitation of the model.

This paper summarizes common traffic models and introduces several recent studies that present new approaches in traffic modeling.

In section 2, we introduce the earliest renewal models that do not account for the correlation of the traffic arrivals. Then, we look at the popular Markovian process models which include Markov chains and Markov modulated Poisson process models in section 3. Section 4 is a description of the fluid flow models which are attractive for high speed asynchronous transfer mode (ATM) networks. In section 5, we present the autoregressive type models and transfer-expand-sample (TES) method that were developed for capturing variable bit rate video traffic possessing strong correlation properties. In section 6, we introduce the self-similar phenomena[16][22] which have been observed both in Ethernet traffic and in variable bit rate (VBR) video sources. This observation provides a totally different view of the traffic characteristics which is not seen in the models presented in the previous literature. Two other methods, chaotic map and frequency domain analysis methods, are described in section 7 and 8 respectively. These two methods have the potential to offer a unified model or measurement for all kinds of traffic sources.

2. Renewal Process Models

A renewal process is defined as a arrival process in which the individual arrivals or the interarrival times are independent from each other. The simplest model for traffic is a renewal process model which does not take the correlation of the traffic arrivals into account. It was used early as an approximation of the voice source.

2.1 Renewal Process Model of Single Voice Source

Event described by the model: packet arrivals from a voice source.

Voice is one of the most basic traffic sources for all kinds of networks. The voice packet stream of a single voice source is characterized by arrivals at fixed intervals of T ms during talkspurts and no arrivals during silences as shown in Fig. 2. 1[2]. The talkspurt has a random length NT, where N is the number of packets arrived in this period, while the silence period has a random length X. If we look at the talkspurt as a single arrival event with a random silence period X, it is reasonable to assume that this arrival process is a renewal process. Similarly, if we look at the silence period as an arrival event with a random talkspurt NT, it is also a renewal process. Furthermore, the number N of packets in a talkspurt can be assumed to be geometrically distributed, i.e., $P(n = N) = p^N(1 - p)$, where a packet arrives with probability p and does not arrive with probability $1 - p$. This assumption is consistent with the measurements indicating that talkspurts are exponentially distributed.

The talkspurt and silence periods have approximately exponential distributions with mean period length α^{-1} and β^{-1} respectively, where $\alpha^{-1} = \bar{N}T = \sum_{N=1}^{\infty} Np^N(1 - p) = pT/(1 - p)$. The mean packet arrival rate from a single source is then $1/(T + \alpha T/\beta)$[1]. A renewal model that approximates this process can be given by an interarrival time distribution of the packet stream as follows:

$$F(t) = [(1 - \alpha T) + \alpha T(1 - e^{-\beta(t-T)})]U(t - T)^{[1]},$$

where $U(t)$ is the unit step function.

Fig.2.1 Speech Process[2]

2.2 Renewal Model of Superposition of Voice Sources

Event: packet arrivals from aggregated independent voice sources

The superposition of independent packet voice sources as described above is usually not a renewal process. However, because of its mathematical simplicity for numerical analysis, one of the earliest works approximates the superposed voice arrival process as a renewal process[2]. In order to capture the dependence among successive interarrival times in the aggregate packet arrival process, , two parameters of the model are chosen to characterize the arrival process: the average arrival rate and the index of dispersion for intervals (IDI). The IDI is defined as the squared coefficient of variation of the interarrival time. It is used to measure the correlations of the interarrival times. This parameter is again used later as a measure of the source burstiness. Although this model is not an accurate one, Sriram and Whitt[2] proved that the superposition of the voice sources is definitely not a Poisson process.

3. Markovian Process Models

3.1 Discrete Time Two-State Markov Chain Model For Single On-Off Traffic

Event: packet arrivals.
Traffic: on-off type traffic.

A more commonly used model for a single on-off source, which includes voice, video, and possibly image, is the Markov chain. This model has an obvious advantage over the renewal model in characterizing the arrival correlation.

Definition of on-off source[8][9]: source that alternates between active emission periods (on period) and idle periods (idle period). During the on period information is generated at a constant peak access rate, while in the off period no information is emitted (see Figure 3.1). A voice source is a typical on-off source. Properties of on-off source are characterized by

- peak access rate, which is the source access rate during its active period;
- average access rate, which is the average rate of the source in steady state.
- rate correlation, which is the sum over all lags of the correlations of packet arrivals (see equation for S in section 3.1.1). For example, a rate correlation value of one means no correlation of arrivals. On the other hand, rate correlation is infinite for a deterministic arrival stream.

The ratio of peak access rate to average rate is then defined as the burstiness of a source.

Because of its modeling simplicity with a clear definition of correlation and burstiness and the simplicity involved in queuing analysis, the discrete time two-state Markov chain is very successful in modeling diverse individual on-off traffic sources. The two-state Markov chain is characterized by three parameters: the transition probability from the off to the on state 1-*q*; the transition probability from the on to the off state 1-*p*; the packet emission rate in the on state. Figure 3.1 shows the two-state Markov chain. Unlike the renewal voice source model, the on-off source is no more a renewal process. There are correlations between packet interarrival times. The following are two examples.

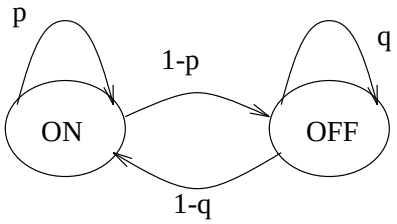

Fig 3.1 Two-state Markov chain

3.1.1 Discrete Time Two-State Markov Chain Model For Voice Source

A discrete-time two-state Markov chain is used to model the on-off type talkspurt and silence periods in a speech process as shown in Figure 3.1. This model is a more accurate model for a voice source than the renewal process model. The transitions of the two-state Markov chain are defined at time instants equally separated by τ, which equals the fixed packetization interval of the voice source. The holding time on each state is geometrically distributed in τ intervals. While in talkspurt, voice packets are periodically generated at fixed τ intervals. τ^{-1} represents the peak access rate (packets/sec). If the average on and off periods are defined by T_{on} and T_{off}, then the average voice activity factor is[9]

$$\bar{\varepsilon} = \frac{T_{on}}{T_{on} + T_{off}} = \frac{1-q}{2-q-p}.$$

The average access rate of the voice source is given by $\bar{\varepsilon}\tau^{-1}$. Its burstiness is then $1/\bar{\varepsilon}$. The normalized autocorrelation function of the packet arrivals in the n^{th} adjacent τ intervals has a geometric form

$$R(\tau n) = \phi_\tau^{|n|}.$$

where $\phi_\tau = q + p - 1$ is the correlation coefficient in time unit τ. The rate correlation can then be defined by

$$S = \sum_{n=0}^{+\infty} R(\tau n) = \frac{1}{1-\phi_\tau}$$

which is the autocorrelation accumulated at all positive τ intervals. The larger the S, the slower the traffic variation.

3.1.2 Multiple Two-State Markov Chain For VBR Video Source

A variable bit rate video source can be modeled as a number, say K, of independently identically distributed (i.i.d.) two-state Markov chains. If the τ and $\bar{\varepsilon}$ are defined same as above, the source access rate is equal to $j\tau^{-1}$ if there are j two-state Markov chains presently in the on state. Peak access rate and average access rate are $K\tau^{-1}$ and $K\bar{\varepsilon}\tau^{-1}$ respectively. The burstiness of the source is therefore equal to $\bar{\varepsilon}^{-1}$, which remains unchanged from that of a single two-state Markov chain. The steady state distribution of the source access rate is binomial with the probability of access rate equal to $j\tau^{-1}$ given by $\binom{K}{j}\bar{\varepsilon}^{j}(1-\bar{\varepsilon})^{K-j}$. Similarly, the rate correlation for the video source is characterized by the same value of S given above.

3.2 Two-State Markov Modulated Poisson Process Model (MMPP) For Superposed Renewal Process

Event: packet arrivals.
Traffic: aggregated voice and aggregated data.

Performance Study: delay vs load.

Besides the renewal model, another more popular approach of the modeling of aggregate packet voice is the two-state Markov modulated Poisson Process[1]. The single voice source used in this case is the renewal model described previously. Heffes and Lucantoni approximate the superposition process as a correlated nonrenewal stream, which was chosen to be the two-state Markov modulated Poisson process for its analytical simplicity and versatility. This model captures the correlation of interarrival times better than the previous renewal model.

The two state Markov modulated Poisson process is a doubly stochastic Poisson process. On the macro level, the superposition arrival process is viewed as a two-state Markov chain with the mean sojourn times in states 1 and 2 are r_1^{-1} and r_2^{-1}, respectively. In each state j (j=1,2), there is another arrival process which is Poisson with rate λ_1 and λ_2 respectively. The rate process is determined by the two-state Markov chain (here the rate is Poisson arrival rate. Notice that this rate process is different from the rate process used in the VBR video source modeling). The process is then fully described by r_1, r_2 and λ_1 and λ_2. Figure 3.2 shows the MMPP process.

The parameters of the model are determined by the following method. Since the single voice source is the renewal process described above, by using renewal theory, its moments of the number of arrivals in an interval can be calculated and the following characteristics of the superposition of n identical independent voice packet processes can therefore be known in terms of parameters of the single source: i) the mean arrival rate; ii) the variance-to -mean ratio of the number of arrivals; iii) the long term variance-to-mean ratio of the number of arrivals; iv) the third moment of the number of arrivals in $(0, t)$[2]. By equating the above four quantities to the corresponding ones of the MMPP, the r_1, r_2, λ_1 and λ_2 can be determined.

The superposition of voice and data sources is a two state MMPP as a whole. This is because the superposition of data streams can be approximated by a Poisson process of rate λ_d and its superposition with a two state MMPP is again a two-state MMPP with arrival rates $\lambda_1 + \lambda_d$, $\lambda_2 + \lambda_d$ and sojourn time r^{-1} and r^{-2}. respectively.

The numerical analysis of an MMPP/G/1 queue is carried out in the paper. The service time distribution is an approximate mixture of the service times of voice and data packets. Results of voice packet delay and data packet delay versus the loads and their delay distributions are compared to the simulated results. Both match reasonably well for commonly used loads.

3.3 Two-State MMPP Model For The Superposition of On-Off Sources

Event: cell arrivals.
Environment: ATM network.
Performance Study: cell loss vs buffer size.

Traffic sources of an ATM network are often assumed to be on-off sources. The superposition of N homogeneous on-off sources can result in a very complicated process. However, if the performance

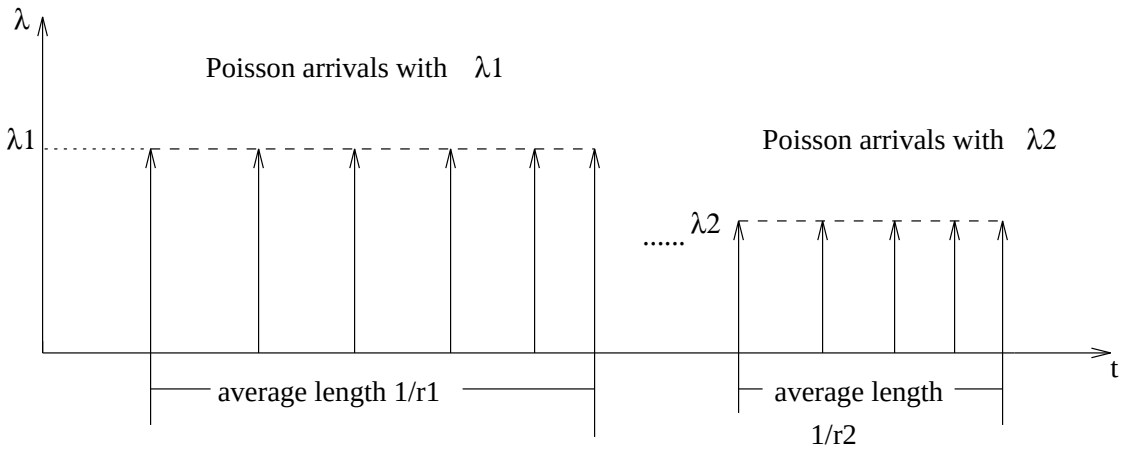

Fig 3.2. MMPP process

analysis is focused only on one parameter, e.g. cell loss rate which is a function of load, it is possible to approximate the superposed N homogeneous on-off sources as a two-state MMPP without losing much information[8].

The two states of the MMPP are chosen to be an overload state and an underload state respectively. In the overload state, the cell emission rate exceeds the output capacity, while in the underload state, cell emission rate is less than the output capacity. The four parameters that describe this MMPP are: mean transition rates of the overload/underload states; the mean arrival rates of the Poisson processes in the overload/underload states. They are determined by a specific analytical method.

The superposed sources and the ATM multiplexer are modeled as a two state MMPP/D/1/K queue. Cell loss behavior vs buffer size was studied for different source parameters and different degrees of superposition. A main result of this model is the evaluation of an upper and an asymptotic bound of the cell loss probability as a function of the buffer size, while the cell loss probability decreases as the buffer size increases. It is also observed that the number of superposed sources alone is not a sufficient parameter to assess the ATM multiplexer loss performance. In fact, it has to be related to the emission characteristics of the single source. In particular, higher activity factors or lower peak bit rates imply better performance as the number of multiplexed sources decreases, and vice versa.

3.4 MMPP Model For Multimedia Traffic

Event: Packet arrivals of multimedia traffic.
Environment: BISDN network.

A much more complicated model than the two-state MMPP model is the multi-dimensional MMPP model for superposed multimedia traffic[9].

Various services will be provided by BISDN such as data, voice, image and video. It is assumed that each source of them can be modeled as either a single two-state Markov chain (voice) or K i.i.d. two-state Markov chains (video), but the sources can have substantially different traffic characteristics and service requirements. The multimedia traffic arriving to a BISDN is the superposition of various packet streams generated by individual input sources. As we described previously, the key properties of each source are, i) average access rate; ii) peak access rate; iii) rate correlation which is defined as cumulative correlation of packet arrivals at different time intervals. Since the multi-media traffic has diverse traffic behavior, these quantities of different sources may differ by many orders of magnitude. So the superposed homogeneous on-off sources will not be a valid model for a general queue analysis.

In [9], each input source, modeled by multiple independent discrete time two-state Markov chains, has its own transition time interval τ_i, (i.e, the peak access rate) to take the diverse burstiness of each source into account. However, there is no analytical technique available for performance evaluation of the queue with multiple time units of τ_i in the arrival process. Since the exact solutions are infeasible, an approximation technique is applied which normalizes the time unit of every arrival to the packet transmission time on the link. The resulting superposed arrival process is then the multi-dimensional discrete Markov process which is a double stochastic process with its rate driven by multiple distinct two-state Markov chains.

A complete queue length distribution is obtained by queue generating function approach. The analytical solution is compared to the simulation results. Furthermore the solution is applied to examine the effect of various correlation and burstiness properties of multi-media traffic on ATM queues. The time unit normalization approximation is therefore verified.

However, due to the limitation of the generating function approach, this method requires infinite buffer and no overload control. This problem can be overcomed by a frequency domain approach presented in section 8. Also this accounts for different traffic characteristics but not different service references.

3.5 Packet Train Model

Event: packet arrivals.
Environment: Token ring local network.

The packet train model was proposed to match the histogram of interarrival times of a total of 11 million packets measured on a token ring local network[3]. It is actually a protocol design-oriented model in which traffic is separated into several streams based on the source and destination. A stream flowing for a given source and destination pair is called a node-pair process. Each node-pair process is then divided into a number of trains. A train consists of packets going in both directions between two nodes. The intertrain time is determined by a so called maximum allowed intercar gap interval. Furthermore each train consists of several tandem trailers, each of which is a sequence of successive packets going from one source to the same destination. A tandem trailer can even be more divided by segments of the same user message. By going through this hierarchy path, the measured correlation of the arrival process is better captured. The parameters of this model which describe the interarrival of trains and cars have concrete physical meanings, therefore are suitable for protocol modeling and designing.

Comment: this model has application in designing bridge, gateway and reservation protocols, but is not

good for performance analysis.

4. Fluid Flow Model

In previous sections, we have shown that the superposed arrival process of multiple on-off sources can be modeled either as the two-state MMPP model, or more accurately the N-state MMPP model. Moreover, the multi-dimensional MMPP model is used for the superposed multimedia traffic. Clearly, in any MMPP model, when the number of states increases, the numerical problem of solving the state equations to determine the performance measures of interest is very involved.

A simpler model, which is considered to retain the long-term correlation characteristics of an arrival process, is the fluid flow model[15]. In a fluid flow model, the arrivals of discrete packets (cells) are treated as a continuous arrival of a liquid flow. The arrivals of individual data units are ignored. For an on-off source, the transmission of data units during an on period is seen at the burst level as a single arrival event with certain flow rate. Thus, the arrivals of a series of on periods form a rate process. If the distribution of the interarrival time is exponentially distributed, the rate process is then a Markov process. This is similar to the MMPP model whose arrivals of each on period at the macro or burst level form a rate process which is a Markov process. If the interarrival time of the underlying on-off source has an nonexponential distribution, the rate process can possibly be a discrete series[3].

An important advantage of this fluid flow model is the analytical simplification of the performance analysis. The fluid flow approximation is reasonable for cases where individual data units are numerous relative to a chosen time scale. In high speed networks, this model is especially useful because the interarrival times of data units transmitted during the on period are small compared to the time between arrival rate changes. Therefore the impact of individual arrivals on the queue performance is negligible.

4.1 Fluid Flow Model of Homogeneous On-Off Sources

Kosten was perhaps the first to employ this approximation for a superposition of exponentially distributed bursts of a Poisson process[23]. Anick later applied this model to a finite number of on-off fluid sources[24] . The following is a brief introduction of the way the model was established[24].

Consider N independent on-off sources which asynchronously alternate between "on" and "off" state. The on periods as well as off periods of all sources are exponentially distributed, though in the analysis below they are not necessarily identical. Assume that the unit of time is the average on period; the average off period with this time unit is denoted by λ^{-1}; the unit of information is chosen to be the amount generated by a source in an average on period. An on source then transmits at the uniform rate of 1 unit of information per unit of time. Thus, when r sources are on simultaneously, the instantaneous receiving rate at the switch is r. The switch stores or buffers the incoming information that is in excess of the maximum transmission rate, c, of an output channel. As long as the buffer is not empty, the instantaneous rate of change of the buffer content is therefore $r - c$.

If at time t the number of on sources equals i, two events can take place during the next small interval Δt, i.e., a new source can start or a source can turn off. Since the on and off periods are exponentially distributed, the arrivals of the successive on periods form a Poisson process with average off time λ^{-1}. Therefore the probability of a single off source turning on in the next Δt is $\lambda \Delta t$ and the probability of any new source being on is $(N - i)\lambda \Delta t$.

Similarly, the arrivals of the off periods form another Poisson process with average on time 1. The probability of any on source turning off is then $i\Delta t$. Compound events have probability $o(\Delta t^2)$. The probability of no change is $1 - \{(N - i)\lambda + i\}\Delta t + o(\Delta t^2)$.

Let $P_i(t, x)$, $0 \le i \le N$, $t \ge 0$, $x \ge 0$, be the probability that at time t, i sources are on and the buffer content does not exceed x. Then

$$P_i(t + \Delta t,\ x) = \{N - (i - 1)\}\lambda\Delta t\ P_{i-1}(t, x) + (i + 1)\Delta t\ P_{i+1}(t, x) + [1 - \{(N - i)\lambda + i\}\Delta t]P_i\{(t,\ x - (i - c)\Delta t\} + o(\Delta t^2).$$

Passing to the limit $\Delta t \to 0$, we get

$$\frac{\partial P_i}{\partial t} + (i - c)\frac{\partial P_i}{\partial x} = (N - i + 1)\lambda P_{i-1} - \{(N - i)\lambda + i\}P_i + (i + 1)P_{i+1}.$$

Since only the time-independent equilibrium probabilities are of interest, we set $\partial P_i/\partial t = 0$ in the above equation and define $F_i(x)$ as the equilibrium probability that i sources are on and buffer content does not exceed x. For $i \subset [0, N]$, we obtain

$$(i - c)\frac{dF_i}{dx} = (N - i\ + 1)\lambda F_{i-1} - \{(N - i)\lambda + i\}F_i + (i + 1)F_{i+1}.$$

In matrix notation,

$$\mathbf{D}\frac{d\mathbf{F}(x)}{dx} = \mathbf{M}\mathbf{F}(x), \qquad x \ge 0,$$

where $\mathbf{D} = \text{diag}\{-c,\ 1 - c,\ 2 - c, \ldots,\ N - c\}$ and

$$\mathbf{M} = \begin{bmatrix} -N\lambda & 1 & & & & \\ N\lambda & -\{(N-1)\lambda + 1\} & 2 & & & \\ & (N-1)\lambda & -\{(N-2)\lambda + 2\} & 3 & & \\ & & \cdot & & & \\ & & \cdot & & & \\ & & \cdot & & & \\ & & & & 2\lambda - \{\lambda + (N-1)\} & N \\ & & & & \lambda & -N \end{bmatrix}$$

Thus, the fluid flow model is established through a set of differential equations that describe the equilibrium buffer distribution. The key approximation used for this model is to assume Poisson processes for the on-period event and off-period event. The individual arrivals of data units are totally neglected. The way of solving the above set of differential equations follows the formal mathematical

method, i.e., finding the eigenvalues and eigenvectors of the matrix equation which is not complicated when the generating function method is used.

Based on this work, models for nonexponentially distributed on and off periods (e.g., Erlang or hyperexponential distributions) were proposed in other papers[24]. Also the Anick model was later generalized for a finite buffer capacity case[24] .

4.2 Fluid Flow Model For Variable Bit Rate Sources

The sources used in section 3.1 are on-off sources which emit data units at a constant rate during the on periods. For ATM networks, traffic can be generated by variable bit rate sources. This means that at the cell level, the cell emission rate may vary continuously in the sense that the interarrival time between cells varies gradually. So there are two correlations that can have impact on the queue performance. At the burst level, there is correlation of the arrival rate process as was considered in section 4.1; at the cell level, there is correlation between interarrival times of successive cells. A fluid flow model which accounts for both burst level and cell level processes of the superposed variable bit rate sources in a ATM multiplexer is presented in [15].

The fluid flow model for the burst level is obtained from a waiting time distribution equation rather than the equilibrium buffer distribution differential equations used in Anick's model.

Let $W(t)$ $(t \geq 0)$ be the amount of work arriving to the system in the interval $(-t, 0)$ and V_t be the virtual waiting time at -t. The virtual waiting time is defined as the time a customer would have to wait for service if he arrived at time -t. It is equivalent to the amount of work remaining to be done at time $-t$[21]. Define $X(t) = W(t) - t$ to be the excess work or the overload arriving in $(-t, 0)$. The virtual waiting time at time 0 is then

$$V_0 = \sup_{t \geq 0} \{X(t)\}^{[15][21]} .$$

If $W(t)$ is a continuous function, the considered system behaves like a reservoir with constant output whenever its content V_t is non-zero. Let Λ_t be the arrival rate at time -t:

$$\Lambda_t = \frac{dW(t)}{dt} = 1 + \frac{dX(t)}{dt} .$$

Let $v(x)$ be the complementary distribution of V_0: $v(x) = P\{V_0 > x\}$. Instant u is defined as a time point at which $\{X(u) = x$ and $V_u = 0\}$. The equation of the complementary distribution of the waiting time is then

$$v(x) = \int_{u>0} \int_{0 \leq \lambda \leq 1} (1 - \lambda) \frac{d^2}{dxdt} P\{X(u) \leq x, \Lambda_u \leq \lambda \text{ and } V_u = 0\} d\lambda du.$$

A fluid input process consisting of a superposition of multiple independent statistically identical on-off sources can be derived from this equation and has a similar form. Thus the model is built at the burst

level.

The cell level fluid flow model, which is a simple stationary fluid process defined by a point process (see [15] pp. 383 for detail), is then coupled to the burst level model such that the virtual waiting time results from the burst level component and virtual waiting time results from the cell level component sum up to the total virtual waiting time V_t. Thus, this composite model is considered to be a quite accurate expression of the real queue rather than just being an approximation. The result of the distribution shows that the first moments of the delay distribution may significantly depend on the cell level component while the buffer dimensioning is given essentially by the burst level component.

5. Variable Bit Rate Video Source Modeling

We introduce the models for VBR video sources as a separate category, since there has been an intense focus on the study of video sources. The importance of video traffic study lies in the fact that it will become a major service provided by high speed networks.

A video source is a variable bit rate source. At the frame level (the individual still images), it has a constant interarrival interval during the active period, but each arrival frame contains a different number of bits. It differs from the data and voice sources in that a single video source can be very bursty in the sense of instantaneous very large bit rate due to the effect of scene complexity and motion. The burstiness also depends on the different coding algorithms used. In an ATM network environment which has a variable bit rate channel, the transmission of variable rate video signals is expected to achieve better quality than what the constant bit rate channel can do. In order to efficiently use the ATM network resources and to get good quality of service for transmission (small queue length and little loss), it is necessary to compress video data to reduce its mean output bit rate. This job is done by a video codec. As a consequence, the codec i) outputs at a lower average rate than CBR codecs, ii) generates a bursty traffic which has correlated cell arrivals, iii) is able to maintain a more consistent user selectable video quality without running at the peak rate. Once the video has been coded, the resultant data is presented to and transported through the ATM network in the form of cells.

A number of variable bit rate video codecs have been proposed in the recent literature. These codecs have used a number of image coding methods, either individually or in combination. These methods are interframe differential pulse code modulation (DPCM), intraframe DPCM, conditional replenishment (CR), motion compensation (MC), transform coding, most notably discrete cosine (DCT), run length coding (RLC), and syntactic coding (SC). The CR codec in section 4.1 is usually used for video teleconferencing. It combines intraframe DPCM for areas of movement and RLC for areas with no movement.

Since real time video applications require large bandwidth and have very stringent delay and loss requirements, the modeling of video sources and the performance study of a video multiplexer become extremely important. Earlier models that use the discrete-time discrete-state Markov chain[14] or the packet-train model[15] are proved to be inadequate to model the autocovariance of the video sample source. The recent studies include the following different models corresponding to different coding schemes. Notice that different coding schemes give rise to different arrival correlations.

5.1 Autoregressive Moving Average Model

Event: cell arrivals of VBR video source.
Environment: ATM variable bit rate channel.

For conditional replenishment (CR) coding, it is measured that the cell arrival process of a single video source is recorrelated at each video frame and line interval and this recorrelation effect decreases with increased lag[4] (Figure 5.1.) The autoregressive moving average (ARMA) model also exhibits this recorrelation feature, so it is used to approximate the single video cell stream with the assumption that the video cell arrival process is a wide sense stationary (WSS) process.

Fig.5.1 Autocorrelation function of the cell arrivals of CR-coded video source[4]

The periodic occurrence of the autocovariance is modeled by the autoregressive part of the ARMA model, while the correlations at different lags are modeled by the moving average part.

The parameters of the ARMA model are estimated in the following way. The long-term mean, variance, and autocovariance function are measured from the cell output of the simulated video codec. These are then used to estimate the parameters of a mathematical model of an ARMA process. The mathematical ARMA model is realized in this paper by a transverse filter of finite order and a recursive filter. Cell interarrival sequence coming out of the filter is fed to a queuing system for simulation study of the queue performance.

Assuming that the cell arrivals from multiple ARMA models on the inlet are statistically independent and the individual arrival process is wide sense stationary, then the multiplexed video stream is also a WSS ARMA process. Its autocovariance function is the sum of the autocovariance functions of the individual cell streams. Therefore the multiplexer is modeled as a G/D/1 queue. G here refers to a general arrival process (in this case ARMA). The performance measurements include the mean, variance, complimentary probability distribution function of the cell waiting time and cell loss due to overflow. Perturbation effect of the arrival's autocorrelation function on the expected cell waiting time is also measured. The performance analysis shows that there exists a critical load. Below this load, e.g. at loads around 0.55, the correlation of the arrival process does not really affect the waiting times, and neither does the perturbation of the input's autocovariance function. This means that within this load range the arrival correlation does not influence the queuing performance. Above this load, the correlation of the input process becomes decisive for the cell waiting time and cell loss probability. Both of them grow abruptly in this case due to overflow. Therefore tradeoff between the correlation and load plays an important role for the design of the network.

5.2 Modeling With Scene Changes: Modulated ARMA Model

Event: frame level rate process (bit rate) of full motion packet video source.
Environment: Broadband packet network.
Study: performance of multiplexer of multiple full motion video sources.

The modeling of full motion video sources needs to take the scene changes into account, since these create considerable source burstiness. In this case a model that only captures the autocorrelation of the arrival process will not be a good approximation. A model of a rate process (bit rate) developed to fit the measured data of a full motion video source with layered coding is presented in [6]. Layered coding is a scheme which is usually used for high bandwidth demanding broadcast video. In addition to the autocorrelation function of the rate process, the model tries to match the empirical sample data path reasonably, especially to account for spikes observed during scene changes.

Let T_n be the bit rate (in Mbits/sec) in the n^{th} frame, the stationary process is defined by

$$T_n = max[0,\ X_n + Y_n + K_n C_n],$$

where X_n and Y_n are two independent first order autoregressive processes. The bits generated by the sum of these two processes are to model the empirical autocorrelation function of the bit rate process. It is noticed that one autoregressive process is not sufficient to mimic the empirical autocorrelation function. The autocorrelation function of the empirical data is shown in Figure 5.2. C_n is a sequence of independent normally distributed random variables with mean $\alpha/2$ and standard deviation $\beta/2$. K_n is a three-state Markov chain $\{K_n = 0, 1, 2, n = 0, 1, 2, 3\ldots\}$ as shown in Figure 5.3. $K_n C_n$ characterizes scene changes by determining the extra bits generated at scene changes. Normally the Markov chain is in state 0, a transition to state 2 with probability p models the scene change. The next transition to state 1 propagates the effect of the scene change and makes the effect of scene change last for two frames. In the first frame after a scene change ($K_n = 2$), the extra bits generated by $K_n C_n$ in Mbits/sec have mean α and standard deviation β. In the next frame ($K_{n+1} = 1$), the mean and standard deviation are $\alpha/2$ and $\beta/2$ respectively. Thereafter, $K_{n+2} = 0$ and there are no extra bits contributed to the bits generated from $X_n + Y_n$.

The parameters of the model are determined by matching its statistics to the correspondingly measured ones. These statistics are mean rate, peak rate, variance, the autocorrelation function, the average rate at which scene changes occur and the frame rate. The model is used for a simulation study of the performance of a video multiplexer with loss priority. The multiplexer multiplexed several variable bit rate full motion video sources. A service strategy with loss priority is proposed. Packets from the codec are marked as essential and non-essential packets respectively, where essential packets are to reproduce the basic picture at the receiver. Therefore they are served with higher priority at the statistical multiplexer and delivered without loss. The nonessential packets are served with lower priority and are dropped in case of congestion. The packet loss rate determines the picture quality. The queue model is shown in Figure 5.4. For a single source, the essential packets of frame n are first transmitted followed by the non-essential packets and then followed by the essential packets of frame $n+1$. When the essential bits from a source arrive for the n^{th} frame, the non-essential bits from the same source for the $(n-1)^{th}$ frame are dropped, if they are still present in the queue. This is done because all the packets have very stringent delay requirement and have to be delivered in sequence. The server either serves the high priority essential buffer at the full capacity of the output line or divides its output line capacity

Fig.5.2 Autocorrelation function of the rate process of layered coded video source[6]

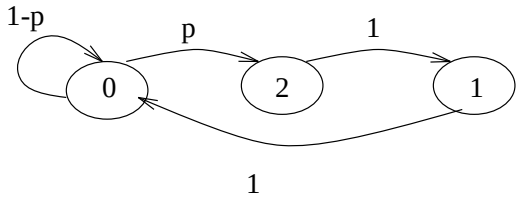

Fig 5.3 Three-state Markov chain

equally among the low-priority nonessential buffers that may need service (when the essential buffer is empty).

The relationship of the number of video sources that can be multiplexed vs the packet loss probability and queuing delay is studied. The result shows that with small threshold values t for essential packets (if frame n contains T_n bits, t bits are assigned to essential packets, and the rest $T_n - t$ bits are assigned to non-essential packets), a large number of video sources are allowed to be multiplexed. In this case all the essential packets are guaranteed to be delivered without loss or excessive delay with only a small loss for non-essential packets. However, as the number of multiplexed sources increases, the threshold for

Fig.5.4 Queue model with loss priority[6]

essential packets must be lowered, which will result in larger queue length or higher loss for non-essential packets. It is also observed that the loss rate of non-essential packets increases as the occurring rate of scene changes increases, which results in a degradation in picture quality. So there is a trade-off between picture quality and the number of sources multiplexed.

By suitable choice of model parameters, a variety of video sources and video applications ranging from video telephony to full motion video with scene changes can be modeled with this method.

5.3 TES Method

Event: general autocorrelated time series.
Environment: not specific.

For more accuracy, it is hoped to devise a model having

- A marginal distribution that matches its empirical counterpart;
- an autocorrelation function that approximates its empirical counterpart; and
- sample paths that "resemble" the empirical data.

The Transform-Expand-Sample (TES) method provides an approach to this aim. It can model general autocorrelation time series and provide a large degree of freedom in approximating the empirical autocorrelation function while simultaneously matching the marginal distribution of the empirical counterpart[7][20].

The main feature of the TES method is using a stochastic transformation technique to construct a time

series from the basic TES process with its autocorrelation function and marginal distribution matched to the empirical ones. The basic TES process provides a wide variety of autocorrelation structures. This feature of the TES method enables its modeling flexibility, which means that TES is a class of models rather than one model for generating autocorrelated random sequences with arbitrary marginals. The models in the literature, however, are flexible only in the sense that the parameters of a model can be selected within some range according to the empirical data.

TES processes are a variation on classical autoregressive schemes, but with modulo-1 arithmetic. Pure TES processes are stationary Markovian and come in two flavors: TES$^+$, which gives rise to a sequence $\{U_n^+\}$, and TES$^-$, which gives rise to a sequence $\{U_n^-\}$. They are defined by

$$U_n^+ = \begin{cases} U_0 & \text{if } n = 0 \\ < U_{n-1}^+ + V_n > & \text{if } n > 0 \end{cases} \tag{5-1}$$

$$U_n^- = \begin{cases} U_n^+ & n \text{ even} \\ 1 - U_n^+ & n \text{ odd} \end{cases} \tag{5-2}$$

where the $\{V_n\}$ is a sequence of i.i.d. random variables with a common density f_v; it is referred to as the innovation sequence. The angle bracket gives the fractional part of the inside quantity. The superscripts suggest that TES methods achieve coverage of the full range of feasible lag-1 autocorrelation; TES$^+$ methods cover the positive range [0, 1], while TES$^-$ methods cover the negative range [-1, 0]. Each of the above equations gives rise to a sequence of stationary random variables with uniform marginals on [0, 1) and a wide variety of autocorrelation structures. $\{U_n^+\}$ and $\{U_n^-\}$ are called background sequences.

The fact that the $\{U_n\}$ all have a uniform marginal distributions on [0, 1) allows a random sequence with essentially arbitrary marginals to be generated if the distribution function of that sequence is known. For a given empirical data, if its marginal distribution function F is known, the corresponding random time sequences are given by

$$X_n^+ = D\,(U_n^+), \qquad X_n^- = D\,(U_n^-)$$

where D is referred to as a distortion function that transforms the random sequence $\{U_n\}$ into sequence $\{X_n\}$. Both $\{X_n^+\}$ and $\{X_n^-\}$ are stationary with marginal distribution F and they are called the foreground sequences. The marginal distribution F obtained from empirical data is often modeled by an empirical histogram H which is a step function density. The transform function D has therefore the following form

$$D_H(x) = F^{-1} = \sum_{k=1}^{N} 1_{[C_{k-1}, C_k)}(x)[l_k + (x - C_{k-1})\,\frac{w_k}{p_k}], \qquad 0 \le x \le 1 \tag{5-3}$$

where N is the number of histogram cells of the form $[l_k, r_k]$, $w_k = r_k - l_k$ is the width of cell k, 1_A is the indicator function of set A which is 1 inside set A and 0 otherwise, p_k is the probability of cell k and $C_k = \sum_{j=1}^{k} p_k$ is the cumulative distribution of $\{p_k\}$ ($C_0 = 0$ and $C_N = 1$). For a given lag τ, the

autocorrelation functions $\rho_X^+(\tau)$ of $\{X_n^+\}$ and $\rho_X^-(\tau)$ of $\{X_n^-\}$ that match the empirical counterpart are given respectively by

$$\rho_X^+(\tau) = \frac{2}{\sigma_X^2} \sum_{\upsilon=1}^{\infty} R[\, \tilde{f}_V^{\tau}(i2\pi\upsilon)] \, |\tilde{D}(i2\pi\upsilon) |^2 \tag{5-4}$$

$$\rho_X^-(\tau) = \begin{cases} \rho_X^+(\tau), & \tau \text{ even} \\ \dfrac{2}{\sigma_X^2} \displaystyle\sum_{\upsilon=1}^{\infty} R[\, \tilde{f}_V^{\tau}(i2\pi\upsilon)] \; R[\tilde{D}(i2\pi\upsilon)^{\,2}], & \tau \text{ odd} \end{cases} \tag{5-5}$$

where σ_X^2 is the common variance of $\{X_n^+\}$ and $\{X_n^-\}$, tilde denotes the Laplace Transform, f_V^{τ} is the τ-fold convolution of the innovation density f_V, R denotes the real part operator , and $i = \sqrt{-1}$. Obviously the innovation plays a role in forming the autocorrelation function. Notice that the series converge rapidly, since the convolution factors decay rapidly in the lag. This fast decaying feature of the autocorrelation function, compared with the autocorrelation function of a self-similar process model, which is another video source model presented in the next section, shows a typical short-range dependence. The self-similar model claims that the video source possesses long-range dependence rather than short-range dependence, because its autocorrelation function decays slower.

Complete TES modeling comprises the following methodology: specify a flavor of TES process; a distortion D; an innovation density f_V. The selection of the TES process sign is based on the experience and familiarity with the fundamental sample path behavior of TES processes. Most real-life data call for a TES^+ model. In contrast, TES^- are rarely used. D depends on the empirical histogram and is given by equation (5-3). The core activity of the TES modeling process is a heuristic search for a suitable innovation density. Different innovation densities usually give rise to different autocorrelation functions, thereby providing a large degree of freedom in fitting autocorrelation functions.

A software package named TEStool[20] has been developed as an interactive computerized modeling environment based on the above methodology. With given empirical time series data, the modeling can be done interactively within a short time.

Comment: The TES method can model the frame level bit rate process of VBR video source as well as its arrival process. In the former case, the $\{X_n\}$ is a random sequence representing bits/frame. In the latter case, $\{X_n\}$ is the discrete interarrival times. Besides the VBR video source, the TES method can be used to model other arrival processes.

6. Self-Similar Phenomena

6.1 Self-Similar Phenomenon

Event: Packet arrivals from interconnected Ethernet traffic.

Motivated by understanding the interconnections between LANs and the proposed B-ISDN interconnection network, recent studies of high-quality high-resolution Ethernet LAN traffic measurements have revealed a new fractal behavior of the traffic which none of the commonly used traffic models is able to capture[16]. The rigorous statistical analysis of hundreds of millions of high quality (high time resolution and low loss rates) Ethernet traffic measurements collected between 1989 and 1992 shows that the traffic display structurally similar traffic bursts across all time scales from a few millisecond to minutes and hours. This is called self-similar phenomena.

A high quality Ethernet monitor was used to record the Ethernet packets without loss (irrespective of the traffic load). Timestamps that represent the arrival time of the end of the packet and packet length are recorded as well. Figure 6. 1[16] shows the measured packet arrivals on different time scales. All the plots have "similar" burstiness.

This self-similar or fractal-like behavior of aggregated Ethernet LAN traffic is very different from currently considered formal models for packet traffic, such as the previously presented MMPP model, packet-train models and fluid-flow models. It provides a new look at traffic modeling and performance characteristics of broadband networks.

The critical characteristic of this self-similar traffic is that there is no natural length of a "burst". This is shown in Figure 6.1 where we can see that as time scale becomes finer the number of packets arrive per time unit still possess a burst nature. Therefore, the burstiness here is a concept relative to the time scale. The more bursty a traffic is, the finer time scale will be from which the traffic still shows the burstiness, which means the more self-similar the traffic is. So the measurement of the degree of burstiness can be measured by the degree of self-similarity. The most striking conclusion obtained is that the burstiness (degree of self-similarity) of LAN traffic typically intensifies as the number of active traffic sources increases, contrary to commonly accepted views from formal models which say that aggregate traffic becomes smoother (less bursty) as the number of traffic sources increases (which is consistent with theoretical conclusion). Furthermore the nature of congestion produced by self-similar network traffic differs drastically from that predicted by traffic models currently considered in the literature and is far more complex than has been typically assumed in the past.

Since LAN interconnection (i.e., interconnecting different local area networks through a connectionless service) will soon become one of the major traffic contributors for B-ISDN, its self-similar behavior has serious implications for the design, control and analysis of the high speed, cell-based networks.

Mathematical definition of a self-similar process is as follows: Let X be a wide-sense stationary process. Assume that X has an autocorrelation function which exhibits long-range dependence with the form

$$r(k) \approx k^{-\beta} L_1(k) \qquad \text{as } k \to \infty, \tag{6-1}$$

where $0 < \beta < 1$ and L_1 is slowly varying at infinity, that is, $\lim_{t\to\infty} L_1(tx)/L_1(t) = 1$. Let

Fig.6.1 "Pictorial" proof of self-similarity: Ethernet traffic on 5 different time scales[16].

$X^{(m)} = (X_k^{(m)}: k = 1, 2, 3, \ldots)$ denote a new time series obtained by averaging (arithmetic mean) the original series X over nonoverlapping blocks of size m. Then process X is called (exactly second-order) self-similar with self-similarity parameter $H = 1 - \beta/2$ if the corresponding aggregated process $X^{(m)}$ has the same correlation structure as X, i.e.,

$$r^{(m)}(k) = r(k), \quad \text{for all } m = 1, 2, 3, \ldots (\, k = 1, 2, 3, \ldots).$$

The basic difference between self-similar traffic and the conventional models lies in that:

- The aggregated processes $X^{(m)}$ possess a nondegenerate correlation structure as $m \to \infty$, while the conventional models all have the property that the autocorrelation functions of their aggregated processes tend to be zero as $m \to \infty$. This is illustrated in the plots in Figure 6.1. If the original time series X represents the number of Ethernet packets per 10 milliseconds (plot (e)), then plots (a) to (d) depict segments of the aggregated time series $X^{(10000)}$, $X^{(1000)}$, $X^{(100)}$, and $X^{(10)}$, respectively. All the plots look "similar" and distinctively different from pure noise, which means they are not independently distributed random variables.
- The autocorrelation function of a self-similar process decays hyperbolically as the lag increases and $\sum_k r(k) = \infty$, which implies that although high-lag correlations are all individually small, their cumulative effect is of importance and gives rise to features which are drastically different from those of the more conventional short-range dependent processes. On the other hand, the latter is characterized by an exponential decay of the autocorrelations resulting in a summable autocorrelation function $0 < \sum_k r(k) < \infty$.
- In the frequency domain, the spectral density of a self-similar process obeys a power-law behavior near the origin. i.e.,

  $$f(\lambda) \approx \lambda^{-\gamma} L_2(\lambda), \quad \text{as } \lambda \to 0, \tag{6-2}$$

  where $0 < \gamma < 1$, L_2 is slowly varying at 0, and $f(\lambda) = \sum_k r(k) e^{ik\lambda}$ denotes the spectral density function. This implies that $f(0) = \sum_k r(k) = \infty$, that is, the spectral density tends to $+\infty$ as the frequency λ approaches 0. On the other hand, processes that only possess short-range dependence is characterized by a spectral density function $f(\lambda)$ which is positive and finite at $\lambda = 0$.
- The variance of the arithmetic mean $X^{(m)}$ decreases more slowly than the reciprocal of the sample size, that is

  $$var(X^{(m)}) \approx am^{-\beta}, \quad as\ m \to \infty, \tag{6-3}$$

  where a is a finite positive constant independent of m, and $0 < \beta < 1$. The β here is the same as in (6-1) and is related to the γ in (6-2) by $\beta = 1 - \gamma$. On the other hand, for stationary processes whose aggregated series $X^{(m)}$ tend to second-order pure noise (i.e., $r^{(m)}(k) \to 0,$) the sequence $\{var(X^{(m)}): m \geq 1\}$ satisfies

$$var(X^{(m)}) \approx bm^{-1},\ as\ m \to \infty,$$

where b is a finite positive constant independent of m.

Equations (6-1), (6-2), (6-3) are equivalent definitions for a self-similar process.

The empirical data sets are tested to be self-similar from three different approaches: i) analysis of the variances of the aggregated processes $X^{(m)}$, ii) time-domain analysis, iii) periodogram-based analysis in the frequency-domain. Since a self-similar process has unique properties in these three aspects, the testing is accurate and reliable. The Hurst parameter H which measures the degree of long-range dependence or the burstiness is also estimated from the above analysis. Different data sets collected from different times and different locations show that the different degrees of self-similarity H depend on the utilization level of the Ethernet and that H increases as the utilization increases.

Comparison of the above three second-order statistical properties of conventional models to measured Ethernet data shows an obvious difference.

Although the data are collected from LAN's in the same company, they are collected over a long period of time as well as from different positions in the network, irrespective of the utilization level of the Ethernet. So some of the characteristics uncovered are likely to be universal for LAN traffic and moreover, are likely to be inherent of packet traffic in many high-speed networks of the future. For example, the analysis of external traffic as a component of internal traffic shows that it has the similar self-similarity as internal traffic only with slightly different Hurst parameter. This external traffic can be viewed as representative for LAN interconnection services which are expected to contribute significantly to future broadband traffic.

Stochastic models for self-similar phenomena can be of two types. One type is an exactly self-similar fractional Gaussian noise process which is a stationary Gaussian process with autocorrelation function

$$r(k) = 1/2\ (\ |k+1|^{2H} - |k|^{2H} + |k-1|^{2H}\ ) \qquad k = 1, 2, 3\ldots.$$

The other type is an asymptotically self-similar fractional autoregressive integrated moving-average (ARIMA) process. The fractional ARIMA processes are much more flexible with regard to the simultaneous modeling of the short-term and long-term behavior of a time series than fractional Gaussian noise[26].

Another interesting feature of the self-similar process is that its burstiness, characterized by the Hurst parameter, can be measured alternatively by the index of dispersion (for counts), which increases monotonically throughout a time span. However, because of the dependency of the time interval and the potential "heavy-tailedness" of the interarrival times, the other two commonly used measures of burstiness, peak-to-mean ratio and coefficient of variation, are not suitable to measure the burstiness of the self-similar process.

Parallel works on congestion management in the presence of self-similar traffic[17][18] suggest that the congestion phenomenon seen in the presence of self-similar traffic differ drastically from those predicted by the formal conventional traffic models. Therefore, they provide convincing evidence for the

significance of self-similar or fractal network traffic for engineering future integrated high-speed data networks.

Considering the access class scheme proposed for switched multimega digital service (SMDS) on a public B-ISDN, Leland and Wilson[17] employ the Ethernet data in a trace-driven simulation of a LAN/B-ISDN interface to observe the effect of the actual aggregate LAN traffic on the behavior of the SMDS service interface buffer, i.e. the relationship between packet delay, packet loss and the amount of buffering at the interface.

The results show that overall packet loss decreases very slowly with increasing buffer capacity, in sharp contrast to Poisson-based models where losses decrease exponentially fast with increasing buffer size[8]. Moreover, packet delay (95th percentile) always increases with buffer capacity, again in contrast to the formal models where delay does not exceed a fixed limit regardless of buffer size[17]. This behavior, according to the author, is typical for self-similar traffic and can be readily explained using its properties.

On the other hand, Fowler and Leland[18] simulate a simple network that provides LAN interconnection and study the congestion behavior due to traffic access contention. The combined traffic is modeled using the measured Ethernet traces. Their study reveals that:

- There exist large variations in the network traffic on time scales of hours, days, or months; this aggravates careful sizing of network components, since small errors in engineering can incur drastic penalties in loss or delay.
- Although some of the standard traffic models suggest that congestion problems essentially disappear with sufficient buffer capacity, realistic network traffic shows that such behavior cannot be expected; large buffers will not prevent congestion from occurring but introduce instead undesirable delay characteristics.
- During congestion periods, congestion persists long enough for the effects of user and protocol responses to be felt.
- A detailed examination of congestion periods shows that when congestion occurs, losses are severely concentrated and are far greater than the background loss rate; losses may exceed the long-run loss probability by an order of magnitude during the first second following the onset of congestion, while the losses are elevated by over two orders of magnitude during the first 100 milliseconds.
- Fortunately, many congestion episodes are preceded by signs of impending danger; whether detecting congestion or activating congestion avoidance responses can be done reliably far enough in advance of an actual congestion period requires further study.

6.2 Self-Similar Video Traffic Source

Event: frame level cell rate (bit rate) process of general video sources.
Environment: ATM network.

Not only is the self-similar phenomena observed in Ethernet traffic, it is also observed in the VBR video source.

All the video source models that are described in section 5 are based on the specific scene (e.g. video phone, video conference, motion picture video) and specific coding scheme. Therefore, one model only fits for one video source.

Motivated by the desire to find a more universal property inherent to VBR video traffic which is independent of scene and codec, a study of 20 sets of actual VBR video data, generated by a variety of different codecs that lasts for 1/2 and 2 hours and representing a wide range of different scenes is carried out in [22]. All data sets are collected with high time resolution. This work reveals:

- A common predominant characteristic of all the VBR video data is that the frame level video rate process $\{X_{k,}\ 0 < k < T\}$ possesses long-range dependence with different degrees for different data sets. X_k denotes the number of ATM cells which contain the compressed and coded information for frame k ($0 < k < T$) over intervals of arbitrary length T;
- The intensity of long-range dependence (measured in terms of Hurst parameter H) depends on the activity level of the recored scene. Video-conference and video phone are low activity scenes, while video TV and full motion pictures are high activity scenes.

These findings, which are not intuitively obvious, challenge the currently proposed models in the literature, all of which concentrate on short-range dependence and are not able to account for long-range dependence. It is also argued that the previous models only tried to capture the marginals and autocorrelations, which is helpful but insufficient, since marginals and autocorrelations are only two of many aspects that contribute to the process.

The long-range dependence is characterized by an autocorrelation function that decays hyperbolically as the lag increases. This in turn results in a non-converging result of the sum of autocorrelation with lags (see section 6.1). On the other hand the short-range dependence is characterized by exponential decay of the autocorrelation function which results in a summable autocorrelation function. This long-range dependence phenomena is exactly the same as what has been observed in the Ethernet aggregated traffic. Therefore, all the analysis (e.g. the long-range dependence testing and the source modeling) follows the same way as was presented in the self-similar process in section 6.1. The Hurst parameter is again used to measure the degree of dependence. The ARIMA model is considered to be one of the possible models for simultaneous modeling of the long-range and short-range behavior, as well as for accommodation of a large class of possible marginal distributions of a video traffic source. Models that combine the proposed short range dependence models with a long-range process are also possible. Parameter estimation techniques for the ARIMA and alternating models are still under study.

Performance implications of this long-range dependence of video source are still unknown and are currently under investigation.

Comment: Similar to the Ethernet self-similar traffic, the long range dependence in the video traffic is observed from data that are collected in high time resolution. It shows again that the way in which data are measured is critical to accurate modeling.

6.3 Acquired Data Sets

Data sets that show the self-similarity of Ethernet traffic and video data sets have been acquired from Bellcore.

There are four data sets of Ethernet packet arrivals collected from different times in Bellcore Morristown Research and Engineering facility. Two of the data sets are records of internal traffic which go within the Lab. The other two are external traffic which are to and from the internet and all of the Bellcore. Each data set is the first 1 million arrivals of the day-long trace. Each line of the data set file contains the time (in seconds) of the packet arrival since the start of a trace and the packet length (in bytes). Packet length ranges from minimum size of 64 bytes to maximum 1518 bytes. The length field does not include the Ethernet preamble, header of CRC.

The video data is collected from the two hour movie ‘‘Star Wars’’ variable bit-rate encoded video which consists of 60 data files. The following information contents are included in each file:

- bytes per frame for 171,000 frames (approximately 2 hours). The frame rate is 24 frames/second. (i.e. the original film rate).
- bytes per slice, which is 16 video lines. There are 30 slices per frame. One slice represents the information transmitted in about 1.4 milliseconds.

These 60 data sets are collected from a specific codec which is simplified to allow computation of a complete movie.

Both the Ethernet data and the video data will be used as real traffic source to feed into existing simulation model for further study. Potential work on that can have two aspects:

- performance comparison with the real traffic to the performance with previously used BONeS traffic models.
- construct BONeS model of the self-similar traffic such that it has the similar pattern of the autocorrelation function possessed by the real traffic, or, it results in the similar performance as the real traffic does.

7. Chaotic Map Approach

While the self-similar phenomena reveals contradictions with conventional models, Erramilli and Singh[19] state that the conventional models in the literature, though they have been used with some degree of success, they do not capture all the relevant statistical characteristics of the process. They address the problem of packet traffic modeling from a fundamentally different perspective by using deterministic, nonlinear chaotic maps as models for packet traffic.

Chaos is a dynamical system phenomenon in which simple, low order deterministic processes can produce behavior that mimics random processes. The connection between randomness and chaos is well established mathematical theory. Chaotic models have been proposed as alternatives to Markov models in the biophysics field[13]. The reason for exploring chaotic models for packet traffic is that there is a degree of determinism in integrated packet traffic sources, particularly circuit emulation type sources, and service times are often deterministic as well. Deterministic chaotic models may therefore allow a more compact description of all the complexities in the traffic that are relevant to performance analysis. Moreover it is capable of analyzing the dynamical behavior of broadband packet networks which is important and of great interest.

According to the authors, the self similar phenomena is a property associated with chaotic phenomena. It is expected that the chaotic map model will complement the variety of stochastic models that have been proposed and lead to more complicated characterization of packet traffic.

The model is still under study.

8. Frequency Domain Queuing Analysis Approach

We have seen in section 3.4 that a key issue in modeling multimedia traffic is how to characterize the input correlation function. In integrated service networks, the multimedia input stream, like voice or video, can be modeled by a Markov modulated Poisson process. The function of the Markov chains is to characterize the diversity of time autocorrelations existing in multimedia traffic streams. However, the queuing analysis becomes very complicated and involves solutions of a multi-dimensional Markov queuing process with an extremely large state space. So far only the two-state Markov chain has been adopted for modeling of voice and video sources. Although it captures the correlation and burstiness properties of each source, a fundamental limit of using two-state Markov chains is that their autocorrelation functions have to be geometric in form. This is certainly insufficient to represent the entire spectrum of multimedia sources. For example, packet video sources can have a pseudo-period input autocorrelation pattern.

Queuing analysis of multimedia traffic in the frequency domain is therefore presented in [11]. The idea is based on the fact that each input traffic stream, as a stationary random process, is mainly captured by both its steady state distribution function and correlation function. So the correlation function in the time-domain can be transformed into a power spectral distribution function in the frequency domain. Queue response to input spectral properties can then be made by spectral theory of random process.

Each input traffic stream in [11] is still modeled as an independent Markov modulated Poisson process, where the underlying Markov chain is to reflect the time autocorrelation properties of the input process at the macro level, while at each state of Markov chain, namely the micro level, the local individual packet arrivals are modeled by a Poisson process with a certain arrival rate. Alternatively the Markov chain can be referred to an input rate process $r(n)$, since each state of chain i corresponds to a different Poisson arrival rate r_i. The autocorrelation function of the input rate process is used instead of that of input arrival process, because it is known that the correlation of the input arrival process has less effect on queues than that of the input rate process.

In the discrete spectral analysis, the input rate process is an N-state discrete-time periodic Markov chain (which is no more a random process). Key parameters that relate the autocorrelation function and power spectral density are the eigenvalues of the transition matrix of the Markov chain. The queue analysis considers a queue with infinite buffer and deterministic service time. The queue length behavior vs input power spectrum reveals that, the higher the positive correlation in the time domain, which causes longer queue length, the greater the concentration of low frequency powers in the spectral domain. Therefore, input power at low frequencies has much more impact on queuing performance than input power at high frequencies. Furthermore, the effect of input powers at high frequencies can generally be neglected in queuing analysis. The same conclusion is made for the continuous spectral case.

A later paper[10] explores the interrelationship between the source second-order dynamics in the time domain and the input power spectrum in the frequency domain, as well as its overall impact on system performance.

For the queuing analysis in the time domain, the multimedia source is described as a general ON/Off binary process. The considered second-order statistics are: variation coefficient of ON period C_t; variation coefficient of OFF period C_s; correlation coefficient in adjacent ON/OFF period C_{ts}; Correlation coefficient in adjacent OFF/ON period C_{st}; variation coefficient of ON-access rate C_r. The

source is modeled as a general Markov modulated Poisson process. The underlying Markov chain has an arbitrary input chain and it is defined by transition rate matrix Q with its state space partitioned into two subspaces: ON and OFF. If the Poisson arrival rate in each state is r_i, then the MMPP is fully characterized by Q and the input rate processes. The system is considered to have finite buffer size of K, exponential service time and N i.i.d binary sources. The queue is therefore a MMPP/M/1/K queue.

The average queue length, queue length standard deviation and mean loss rate vs the second-order source statistics are studied through numerical analysis, i.e, the individual effect of C_t, C_s, C_{ts}, C_{st}, C_r on the above queue performance and the joint effect of (C_t, C_s) and (C_{ts}, C_{st}) on the same queue. On the other hand, in the frequency domain, the corresponding change of input power spectrum is studied. It is observed that the more the input powers lie in low frequencies, the longer the queue length and the higher the loss rate would be. Therefore, the input power spectrum is recommended as a unified source measurement for multimedia traffic queuing analysis.

An important implication of the above study is in the effective policies for network resource allocation. A link capacity allocation method based on the input power spectrum for a finite buffer system that transmits multimedia traffic is proposed in [11]. The idea is that, since the low frequency band of the input signal introduces the queue delay and loss, a way of avoiding the delay and loss during the transmission is to let the low frequency component travel intact through the network without being buffered, while the high frequency component is to be effectively delivered with the finite buffer capacity subject to no loss. The key question here is how to pick the minimum link capacity.

Intuitively, with no buffer, the link capacity should be assigned to the input peak rate. In the frequency domain, this is equivalent to a filtered input peak rate which is defined as maximum output of a filtered input signal at a cutoff frequency. The simulation study with a sample sequence of coded video data shows that the minimum link rate of a finite buffer system with no cell loss can be assigned by filtered peak input rate at a properly selected cutoff frequency. With this link rate, the low frequency component of the input signal can be guaranteed stay intact while traveling through the network. This gives a means of static link rate allocation. However the cutoff frequency depends on the buffer size, which in turn depends on the queue system subject to no loss. So how to correctly select the cutoff frequency to get the filtered peak rate is still subject to further study.

The dynamic link allocation as a more practical approach is discussed as well. Since the low frequency component has a slow and predictable time variation, it is possible to assign the link rate based on the on-line observation of the filtered input signal.

The use of the frequency domain analysis is not limited by traffic models. The algorithm of the link capacity allocation can be applied to various traffic sources, even the self-similar traffic.

9. Summary

Typical models of different types of traffic in the literature are presented in this paper. One purpose of this paper is to show the regular procedure that is involved in traffic modeling. Another purpose is to provide an overview of the historical development of traffic modeling in the sense of accuracy in capturing the arrival correlation (or rate correlation) and burstiness of traffic. These two characteristics are basic factors that affect the queue performance. The development of the models from the renewal process to Markovian chains and then to the MMPP is an improvement in the sophistication of accurately modeling the correlation of the on-off sources; Video traffic models from ARMA to modulated ARMA and then to TES methods is another example of the progress in describing the correlation and burstiness of the traffic. The self-similar process, in contrast, is a model whose traffic characteristics differ drastically from almost all the conventional models. This will consequently attract more attention to the research in the area of traffic modeling. This provides a different view to the conventional ideas about traffic characteristics. It also implies an important impact on network design and traffic control.

The diverse approaches to traffic modeling are presented in the paper as well. For the purpose of analytical simplification, the fluid flow model is used as an alternative to the models that describe the arrivals of individual data units. It is an attractive model for high speed ATM network traffic. TES methods, on the other hand, use the basic flexible TES process to build a traffic model such that the autocorrelation function and the marginal distribution of the model are matched to the measured data. Moreover, the frequency domain analysis of the traffic sources not only implies a unified method of dealing with the traffic modeling from a different perspective, it also provides an effective way of network resource allocation. Finally, the chaotic map method, though still under study, is expected to be able to complement the various models that have been proposed so far in the literature (models presented in this paper are the typical ones) and is able to provide a compact description of the complicated properties of different types of traffic including the self-similar traffic.

10. References

[1] Harry Heffes, David M. Lucantoni, "Markov Modulated Characterization of Packetized and Data Traffic and Related Statistical Multiplexer Performance", IEEE JSAC vol.sac-4 No.6, Sept 1986, pp.856

[2] Kotikalapudi Sriram, Ward Whitt, "Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data", Intergrated Broadband Networks, pp.172

[3] Raj Jain, Shawn A. Routhier, "Packet Trains-Measurements and a New Model for Computer Network Traffic", Integrated Broadband Networks, pp.157

[4] Reto Gruenenfelder, John P. Cosmas, Sam Manthorpe and Augustine Odinma-Okafor, "Characterzation of Video Codecs as Autoregressive Moving Average Processes and Related Queuing System Performance", IEEE JSAC, vol.9, No.3, April 1991

[5] B. Maglaris, D.Anastassiou, P.Sen, G. Karlsson and J. Robins, "Performance models of statistical multiplexing in packet video communications," IEEE Trans. Commun., vol. 36, No.7, pp.834-844, July 1988

[6] G. Ramamurthy and B. Sengupta, "Modeling and Analysis of A Variable Bit Rate Video Multiplexer", 7th ITC Seminar, Session 8, Oct 1990

[7] Daniel Geist, Benjamin Melamed, "TEStool: An Environment for Visual Interactive Modeling of Autocorrelated Traffic" ICC'92 vol.3, pp.1285

[8] A. Baiocchi, N. Blefari Melazzi, M. Listanti, A. Roveri, R. Winkler, "Modeling Issues on an ATM Multiplexer Within a Bursty Traffic Environment", INFOCOM'91, vol.1, pp.2c.2.1.

[9] San-qi Li, Hong-Dah Sheng "Discrete Queuing Analysis of Multi-Media Traffic with Diversity of Correlation and Burstiness Properties", INFOCOM'91, vol.1, pp.4c.1.1.

[10] Hong-Dah Sheng, San-qi, Li, "Second order effect of binary sources on characteristics of queue and loss rate", IEEE INFOCOM'93, pp 18-27

[11] San-qi Li, Chaia-Lin Hwang, "Queue Response to Input Correlation Functions: Discrete Spectral Analysis", INFOCOM'92, vol.3, pp.0382.

[12] L. Kosten, "Stochastic theory of a multi-entry buffer(II),(III)," Delft Progress Rep, Series F, vol.1, pp.44-50, 1974., vol.1, pp.103-115, 1975

[13] L. S. Liebovitch andk T.I. Toth, "Ion channel kinetics: random or deterministic process:", Biophysical Journal, vol.57, pp.317, 1990

[14] P. Sen, B. Maglaris, N-E. Rikli, and D. Anastassiou,"Models for packet switching of variable bit rate sources,: IEEE JSAC, vol.7, no.5, pp.865-869, June 1989

[15] Ilkka Norros, James W. Roberts, Alain Simonian, and Korma T. Virtamo, " Superposition of Variable Bit Rate Sources in an ATM Multiplexer," IEEE JSAC, vol.9, no.3, April,1991

[16] Will E. Leland, Murad S. Taqqu, Walter Willinger, Daniel V. Wilson, "Ethernet Traffic is Self-Similar: Stochastic Modeling of Packet Traffic Data," to be published

[17] Will E. Leland, Daciel V. Wilson, "High Time-Resolution Measurement and Analysis of LAN Traffic: Implication for LAN Interconnection," INFOCOM'91 vol.3, pp.11d.3.1.

[18] Henry J. Fowler and Will E. Leland, " Local Area Network Traffic Characteristics, with Implications for Broadband Network Congestion Management", IEEE JSAC, vol.9, no.7, Sept 1991

[19] A. Erramilli and R.P. Singh, "The Application of Deterministic Chaotic Maps to Characterize Traffic in Broadband Packet Networks," 7th ITC Seminar, Session 8, Oct 1990

[20] Benjamin Melamed, B. Sengupta,"TES Modeling of Video Traffic," IEICE Trans. Commun., Vol. E75-b, No.12, December 1992

[21] V$\tilde{a}$clav E. Bene$\tilde{s}$, "General Stochastic Process in the Theory of Queues," Addison-Wesley Publishing Company, Inc., 1963

[22] J. Beran, R. Sherman, Murad S. Taqqu, W. Willinger, "Variable-Bit-Rate Video Traffic and Long-Range Dependence", to be published

[23] L. Kosten, "Stochastic theory of a multi-entry buffer(I)", Delft Progress Report Series F, vol.1, pp.10-18, 1974

[24] D. Anick, D. Mitra and M.M. Sondhi, "Stochastic theory of a data handling system with multiple resources," Bell Syst. Tech. J. vol.61, no.*, pp.1871-1894, 1982

[25] R.C. Tucker, "Accurate method for analysis of a packet speech multiplexer with limited delay," IEEE Trans. Commun., vol.36, no.4, Apr. 1988

[26] J.R.M. Hosking, "Fractional differencing," Biometric, vol.68, No.1, pp. 165-176,1981

CONTENTS