J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*
Copyright © 1999 John Wiley & Sons, Inc.

TELEPHONE TRAFFIC

The users and providers of telecommunication services are in the midst of far-reaching technical, operational, and regulatory changes. Voice is no longer transferred on a section of copper cable virtually dedicated to connecting two subscribers. In an attempt to optimize utilization, voice, data, internet connections, multimedia, surfers on the World Wide Web, and many other sources of telecommunication traffic have to share resources. These sources are manipulated, compressed, packetized, and multiplexed into a mixed and complex stream of aggregated information. In many countries around the world, the national telecommunications monopolies either no longer exist or have been reduced to provide so-called basic services within restricted parts of the network (1). In such an open market, the most important criteria for choosing a service provider are price and grade of service (*GOS*). Understanding the nature of telephone traffic and predicting the number of resources that will be necessary to handle the traffic efficiently are important for the provision of quality service and cost-effective networks.

The subject of telephone traffic or teletraffic refers to the study and modeling of traffic sources and circuits, both at an individual level and en masse. Applications include the efficient dimensioning of transmission links, routes, switching systems, and the design of network topologies and protocols. Fundamental quantities arising from this knowledge are factors such as system capacity, congestion levels, and identification of congestion points. The subject of teletraffic theory has a long history dating back to the work of the Danish mathematician Agner Krarup Erlang (1878–1929) (2).

According to the basic definitions of telephone traffic or teletraffic engineering, provided by the *ITU* (International Telecommunications Union), *GOS* is defined as “a number of traffic engineering variables used to provide a measure of adequacy of a group of resources under specified conditions.” The *GOS* variables can be defined as any relevant indicators of quality, such as probability of call blocking, drop call rate, or call set-up delay (3).

As far as traffic handling is concerned, the resources mentioned in the above definition may be any kind of user terminal, its connections to a node, such as a switch, combined with the nodes and links to networks. This implies that both the information-carrying network and the corresponding signaling network should be viewed as resources.

Practical traffic-carrying systems (e.g., telecommunication systems) generally consist of multiple (N) traffic sources that load, or are served by, several (n) devices or *resources*, as illustrated in Fig. 1. Typical sources are subscribers generating telephone calls or automatic call generators such as telemetry systems and data links. Typical *resources* are devices such as lines, trunks, circuits, switches, inputs, outputs, signal receivers, or radio channels.

The load placed on an individual *resource* by any particular source is of a binary nature (either loaded or unloaded) and intermittent, of varying duration, and unrelated to the activity of the other sources.

For economic reasons, in most cases $n < N$, and therefore the possibility of call blocking or congestion arises. The system consequences of $n < N$ provides the motivation and basis for teletraffic theory. A generic system of this type is illustrated in Fig. 1.

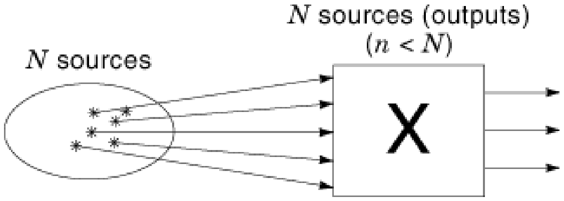


Fig. 1. Generic network with N sources and n resources.

Teletraffic engineering is necessary to predict and determine the adequacy of available resources according to acceptable values of *GOS* variables. There is a complex interaction between the telephone traffic processes and interconnected groups of resources (1). Teletraffic engineering describes all the recourses that are relevant to end-to-end (user-to-user) connections.

As a result of the complexity of telephone traffic processes, the resources are usually studied in selected groups, such as switches or links. For practical reasons, some complex traffic processes, such as call arrival rate, are usually described by means of simplified theoretical methods (4). The effects of failure-free recourses may be assumed or in some cases omitted. These simplifications are only acceptable if the end user still experiences the full complexity of the *GOS*, that is, the *GOS* as it appears in real circumstances.

Service users and service providers usually have different evaluation criteria. The user is mostly interested in network access and end-to-end aspects that are directly experienced by them. The service provider is mainly focused on traffic performance of the network between bearer service access points, or from a development, planning, or operation and maintenance point of view. These would include important issues such as route dimensioning, switch dimensioning, and technology selection.

In simple terms, teletraffic studies are devoted to the following points:

(1) The volume of traffic to be handled by the network resources
(2) The *GOS* that should be maintained
(3) The type and number of resources that are required and how they should be configured to accommodate the permissible traffic
(4) The functional relationship between *GOS* and traffic resources

Many of the mentioned considerations have been described successfully in previous years but modern digital and packet switched systems have introduced new complexities that require new approaches. The following section focuses on different types and sources of telephone traffic.

Sources of Traffic

Voice Traffic.

Variation of Traffic Level. Traffic levels in telecommunication networks are, in general, related to human activities (or the lack thereof). Higher levels of both telephone and data traffic can be expected during weekday business hours; however, also for precisely this reason, some automatic data transmission may be scheduled for off-peak periods. A good understanding of the behavior of the system's users will be of value in both designing the system and devising (marketing) strategies to make use of periods when the system is underloaded.

Sources of Telephone Traffic. The traffic load on a telephone system displays regular variations on differing time scales. This can be explained be studying the behavior of the traffic sources and the system response.

In a well-dimensioned telecommunication system, a time function of the number of busy servers depicts the stochastic variation of carried traffic. This is usually a discontinuous curve with steps of ±1 occurring at irregular intervals, as is shown at a later stage in Fig. 4(b).

Traffic models of this type are not easy to develop and are not useful in practice, as systems dimensioned on instantaneous or peak values are economically impractical. As an alternative, various time averages will be considered.

A sliding one hour average (based on traffic measurements with an interval of several minutes) will produce a smooth curve providing an accurate indication of the busy hour. The standardized method is based on 15 min measurements; the four consecutive readings that produce the highest sum identify the busy hour, with results very similar to that of the sliding one hour method. Care should be taken to use a representative data set, for example, 10 days (excluding weekends and special traffic conditions). It may be noted that both the busy hour and the traffic volume hour change according to the day of the week, reflecting slightly changing subscriber habits as the week progresses.

Factors Affecting Traffic Volumes. Traffic levels give a good macrolevel indication of the activities of subscribers. Figure 2 provides an example of a traffic profile during a day. The traffic profile in Fig. 2 will be affected according to the geographic location of the switch and mix of traffic sources (e.g., a switch based in the central business district will not show significant after-hours traffic levels). Figure 2 shows the following significant traffic events:

(1) Very low traffic levels in the early morning
(2) Rapid rise in traffic up to the start of the commercial day
(3) Midmorning peak (traveling salespeople on the road, etc.)
(4) Lunchtime dip
(5) Late afternoon peak before the close of the day
(6) Traffic jam causing high levels (typical of mobile telephone systems)
(7) Dinner dip
(8) Increase as off-peak tariff encourages domestic calls
(9) Television news or popular program
(10) Decrease as night progresses
(11) Possible increase on selected links as automatic data transfers take place

Exceptions in traffic levels are commonly observed on a daily basis, especially at a local scale, and may be due to factors such as:

- Failure of a section of a network (cellular or fixed) increases load on the remaining operational ones.
- Road traffic jams typically increase load on cellular networks.
- Natural or manmade disasters (e.g., storm or fire) cause traffic to increase provided the network is not damaged.
- Major sports events cause traffic levels to be low during play but to be higher afterward and to last longer especially if the home team wins.
- Major newsworthy events such as wars, political changes, economic collapse, or court cases cause changes.

Traditional holiday periods are reflected in seasonal changes in traffic. Significant geographic changes in traffic occur, especially in cellular telephone networks as customers move from commercial centers to holiday

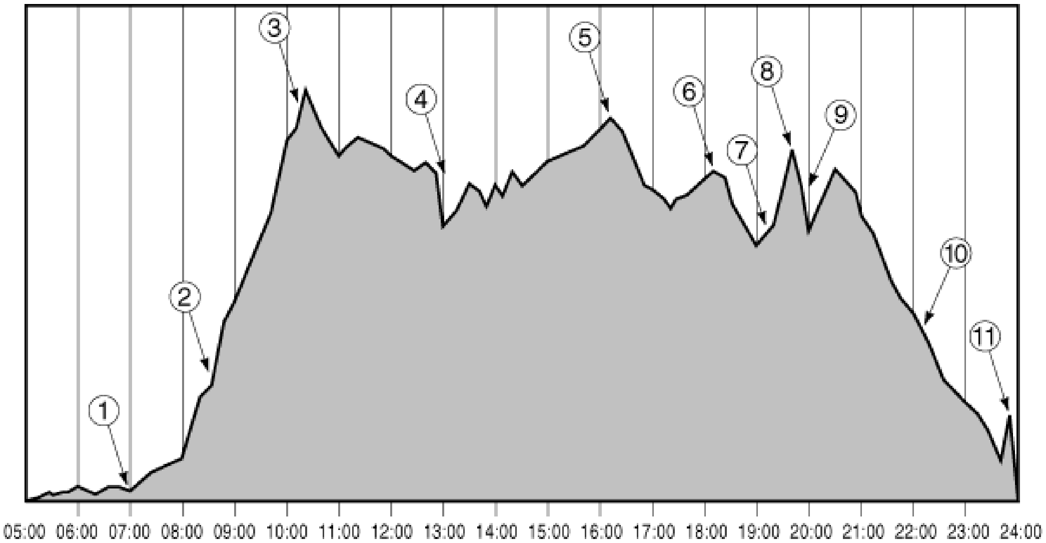


Fig. 2. Typical daily telephone traffic profile.

destinations—in some cases, raising the traffic levels in destination cities above that of the normal nonholiday levels. This effect is greater in smaller towns.

Long-term trends, also known to as "organic growth" of the network, are visible when traffic levels are plotted over longer periods (typically years). Medium-term and short-term trends or "nonorganic growth" may be attributed to the introduction of new services such as internet, cellular networks or regulatory changes, network expansion, (wireless local loop), or simply service quality improvement.

Traffic Per Subscriber. A key parameter in dimensioning telecommunication systems is to determine the amount of traffic that the average subscriber will generate in the busy hour. This is done by estimating the number of calls and their duration that such a subscriber will place during that hour. The busy hour traffic (BHT) is then

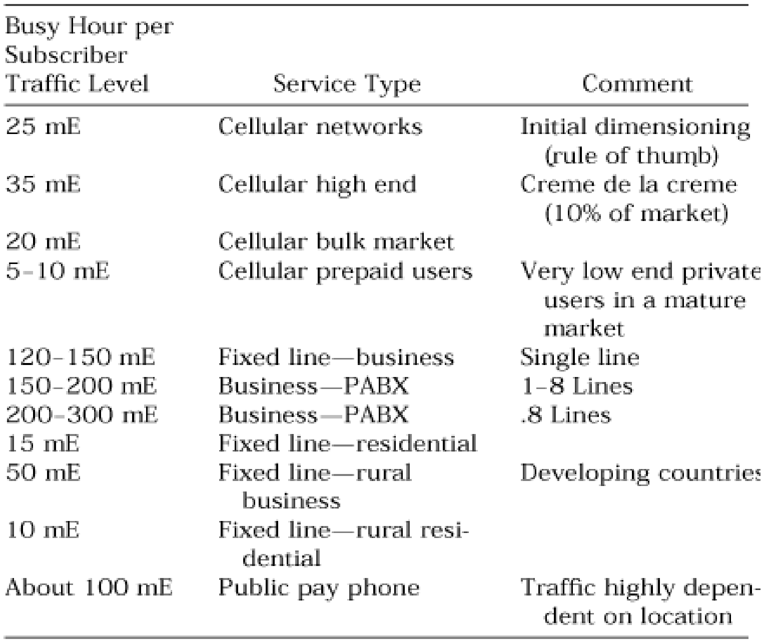
$$\mathrm{BHT} = nT/3600 \qquad (1)$$

where n is the number of calls placed during the hour, and T is the average duration of each call (seconds).

For example, in the initial dimensioning of cellular systems it is typical to use $n = 1$ and $T = 90$ s. This gives a busy hour traffic of 0.025 E (25 mE). The traffic per subscriber value will vary according to the type of service offered, the tariff level, whether calls are business or private, who pays the bill, and the social customs of the country (in some societies it is considered impolite if business is conducted without first making mutual inquiries regarding the well-being of the correspondents and their relatives, or discussing the weather). Table 1 provides some indicative values.

Nontelephone Traffic or Data Traffic. Nontelephone based traffic follows similar profiles and obeys similar traffic laws, although factors such as holding time, data rates, and data symmetry may be very different. New traffic sources such as voice over the Internet, voice over packet-networks (5), and voice over ATM (6) have blurred the dividing line between voice and data networks. Consequently, the investigation of data traffic is very relevant to describe the complex nature of modern heterogeneous network traffic, where voice and many other services reside on a common backbone. Traffic might even change in characteristics as it advances

Table 1. Busy Hour per Subscriber Traffic Levels for Different Telephone Services



Busy Hour per Subscriber Traffic Level	Service Type	Comment
25 mE	Cellular networks	Initial dimensioning (rule of thumb)
35 mE	Cellular high end	Creme de la creme (10% of market)
20 mE	Cellular bulk market	
5–10 mE	Cellular prepaid users	Very low end private users in a mature market
120–150 mE	Fixed line—business	Single line
150–200 mE	Business—PABX	1–8 Lines
200–300 mE	Business—PABX	.8 Lines
15 mE	Fixed line—residential	
50 mE	Fixed line—rural business	Developing countries
10 mE	Fixed line—rural residential	
About 100 mE	Public pay phone	Traffic highly dependent on location

through the different nodes of a network, for example, wireless cellular calls. The following examples illustrate the differences between conventional voice and heterogeneous data sources:

- Facsimile transmissions, although relatively short in duration, produce asymmetric data transfer.
- A dial-up Internet session typically involves a long holding time with a high asymmetry in data rates between the user's computer and the Internet. Most sessions involve large amounts of data being sent to the user (graphics and text) compared to the few, infrequent, navigation key strokes from the user.
- File transfers between computer systems typically involve sustained and high volume unidirectional data flow. Returning data may simply be reception acknowledgments.
- Video transmissions for entertainment require lengthy unidirectional data transmission while a video conference may involve a bidirectional transfer.

The introduction of packet-switched digital communication has resulted in many advantages but has also brought forward new challenges, especially as far as modeling and management of present and future networks are concerned. The initial application of narrow-band services such as *ISDN*s (Integrated Services Digital Networks) proved to be inadequate in respect to digital throughput, especially for high-bandwidth-consuming moving video services, such as video conferencing, video phones, and video-on-demand (6,7,8).

New signaling systems, like Common Channel Signalling System No. 7 (CCSS#7), are considered to be a prerequisite for the efficient national and global management of *ISDN* and future broadband *ISDN* (BISDNs) services (9). CCSS#7 was defined to provide a comprehensive signaling method in order to accommodate digital telephone traffic at multiples of 64 kbit/s. These signaling systems make enough information available to allow further automation of network management. It is expected that there will be a considerable increase in management and signaling traffic as a result of the services associated with CCSS#7, like Mobile Communications and Intelligent Network (*IN*) services such as Freephone, 800 service, credit card calling, and user-to-user signaling (10). As a consequence, more emphasis is being placed on automatic network management techniques built into stored program controlled (*SPC*) switching and signaling systems. Recent advances are accomplished primarily because of the introduction of automated, computer-based, and distributed network management. The current approach attempts to provide an economic balance between automatic and manual network management (11).

Changes in network architecture have also brought forward changes in the characteristics of network traffic. Newly conceptualized broadband technologies such as BISDNs and asynchronous transfer mode (*ATM*) have to accommodate considerably diverse payload types on the same virtual channel within a common packet-switched network, giving rise to increasingly bursty traffic (12). Flexible user to network interfaces (*UNI*s), in the range of up to 600 Mbit/s, will provide the possibility to support connections ranging from low (e.g., slow terminal) to very high bit rates (e.g., moving entertainment video) (13). The statistical multiplexing of *ATM* cells that belong to different virtual connections (*VC*s) give network users the further possibility of services varying in a wide range of bit rates, according to their changing needs during a connection. The changing process dynamics of these new types of variable bit rate (*VBR*) and available bit rate (*ABR*) traffic make it particularly difficult to model and capture traffic characteristics (14,15).

It has been shown that the statistical multiplexing of multiple packet-switched sources do not give rise to a more homogeneous aggregate, but that properties such as burstiness are conserved (15). Figure 3 shows a sample of *VBR* video traffic that serves as an example as to how bursty these types of heterogeneous traffic can be. The traffic could be continuous (e.g., large data transfer), predominantly a monologue (e.g., image transfer), highly bursty (e.g., compressed entertainment video), or might require bidirectional communication (e.g., interactive video conferencing). Broadband services also require diverse performance requirements. Real-time voice, for instance, requires rapid transfer through the network but a small amount of lost data is tolerable. In many alternative instances, real-time delivery is not important but strict error control is of the essence. On the other hand, real-time video communication (Fig. 3) is one of the most challenging cases because it requires error-free as well as rapid transfer (16).

Teletraffic Theory

The processes of teletraffic systems are normally divided into two fundamental subsections—the call origination process and the service process. Furthermore, it is important to realize that the complete description of these two process subsections are essential in order to achieve a methodical characterization of the total teletraffic process. Markovian properties, process stationarity, and exponential distributions (Poisson arrivals) are fundamental requirements that are necessary in the application of the methods reviewed in this section and for conventional teletraffic models in general.

This section provides an overview of the basics of telephone traffic theory. For a more complete discussion and applications of these concepts, the reader is referred to the publications from which this section was largely abstracted (4,17,18,19) and to (20 and 21) for more background on probability theory.

Traffic Volume and Intensity. *Sources* or traffic are typically individual customers that ask for service, usually in an uncoordinated manner. A request for service is called a *call-attempt* (or more generally *occupation attempt*) and, if granted, will occupy the resource as a *call* (or *occupation*). This is illustrated in Fig. 4.

The *traffic volume* carried by a system is usually therefore measured over a period. This will be developed using Fig. 4 as follows.

Let t_{ij} be the *holding time* of the ith of n channels on the jth of m occasions of it being loaded during the period T.

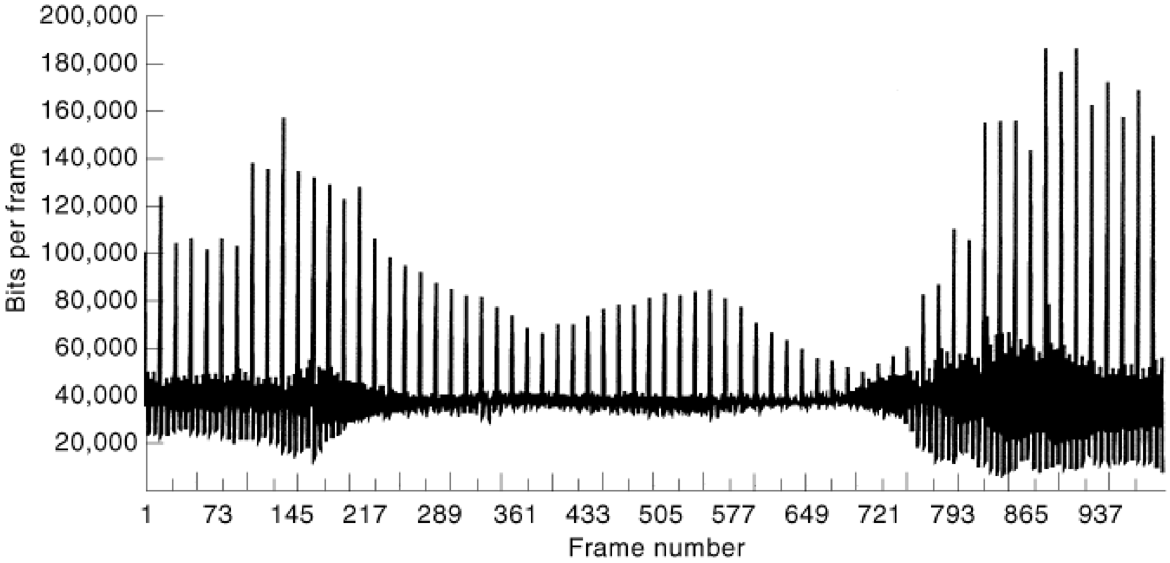


Fig. 3. Bursty traffic sample from a variable bit rate video source shows how bursty these types of heterogeneous traffic can be.

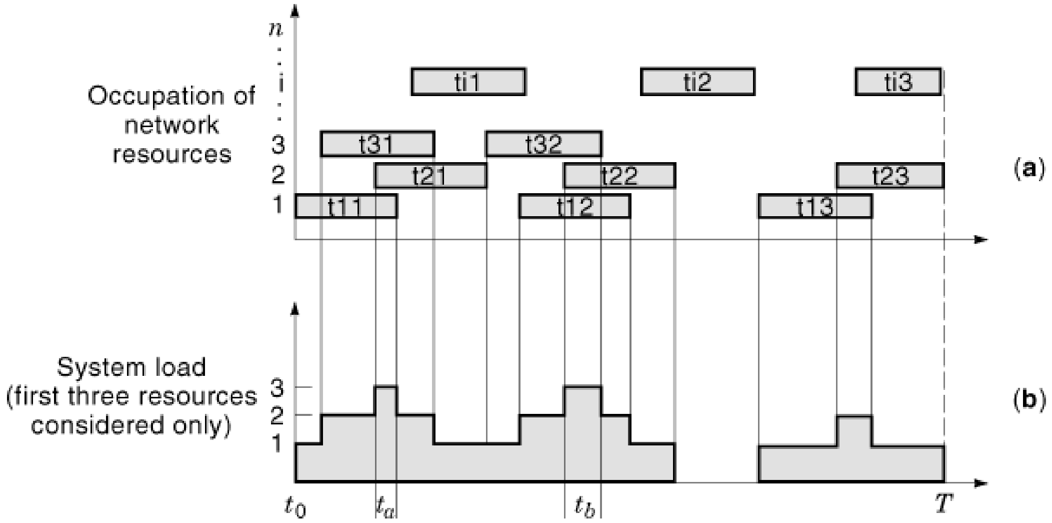

n
i
3
2
1
Occupation of network resources
ti1
ti2
ti3
t31
t32
t21
t22
t23
t11
t12
t13
(a)
System load (first three resources considered only)
3
2
1
(b)
t_0
t_a
t_b
T

Fig. 4. (a) Occupation of network resources and (b) corresponding system load over measurement period T .

The traffic volume on the ith channel is defined as the sum of the times for which the resource is occupied over the period, that is,

$$(\text{Traffic volume})_i = V_i = \sum_{j=1}^{m} t_{ij} \qquad (2)$$

The total traffic volume of the system is then the sum of the traffic of each channel:

$$\text{Total system traffic volume} = V_\text{s} = \sum_{i=1}^{n} B_j = \sum_{i=1}^{n}\sum_{j=1}^{m} t_{ij} \quad (3)$$

Alternatively, the total system traffic volume can be given by the area under the discontinuous curve of the system load shown in Fig. 4(b).

$$V_\text{s} = \int_T r(t)\,dt \quad (4)$$

where *r(t)* is the number of busy servers (i.e., the instantaneous traffic level).

A more important term is *traffic flow* or *traffic intensity*. This can be expressed in three ways.

First, this is given by the total system traffic volume divided by the duration of the specified period:

$$\text{A} = \text{Traffic intensity } V_\text{s}/T = \frac{1}{T}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} t_{ij}\right) \quad (5)$$

This is a dimensionless quantity, but because of its importance, the term “erlang” has been assigned as the unit of traffic. This means that if a single resource is occupied continuously or intermittently for a total time of t during a period of T, then it carries t/T erlang; its maximum possible load (*traffic capacity*) being 1 erlang. Thus the maximum traffic capacity of a system in erlang equals the number of resources in the system. This could be interpreted as one call-hour per hour or one call-minute per minute.

Second, let t_x be the sum of the time for which exactly x out of the N resources are occupied simultaneously within a period T. For example, in Fig. 4 we get $T_3 = t_a + t_b$. Then,

$$\sum_{x=0}^{n} t_x = T \quad (6)$$

The sum of the holding times of all resources is

$$\sum_{x=1}^{n} x t_x \quad (7)$$

This is the same as the total system traffic defined in Eq. (3) above, so the traffic intensity is given as

$$A = \sum_{x=1}^{n} x t_x \quad (8)$$

Since t_x/T is the proportion of time for which x devices are held simultaneously, the right-hand side of this equation gives *the average number of resources held simultaneously during the specified period* as the second definition of traffic intensity.

Third, if T is so long that the effect of unexpired calls at the beginning and end of the period is negligible, the traffic intensity is approximately.

$$A \cong (CT)h/T = Ch \text{ erlang} \qquad (9)$$

where C is the average number of resource occupations per unit time and H is the average holding time per occupation. Hence the traffic intensity is *approximately equal to the average number of occupations occurring during a period equal to one average holding time* (22).

It should be noted that the definition of traffic does not refer directly to data capacity of the system (e.g., in terms of the amounts of data transmitted). A *resource* capable of transmitting 9600 bits/s and one of 64 kbit/s are each said to be carrying 1 erlang if occupied, although the 64 kbit/s recourse would have carried a greater volume of data.

Call Origination Process. There are various patterns of call origination, but the most fundamental is random origination and is modeled as follows (with λ the arrival or origination rate and $\Delta t \to 0$):

- The probability that a single call originates in time interval $(t,\ t + \Delta t]$ tends to $\lambda\Delta t$ independent of the value of t, where λ stays constant during the time interval.
- The probability that two or more calls originate in $(t,\ t + \Delta t]$ tends to zero.
- Calls originate independently of each other.

Within these limitations the probability, $p_k(t)$, that k calls originate in the time interval $(0,\ t]$ can be calculated according to Eq. (10) (4, 19):

$$p_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \qquad (10)$$

Equation (10) describes the *Poisson distribution* with mean λt and is fundamental to teletraffic theory (4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19). The fact that λ is constant and independent of time is a feature of the random origination assumption, and consequently this model is also referred to as the Poissonian arrival process.

Since the mean number of calls originated in $(0,\ t]$ is λt, λ is interpreted as the mean number of arrivals in unit time. This leads to the fact that the arrival rate depends on the choice of unit time. This property is indirectly exploited in the fractal approach (23) mentioned in the section entitled “Alternative Modeling Methods” to overcome the limitations of coarsely sampled time scales.

From Eq. (10) it follows that the probability that no calls originate in $(0,\ t]$ is given by

$$p_0(t) = e^{-\lambda t} \qquad (11)$$

and consequently the distribution function of the interarrival time (probability that the interarrival time is no greater than t) is given by

$$A(t) = 1 - e^{-\lambda t} \qquad (12)$$

which is called the *exponential distribution of interarrival time* with mean λ^{-1} and is also a feature of the assumption of random origination.

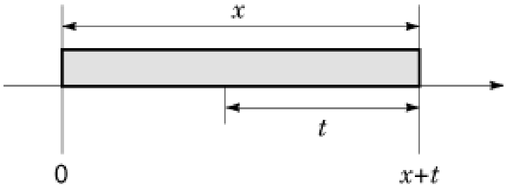


Fig. 5. Markov property (memoryless).

Service Distribution Process. The simplest case in the description of the service time distribution is the assumption that a call is terminated at random (19). It follows from the random termination assumption that the probability that a call is terminated in time interval $(t, t + \Delta t]$ is $\mu\Delta t$ and is independent of t, with μ the service time or termination rate. The complementary distribution function *H(t)* (probability that the service time is greater than t) is the same as the probability that a call is terminated in $(0, t]$.

If the time interval $(0, t]$ is partitioned into a sufficiently large number n of subsections, in the same manner as noted in the previous section, then the probability *H(t)* is given by (with $\Delta t = t/n$):

$$H(t) = (1 - \mu\Delta t) \tag{13}$$

$$H(t) = \lim_{n\to\infty}\left(1 - \frac{\mu t}{n}\right)^n = e^{-\mu t} \tag{14}$$

Thus the service time is exponentially distributed with mean μ^{-1}.

From Eqs. (10) and (14)) it is evident that the basic assumptions describing the arrival and service processes are based in some way on exponential functions. The assumption of exponentially distributed service times agrees fairly well with conventional circuit-switched telephone service times (19). Primarily because of the consequent simplicity of analysis, this assumption has widely been implemented in traffic theory (4,17,19). Unfortunately, this simplification has led to some discrepancies in the description of modern packet-switched processes (14,15,23,24,25).

Fundamental Relations.

Markov Property. A stochastic process is called a Markov process if it exhibits the Markov or memoryless property: that is, the stochastic behavior of the process in the future is dependent only on the present state and is independent of past progress (19).

To explain this property in more detail, consider the duration time X of a phenomenon, say, service time, as shown in Fig. 5. If X is exponentially distributed with mean μ^{-1}, then the probability that the phenomenon continues after time instant x, is given by

$$P\{X > x\} = e^{-\mu x} \tag{15}$$

Furthermore, the conditional probability that the phenomenon continues, after initial occurrence at time x, for duration period t, is calculated as

$$\begin{aligned} P\{X > x + t | X > x\} &= \frac{P\{X > x + t\}}{P\{X > x\}} = \frac{e^{-\mu(x+t)}}{e^{-\mu x}} \\ &= e^{-\mu t} = P\{X > t\} \end{aligned} \tag{16}$$

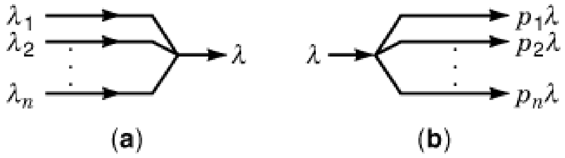


Fig. 6. (a) Aggregation and (b) decomposition of a Poisson stream.

Now it can be noted that the probability given by Eq. (16) is independent of x. This implies that the stochastic behavior of the phenomenon after (future) time x is only dependent on the state at time x (present) and independent of the progress before (the past) time x.

It is important to realize that only the exponential distribution exhibits this property for continuous distributions and that this assumption is fundamental in many teletraffic models.

Aggregation and Decomposition of Poissonian Streams and The Law of Large Numbers. In order to discuss the principles relevant to traffic stream *aggregation* and *decomposition*, it is first important to note the so-called PASTA (Poisson arrivals see time average) principle (19). This principle specifies that in the exceptional case were system interarrival times are exponential (Poisson arrivals), then it follows that

$$\Pi_j = P_j \tag{17}$$

where P_j is the probability that j calls exist at an arbitrary instant in the steady state of a system, and Π_j is the corresponding probability just prior to the call arrival epoch.

The PASTA property leads to a very important condition as far as the aggregation and decomposition of Poisson streams are concerned (19). If n independent Poisson streams with rates $\lambda_j, j = 1,2, \dots, n$, as shown in Fig. 6(a), are aggregated, then the resultant stream again becomes a Poisson stream, with rate $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$. This is the result of the fact that the convolution of Poisson distributions is again Poissonian.

Conversely, if a Poisson stream with rate λ is directed to route j with probability p_j, Fig. 6(b), then the stream in j again becomes Poissonian.

The aggregation and decomposition properties are very useful for the analysis of systems with Poissonian inputs. The mentioned properties are exploited when a large number of independent sources are aggregated and the *law of large numbers* may be assumed to apply (19). The law of large numbers is not a completely analytical method but is of some importance in the characterization of telephone traffic. As an example, this "law" states that if a large number of bursty sources are aggregated, then the burstiness of the resultant stream diminishes and therefore would significantly simplify the analysis process (4,19). Unfortunately, this property has been shown to be absent in modern packet-switched traffic and, for instance, burstiness to be conserved when sources are aggregated (15,23,26).

Little's Formula. The principles associated with the aggregation and decomposition of traffic streams lead us to one of the most fundamental relations in the field of telephone traffic. Previously, we defined λ as the arrival or origination rate. In simple terms, λ could be seen as the number of calls that arrive during a certain time span and have to be serviced by the system within an acceptable service interval, h. Little's formula states that the number of calls, L, that will have to wait in a queue to be serviced is given by Eq. (18):

$$L = \lambda h \tag{18}$$

Little's formula applies only for mean values and requires that the processes described are stationary in time. Note that there are no restrictions on the statistical distribution of the variables.

Non-Markovian and Advanced Conventional Teletraffic Models. Although most conventional models are Markovian in nature, there is a rich family of non-Markovian methods that do not completely rely on the Markovian property and also incorporate a number of nonexponential arrival or service distributions. According to the Kendall notation (19), we define the following:

M Exponential (conventional Markovian—most common) and some non-Markovian distributions
E_k Phase kErlangian (convolution of k exponentials with identical mean)
H_a Order n hyperexponential (alternative of n exponentials)
D Deterministic (fixed, e.g., *ATM*)
G General (arbitrary—e.g., defined by mean and variance)
GI General independent (renewal process)

And some “advanced” teletraffic distributions such as:

Phase-type Markov renewal process
Markov modulated Poisson process

As can be seen from the above examples, in practice most non-Markovian distributions still maintain some connection with the Markovian and exponential assumptions in order to limit analytical complexity. Indeed, in most cases a non-Markovian model is defined as a system where either the arrival or the service distribution is still assumed to be exponential (17,19). Here we recognize well-known models such as Poisson input general service time (M/G/1(m)), Poisson input constant service time especially applicable to *ATM* networks, M/D/1), renewal input exponential server (GI/M/s), and many others such as GI/G/1, H_2/G/1, and E_k/G/1 .

Advanced Teletraffic Models. The increasing complexity of network traffic characteristics and the possibility to combine diverse and heterogeneous traffic on common virtual circuits within a network, such as ISDN and *ATM* services, have prompted more interest in multiclass input models and more complete and advanced teletraffic models. Under the scope of multiclass input models we recognize methods such as batch arrival models ($M^{[X]}$/m/s(0), $M^{[X]}$/G/1, $GI^{[X]}$/G/1, etc., priority models, multidimensional models (trunk systems), mixed loss and delay systems ($M_1 + M_2$/M/s $(0, \infty)$; GI + M/M/s$(\infty, 0)$; GI + M/M/s$(0, \infty)$, etc.), and multiqueue models (4,17,19).

In the field of so-called more advanced teletraffic models there are numerous approaches, such as the renewal input multiserver model with the diffusion approximation, the phase-type Markov renewal process (PH-MRP inputs), and the MMPP input models (MMPP/G/1, MMPP/D/1) (4,19).

The MMPP method has been very successfully applied to the characterization of packet-switched traffic networks such as ISDN and *ATM* when simplistic exponential assumptions fail. The MMPP/D/1 model is accepted as a state of the art approximation of the arrival rate and fixed packet length transfer processes of *ATM* traffic and networks (4,19). Especially in applications where the statistical multiplexing of packets is modeled, the MMPP method (MMPP+M/G//1) is considered to provide an acceptable solution (19,27).

Markov Modulated Poisson Process (MMPP). The Markov modulated Poisson process is a doubly stochastic Poisson process with arrival rates depending on the phases or states, which are dictated by a continuous Markov chain. The MMPP method is a special case of the PH-MRP and because of its tractability it is widely used for the modeling of bursty traffic such as packetized voice and video in *ATM* networks (6,19). For a more complete discussion and analysis of the MMPP method, the reader may want to refer to the following excellent works (4,14,19).

The simplest case of an MMPP is shown in Fig. 7, where two phases of Poisson arrival rates are given by $\lambda_j, j = 1, 2$, and appear alternatively with exponentially distributed lifetimes with means $r^{-1}{}_j$. This is described

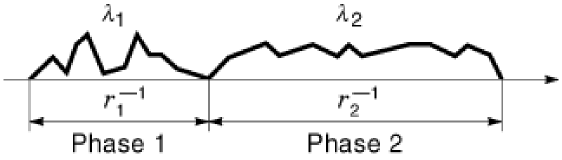


Fig. 7. Two-phase MMPP.

by (R, Λ), where R is the infinitesimal generator matrix (transition probabilities) of the underlying Markov chain and Λ is the arrival rate matrix, given as

$$R = \begin{bmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \tag{19}$$

The n-phase MMPP is similarly characterized by (R, Λ) with each matrix of $n \times n$ size. Although the matrix analysis (14) is valid for general phase MMPPs, the two-phase MMPP is most often used because of tractability and complexity considerations. Significant improvements in the accuracy of characterization can be obtained by using a three-phase MMPP, but with an accompanying increase in complexity (19).

It should be stated clearly that an MMPP-based model can achieve very high accuracy in the description of an arrival rate distribution. However, accuracy usually depends on the number of states used to create the Markov chain controlling the intensities of the Poissonian arrivals. As a result of the number of variables, precise tuning and optimization of an MMPP model can be difficult and sometimes unstable, with local minima easily reached.

Call Blocking. Calls must have some destination and a number of possible options exist: nonaddressed (broadcast), multiaddressed (several addresses), single address multiresource, or single address single resource. The term *tele* in teletraffic implies some distance aspect. This together with the destination options mentioned implies the need for some switching capability in such systems. Traffic studies are therefore typically carried out at well-defined interfaces between the nodes of such systems.

Although most connections are full duplex (information flowing in both directions), the connection is usually directional, that is, from party A to party B. In practice, traffic usually diminishes in moving from the A-side to the B-side due to delays along the path, leading to diminishing holding times and calls being aborted along the path for a variety of reasons (e.g., congestion).

$$\begin{aligned} \mathrm{A}_{\text{source}} &> A_{\text{transmission link}} > A_{\text{switch}} > A_{\text{transmission link}} \\ &> A_{\text{switch}} > A_{\text{transmission link}} > A_{\text{destination}} \end{aligned}$$

A good understanding of the characteristics of the traffic source will allow efficient dimensioning of the transmission links and switches in such a system to ensure that the total system responds to the required levels of service.

If the response to the system is very poor (highly congested), there may be a significant difference between call intent and call demand (indicating a “hidden” traffic demand). The upgrading of such a system is typically followed by an increase in traffic, as users respond to the improved quality.

Call blocking can be defined as the probability that not enough resources are available in the telephone network to allow a new connection to be set up while still maintaining the quality of existing connections. The loss rate or blocking probability is often utilized as a measure of the grade of service (*GOS*) of a network and can be defined as the ratio between the number of calls that could not be carried and the offered traffic. If we denote the offered Erlang traffic by E and the carried traffic by E_c then the blocking probability is given by Eq. (20):

$$B = \frac{E - E_c}{E} \tag{20}$$

Blocking has a direct impact on the system performance and service quality of a telephone network. There are mainly three schools of thought as far as the acceptable levels of call blocking are concerned. The first approach states that no blocking should be tolerated in a network because this results in lost revenue. The second approach will tolerate any level of blocking up to a certain maximum level. The third school of thought states that blocking levels should be maintained within a certain acceptable band. The upper limit of the band provides an acceptable *GOS* while the lower limit assures that the available resources are fully utilized.

The acceptable levels of blocking are usually based on the offered Erlang traffic as described by equations such as Erlang B, Poisson, and Erlang C. These formulas are based on certain fundamental assumptions:

(1) Calls are assigned to the available channels at random.
(2) The number of sources are significantly large when compared to the number of channels.
(3) In the event that a channel is not available to service a call, the call is either lost to the system or is put in a queue to be serviced at a later stage.

Erlang B Loss Formula. *Assumption: Lost Calls Cleared* (*LCC*)

Erlang B is the normal industry standard and is usually applied to determine the trigger points to act upon. As is given in Eq. (21), Erlang B estimates the statistical probability that all the available circuits will be busy when a new call is attempted.

$$B = \frac{E^N/N!}{\sum\limits_{n=0}^{N} E^n/n!} \tag{21}$$

where

B = the probability that a call will be lost or blocked because of insufficient channels
E = the offered Erlang traffic
N = the total number of channels or servers in the group

The intensity of carried traffic, E_c, is then given by Eq. (22):

$$E_c = (1 - B)E \tag{22}$$

The relationship between E and E_c is shown in Fig. 8 for various values of N. For small loads the carried traffic is equal to the offered traffic, $E_c = E$. As the load increases, some percentage of calls will experience blocking. It can also be observed that the greater the number of servers the better the system utilization. For $N = 8$, calls experience 2% blocking at levels of E greater than 2 erlang, which equates to a system utilization of 25%; while for $N = 64$, the system experiences only 2% blocking at 39 erlang or about 60% utilization.

Erlang B relies on the assumption that if a call could not be assigned a circuit during the period of consideration, it is permanently lost to the network (*LCC*). As a result of the assumption that a call is lost if it is blocked, the Erlang B approach has been shown to yield a pessimistic estimation of the *GOS* (the observed *GOS* will be better). There are numerous reviews and technical papers on the statistical nature and accuracy of

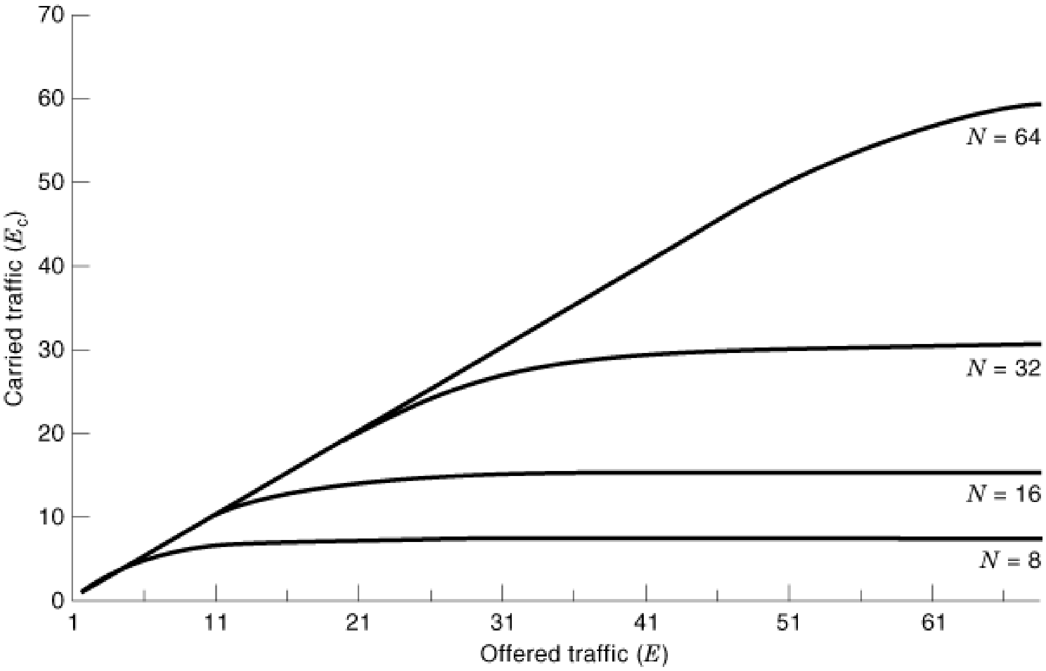


Fig. 8. Relationship between carried and offered traffic.

the Erlang B description of blocking, but it still remains the accepted industry standard for traffic calculations and growth prediction (4,19,23,28).

Poisson and Erlang C Formulas. *Assumption: Lost Calls Held* (*LCH*)

Lost Calls Delayed (*LCD*)

Other familiar approaches such as Poisson and Erlang C assume that blocked calls are not lost but merely queued for a certain time until network resources become available to service the attempted connection. The Poisson approach assumes that calls are held in the system for a period not to exceed the average holding time (*LCH*). The Poisson formula also assumes that call retries will be distributed according to a negative exponential or Poissonian distribution, as given by Eq. (23):

$$W = e^{-E} \sum_{n=c}^{\infty} \frac{E^n}{n!} \tag{23}$$

where

W = the probability that a call will be lost or have to wait because of insufficient channels
E = the offered Erlang traffic
c = the total number of servers or channels,

The Erlang C formula also defines the probability that a call has to wait until the network resources become available to service it. Calls not immediately satisfied are held or delayed by the system until a channel becomes available (*LCD*). An incoming call will only have to wait if the available channels are less than the number of arriving calls. It is assumed that the caller continues to demand service until a channel becomes available and in this way enters a service queue. If we keep the PASTA property in mind, it follows that the

Erlang C formula is given by Eq. (24):

$$W = \frac{\dfrac{E^c}{c!}\dfrac{c}{c-E}}{\displaystyle\sum_{n=0}^{c-1}\dfrac{E^n}{n!}\dfrac{c}{c-E}} \tag{24}$$

where

W = the waiting probability
c = the total number of servers or channels
E = the offered Erlang traffic

Engset Method. *Assumption: Finite Traffic Sources*

This approach differs from the previously mentioned methods in the sense that a limited number of sources are assumed. Furthermore, the Engset method assumes that call arrivals are random and that holding times can be represented by exponential functions.

This method assumes that if calls could not immediately be assigned resources they are lost to the system (*LCC*). The Engset approach, given by Eq. (25), is a variation of Erlang B for a finite number of sources. It is sometimes referred to as the Engset distribution because it finds its roots in a truncated form of the binomial distribution.

$$B = \frac{\frac{(c-1)!}{N!(c-1-N)!}\left[\frac{E}{c-E(1-B)}\right]^N}{\displaystyle\sum_{i=0}^{N}\frac{(c-1)!}{i!(c-1-i)!}\left[\frac{E}{c-E(1-B)}\right]^i} \tag{25}$$

where

B = the probability of calls being lost or blocked because of insufficient channels
c = the total number of servers or channels
E = the offered Erlang traffic
N = the total number of channels in the group

The Engset formula involves extensive calculation and yields values that approximate Erlang B, except in the nonblocking case, where the number of channels exceeds the number of sources.

Alternative Modeling Methods. Modern networks and especially packet-switched traffic exhibit characteristics that are vastly different from conventional circuit-switched voice traffic. The introduction of packet-oriented and cell-based technologies such as Ethernet, ISDN, BISDN, CCSS # 7, and *ATM* have already drastically changed the traffic characteristics of the modern telecommunications network. These types of traffic are highly bursty, less homogeneous, more nonlinear, and more nonstationary compared to conventional telephone traffic (15,29,30). Simply stated, the differences between the characteristics of so-called smooth and rough traffic streams (31) and the consequent difficulties in network dimensioning motivate the utilization of alternative methods to describe modern network traffic.

A number of classical models, based on the fundamentals highlighted in the previous sections, have been proposed to describe traffic in packet-switched networks (4,19). The well-known Markov process has been utilized, with great success, in queuing theory for many years. Although this approach and related models are well established, especially in the realm of circuit-switched traffic, the rapid changes in network traffic characteristics have prompted an influx of publications (15,24 25,26) that highlight the limitations of Markovian and exponential assumptions. These methods and assumptions include Poissonian arrival distributions and

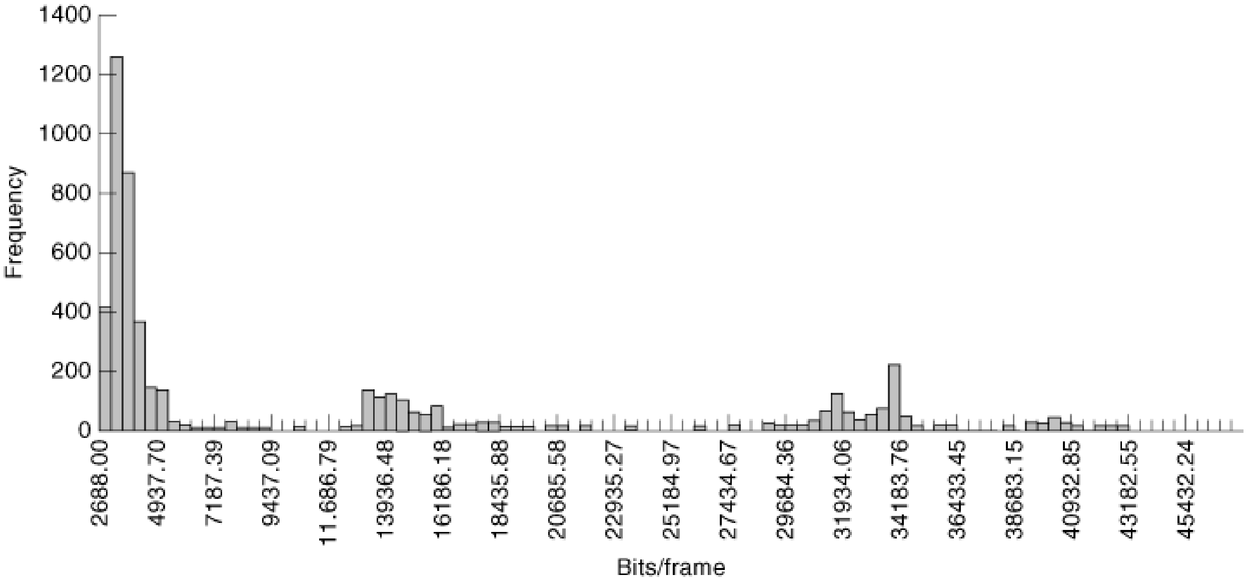


Fig. 9. Traffic clustering of a *VBR* video conferencing source.

related methods such as the M/D/1 queue (17,19), Markov modulated poisson process (MMPP) (19,32), packet train models, and fluid flow models.

Packetized traffic and especially *VBR* sources are difficult to model with conventional analytic math models and Markovian approaches because of the bursty and nonlinear nature of these processes (12). In the construction of analytical models, some assumptions, such as exponential decays, uniform sources, or process stationarity, must be made in order to simplify the mathematical description of the traffic characteristic (4,19).

These assumptions usually lead to degraded generality and robustness of the models (4). (For a truly random process to be stationary, the statistical characteristics of the sample functions must not change with time. Similarly, for a process to be ergodic, the time average of a single sample function must be the same as for the whole ensemble (21).

The mentioned assumptions and methods unfortunately disregard the complex process structure of future broadband packetized services, and as a result, many of the upper layers of the arrival process are neglected (4). This state of affairs is particularly accentuated if the traffic is analyzed in the cell and burst layers (4,5). Figure 9 shows how *VBR* traffic from a video conference session is clustered into distinctly separate traffic intensity layers that are distributed over the complete spectrum of the arrival process.

Previously, it was thought that conventional assumptions approach the observed values when the so-called law of large numbers applies. In other words, if a large number of virtually identical independent sources feed into a common node (19), the consequence will be that the nonlinearity and burstiness of the aggregate process are reduced. In modern networks, this assumption has been shown not to apply in many cases and burstiness is conserved when sources are aggregated. This observation is mostly a result of the relatively high levels of traffic burstiness and the low levels of source aggregation (23) (low number of aggregated sources).

Several researchers have recently produced significant results that show traffic patterns, from diverse packet-switched networks and services, exhibit the presence of properties such as self-similarity, long-range dependencies (26), slowly decaying variances, “heavy-tailed” or power law distributions (24), and fractal structures. The networks and services most frequently mentioned include ISDN packet networks, Ethernet LANs,

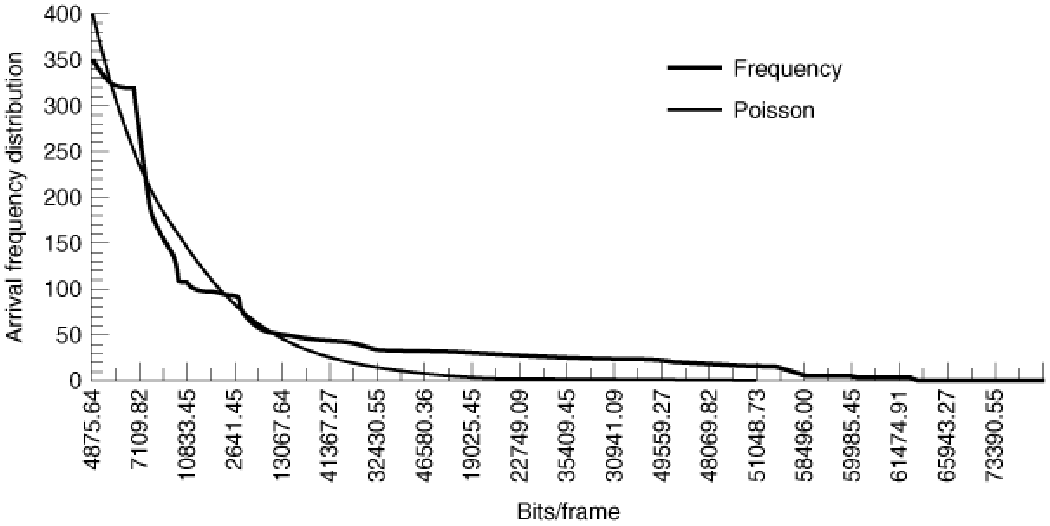


Fig. 10. Heavy-tailed arrival distribution.

and CCSS#7 networks (15,24). These results accentuate the discrepancies between classical assumptions and the characteristics of newly emerging technologies.

Figure 10 reiterates the mentioned "heavy-tailed" distributions and shows a Poissonian fit to the arrival frequency distribution of a sampled *VBR* video source. Heavy-tailed distributions such as these are problematic in that they decay slower than predicted by conventional exponential assumptions. As a result, models based on these assumptions can significantly underestimate network traffic descriptors, such as cell loss probability (15). The heavy tail of packet-switched traffic is primarily a result of high bandwidth consuming traffic that occurs relatively seldom, that is, highly bursty. Unfortunately, bursty high bandwidth sources produce the type of traffic that is problematic to describe with conventional Markovian approaches. Indeed, the results of the above-mentioned studies have shown that none of the commonly used traffic models are able to capture the bursty behavior of cell-based networks, and that this situation may have serious implications in network design, analysis, and management (15,23).

Promising alternative traffic modeling methods that have been mentioned in the relevant literature are listed in Table 2. Most of these methods have not been developed to their full potential and provide stimulating fields for further investigation and research.

Traffic Management and Enforcement Techniques

Introduction. In general, network management can be defined as a real-time surveillance and control activity, to optimize the *GOS* of a network under stress due to traffic overload or failure (11). The ITU-T proposed a number of traffic control algorithms in Recommendation E.412 (4). These methods can loosely be divided into two general groups that address (1) link congestion and (2) node congestion. Trunk protection is suggested as the main method to handle link congestion but counteracting node congestion requires more advanced methods, such as extension and restriction of alternative routes, code blocking, and call gapping.

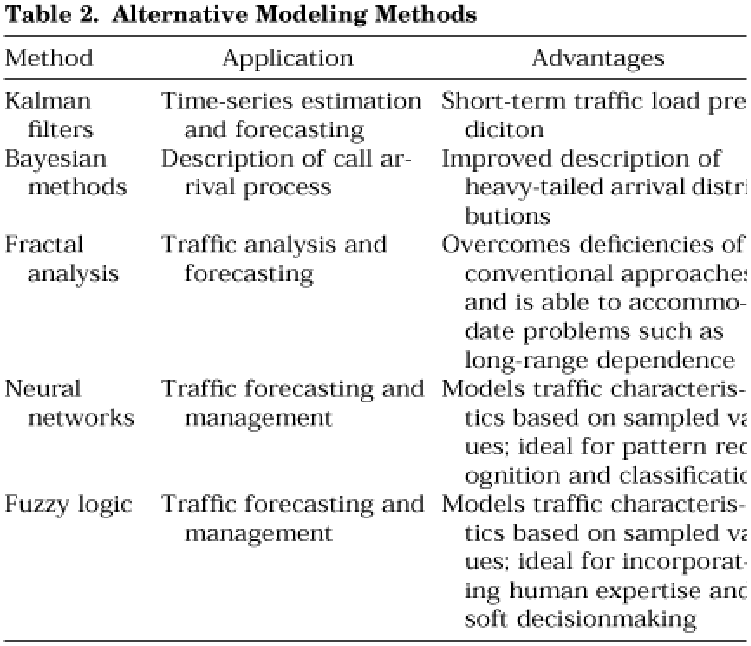


Table 2. Alternative Modeling Methods

Method	Application	Advantages
Kalman filters	Time-series estimation and forecasting	Short-term traffic load prediciton
Bayesian methods	Description of call arrival process	Improved description of heavy-tailed arrival distributions
Fractal analysis	Traffic analysis and forecasting	Overcomes deficiencies of conventional approaches and is able to accommodate problems such as long-range dependence
Neural networks	Traffic forecasting and management	Models traffic characteristics based on sampled va ues; ideal for pattern recognition and classificatic
Fuzzy logic	Traffic forecasting and management	Models traffic characteristics based on sampled va ues; ideal for incorporating human expertise anc soft decisionmaking

The method of alternative route extension and restriction utilizes an automatic rerouting algorithm to extend the alternative route set. In the event that a high blocking condition is detected, the intensity of the overflow traffic is compared to the residual capacities of lightly loaded alternative routes. Up to eight alternative routes can be determined and should the proposed route be able to accommodate the additional traffic, the routing hierarchy is extended to include the relevant nodes. Restricting calls is only allowed in extreme circumstances. Newly established calls are rejected first along with calls that are destined for distant nodes. Traffic intended for adjacent nodes is restricted next and lastly transit traffic is throttled.

Code blocking is used in the case of focused or localized network overloads to block calls that originate from or are destined for overloaded areas. Calls are blocked based on their probability to be completed successfully and in this way network resources are made available for effective calls.

Lastly, the call gapping procedure monitors the number of calls in a certain time window. If the number of calls that arrive in this time window rises above a predefined threshold, it indicates that congestion is imminent and as a result new arrivals are rejected.

Alternative Network Management Methods. The problems of characterizing modern bursty traffic sources have led to some concerns in the field of network management, especially if the stringent *GOS* standards required for emerging networks are considered, for example, cell loss of 1 in 10^{10} for *ATM* networks (6). With this in mind, it is essential to fully utilize the greater flexibility in traffic management that modern electronic line-switching exchanges and network switches allow (16,33). Static alternative routing with “nailed” or fixed alternative routes is common in classical telephone networks. Modern *SPC* (stored program control) switches make it possible to implement dynamic or adaptive routing. This approach uses measurements of carried traffic and network state to determine the most optimum alternative routes dynamically. Advanced routing methods like these alter the characteristic of traffic streams and tend to smooth out traffic bursts, which results in decreased blocking levels (22).

Increased bandwidth flexibility and variable traffic rates, which are associated with future telephone networks, require extended procedures for call establishment and monitoring. Acceptable parameter values, such as mean and peak cell rate as well as *GOS* required, will have to be negotiated during call setup and defined in a "traffic contract" between the user and the network provider (12). A network function called call admission control (*CAC*) is responsible for deciding whether a new call can be introduced into the network. The decision must be made in real-time and must take into account the characteristics of a new call, as well as the requirements of the calls that already exist and will have to share network resources. *CAC* methods are necessary to fully utilize statistical multiplexing principles in order that more calls could coexist on the same network without infringing on the requirements of existing calls. A *CAC* mechanism relies on the parameters as defined in the negotiation phase during call establishment. Thus these parameters must be enforced during the duration of the connection to ensure that there is no violation of the initial agreement.

Newly emerging multimedia and broadband transfer methods, such as *ATM*, make provision for packetized information transfer without flow control between the user and the network. Due to this property and the utilization of statistical multiplexing, it is possible for a call to exceed the negotiated rate up to the maximum capacity as limited by the user to network interface (*UNI*). With this kept in mind, it is necessary that the agreement between the network user and service provider be upheld by a function called usage parameter control or "traffic policing" (6,12,34). This function restricts the traffic during an active connection in order to ensure that the information stream will abide by the negotiated contract. Several methods have been proposed such as the leaky bucket (*LB*) mechanism (6,35), jumping window (*JW*) (36), triggered jumping window (*TJW*) (37), exponentially weighted moving average (*EWMA*), and moving window (*MW*).

It is quite a challenge to cope effectively with the conflicting demands of an ideal policing mechanism that requires a low false alarm probability but also high responsiveness (12,18). Although the mentioned methods have been accepted by the telecommunications industry, they are all limited in the fact that statistical variables, such as traffic mean, are controlled and manipulated by thresholds defined by "crisp" values (38). This type of deterministic decisionmaking is a limitation in an environment, such as exists in modern packet-switched networks, where the process is ill-defined, highly bursty, and nonlinear (34,39).

Recently, there has been noticeable interest in the application of artificial intelligence (*AI*) and soft decisionmaking in the field of packet-orientated networks (34,38,40). An operational telecommunications network typically requires $>10^6$ lines of conventional code to function properly. Software of this scale is usually difficult to fully understand and debug: so much so that there are now increasingly more suboptimal network and system solutions (40). In contrast, we are surrounded by organisms (e.g., ants) that are capable of performing complex cooperative tasks based on a very small knowledge base. AI systems and methods provide the capability to extract the inherent characteristics of a system and in this way make it possible to compress large knowledge bases into more comprehensible and manageable decisionmaking sets of rules, typically < 1000 lines of code (39,40).

Especially in the field of network management, modern AI methods that rely on the ability to utilize soft decisionmaking, such as fuzzy logic, have become more popular. Most of the concepts that humans use in daily life, such as large, small, heavy, and light, are vague and ill defined or "fuzzy" (39,41). Fuzzy logic offers a methodology, with a firm mathematical basis, to accommodate vagueness and qualitative, inexact, imprecise information (42). The fuzzy decisionmaking process is not limited by crisp threshold values (False $= 0$, True $= 1$) but may include the full range of values in between thresholds [0,1] and in this way is much closer to human decisionmaking (39,42). In other words, this approach is a possible method to describe nonlinear and bursty process characteristics, like those encountered in modern packet-switched networks, and to be utilized in applications such as process modeling, time-series estimation, and soft decisionmaking schemes, in order to achieve the ultimate goal of improved traffic management.

Summary

Service quality is one of the most important discriminators in the modern telecommunications network. Consequently, understanding the nature of telephone traffic and the dimensioning of resources are important elements to provide improved *GOS*. Teletraffic engineering describes all the resources that are involved in user-to-user network connections. As a result of the complexity of telephone traffic processes, the related variables are usually defined by means of simplified theoretical methods. In simplified terms, teletraffic studies are devoted to maintaining the ideal relationship between the required *GOS* and the network resources.

Sources of modern and developing network traffic are heterogeneous in nature and consist of multimedia voice and data traffic. The nature of conventional circuit-switched voice traffic is affected by various periodic functions that are mostly based on hourly, daily, weekly, monthly, and yearly observations. To the contrary, modern broadband traffic exhibits vastly different characteristics than circuit-switched voice traffic and is highly bursty and nonlinear.

Telephone traffic or traffic intensity is a measure of the load that active calls put on the network. Although traffic intensity is a nonphysical and dimensionless descriptor, it was agreed to assign the erlang as its fundamental measure. The most basic subsections of teletraffic theory are the call origination and service processes.

Markovian properties, process stationarity, and exponential distributions are essential assumptions to derive conventional teletraffic models and to achieve a methodical characterization of the total teletraffic process. The so-called conventional and Markovian methods have been successfully utilized for many years to describe voice and circuit-switched telephone traffic, but recently there has been an avalanche of publications stating that many of the upper layers of modern and packet-switched network processes are neglected by these approaches. These types of traffic are especially difficult to describe with conventional Markovian methods because of the bursty and nonlinear nature of these technologies. Heavy-tailed arrival distributions that are associated with many of the modern traffic sources reiterate the need for new and alternative approaches to traffic modeling. Some possible alternative methods, intended for traffic modeling and management, are Kalman filters, Bayesian methods, fractal analysis, neural networks, and fuzzy logic.

Increased bandwidth flexibility, associated with future networks, requires extended procedures for call establishment and monitoring. Call admission control has been defined as the network function that is responsible for the decision whether a new call could be introduced into the network, without infringing on other existing calls. Newly emerging multimedia and broadband networks have prompted investigation into soft decisionmaking methods to accommodate the nature of modern network traffic.

It seems that as far as the description of modern telephone and network traffic is concerned, we have just started to understand the complexities involved in this process and many new challenges have still to be faced.

BIBLIOGRAPHY

1. G. Gosztony CCITT work in teletraffic engineering, *IEEE J. Sel. Areas Commun.*, **9**: 131–134, 1991.
2. E. Blockmeyer H. L. Halstrom A. Jensen *The Life and Works of A. K. Erlang*, Copenhagen: Academy of Technical Sciences, 1948.
3. ITU-T, *Recommendation E.600*, Geneva: ITU, 1989.
4. J. Filipiak *Real Time Network Management*, Amsterdam: North-Holland, 1991.
5. S. Li J. W. Mark Traffic characterization for integrated services networks, *IEEE Trans. Commun.*, **38**: 1231–1243, 1990.
6. M. de Prycker *Asynchronous Transfer Mode, Solution for B-ISDN*, Chichester, England: Ellis Horwood, 1993.
7. R. Griffiths P. Key *Adaptive Call Admission Control in ATM Networks*, ITC 14, Amsterdam: Elsevier, 1994, pp. 1065-1076.

8. P. Skelly M. Schwartz S. Dixit A histogram-based model of video traffic behavior in an *ATM* multiplexer, *IEEE/ACM Trans. Netw.* **1**: 446–459, 1993.
9. J. Ronayne *The Integrated Services Digital Network: From Concept to Application*, London: Pitman, 1987.
10. ITU-T, Recommendations Q. 1200, *Intelligent Network*, Geneva: ITU, 1993.
11. D. G. Haenschke D. A. Kettler E. Oberer Network management and congestion in the U.S. telecommunications network, *IEEE Trans. Commun.*, **COM-29**: 376–385, 1981.
12. E. P. Rathgeb Modeling and performance comparison of policing mechanisms for *ATM* networks, *IEEE J. Sel. Areas Commun.*, **9**: 325–334, 1991.
13. N. G. Bean *Robust Connection Acceptance Control in ATM Networks with Complete Source Information*, Stat. Lab. Rep. 93–1, Cambridge, England: University of Cambridge, 1993.
14. M. F. Neuts *Structured Stochastic Matrices of M/ G/ 1 Type and Their Applications*, New York: Dekker, 1989.
15. W. E. Leland *et al.* On the self-similar nature of ethernet traffic (extended version), *IEEE/ACM Trans. Netw.*, **2**: 1–15, 1994.
16. J. J. Bae T. Suda Survey of traffic control schemes and protocols in *ATM* networks, *Proc. IEEE*, **79**: 170–189, 1991.
17. D. Bertsekas R. Gallager *Data Networks*, Englewood Cliffs, NJ: Prentice-Hall, 1987.
18. M. F. Scheffer J. S. Kunicki Fuzzy adaptive traffic enforcement for *ATM* networks, *IEEE Melecon '96*, Bari, Italy, pp. 1047–1050, 1996.
19. H. Akimaru K. Kawashima *Telletraffic: Theory and Applications*, New York: Springer-Verlag, 1993.
20. A. O. Allen *Probability, Statistics, and Queueing Theory with Computer Science Applications*, San Diego, CA: Academic Press, 1990.
21. F. C. Stremler *Communication Systems*, Reading MA: Addison Wesley, 1990.
22. D. Bear Principles of telecommunication engineering, *IEE Telecommun. Ser. 2*, **2**: 1988.
23. W. Willinger *et al.* Self-similarity through high-variability: Statistical analysis of ethernet LAN traffic at the source level, *Proc. ACM/SIGCOMM '95*, 1995.
24. K. Meier-Hellstern *et al. Traffic Models for ISDN Data Users: Office Automation Application*, ITC-13, Copenhagen, 1991.
25. D. E. Duffy *et al.* Statistical analysis of CCSN/SS7 traffic data from working CCS subnetworks, *IEEE J. Sel. Areas Commun.*, **12**: 544–551, 1994.
26. J. Beran *et al.* Long-range dependence in variable-bit-rate video traffic, *IEEE Trans. Commun.*, **43**: 1566–1579, 1995.
27. J. Filipiak Accuracy of traffic modeling in fast packet switching, *IEEE Trans. Commun.*, **40**: 835–846, 1992.
28. U. Black *Data Networks: Concepts, Theory, and Practice*, Englewood Cliffs, NJ: Prentice-Hall, 1989.
29. M. F. Scheffer J. S. Kunicki Comparative analysis of modeling techniques for packetized data, *ITC '95*, St. Petersburg, 1995.
30. M. F. Scheffer *et al.* Improved modeling techniques for packetized ISDN traffic, *RITS '95* Pretoria, 1995.
31. E. Szybicki A. E. Bean Advanced traffic routing in local telephone networks: Performance and proposed routing algorithms, *ITC'9*, 1979.
32. H. Heffes D. M. Lucantoni A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, *IEEE J. Sel. Areas Commun.*, **SAC-4**: 856-868, 1986.
33. B. Jabbari Common channel signaling system number 7 for ISDN and intelligent networks, *Proc. IEEE*, **79**: 155–169, 1991.
34. V. Catania *et al.* A fuzzy expert system for usage parameter control in *ATM* networks, *Proc. GLOBECOM'95*, Singapore, 1995.
35. A. Eckberg D. Luan D. Lucantoni Bandwidth management: A congestion control strategy for broadband packet networks—Characterizing the throughput-burstiness filter, *ITC Sect. Semin. Adelaide '89*, Adelaide, Paper 4.4, 1989.
36. F. Vand Den Dool *Policing and Load Control and Related Functions*, RACE R1022, DNL-311&321–008-CD-CC, 1988.
37. F. Denissen E. Desmet G. H. Petit The policing function in an *ATM* network, *Proc. 1990 Int. Zurich Semin. Digit. Commun.*, Zurich, pp. 131–144, 1990.
38. L. A. Zadeh A rational for fuzzy control, *Trans. ASME, J. Dyn. Syst. Meas. Control*, **94**: 3–4, 1972.
39. B. Kosko *Neural Networks and Fuzzy Systems*, Englewood Cliffs, NJ: Prentice-Hall, 1992.
40. P. Cochrane D. J. T. Heatley *Modelling Future Communications Systems*, London: Chapman & Hall, 1996.

41. P. Chemouil J. Khalfet M. Lebourges A fuzzy control approach for adaptive traffic routing, *IEEE Commun. Mag.* **33**(7): 70–76, 1995.
42. C. C. Lee Fuzzy logic in control systems: Fuzzy logic controller. Parts 1 and 2, *IEEE Trans. Man Cybern.*, **20**: 404–435, 1990.
43. R. I. Wilkenson Theories for toll traffic engineering in the USA, *Bell Syst. Tech. J.* **35**: 421–454, 1956.

MARTEN F. SCHEFFER
Motorola South Africa
HILTON GOODHEAD
Mobile Telephone Networks