3.3. Traffic models and teletraffic dimensioning

Sándor Molnár, author

Béla Frajka: reviewer

3.3.1. Introduction

The basic teletraffic principles, equations and an overview of the nature of network traffic are given in Chapter 1.7. Based on these preliminaries this chapter presents the most important traffic models, which are possible candidates to capture the main characteristics of network traffic. Models range from very simple to very complex. In practice, it is a compromise how complex a model we chose to capture more accurately traffic characteristics but also keeping mathematical tractability and computability as simple as possible. With a chosen model applied in a queueing context the aim is generally to find the performance measures in the investigated scenario.

We also provide a survey about the dimensioning principles in both classical telephony and also in data networks with the Internet in the focus.

3.3.2. Traffic models

Traffic consists of single or multiple arrivals of discrete entities (packets, cells, etc.). It can be mathematically described by a *point process*. There are two characterizations of point processes: by the *counting processes* or the *interarrival time processes* [3.3.1]. The counting process $\{N(t)\}_{t=0}^{\infty}$ is a continuous-time, non-negative integer-valued stochastic process, where $N(t) = \max\{n : T_n \le t\}$ is the number of arrivals in the interval (0,*t*]. The interarrival time process is a real-valued random sequence $\{A_n\}_{n=1}^{\infty}$, where $A_n = T_n - T_{n-1}$ is the length of the time interval between the n^{th} arrival from the previous one. The traffic is called compound traffic in case of batch arrivals. In order to characterize compound traffic the batch arrival process $\{B_n\}_{n=1}^{\infty}$ is defined, where B_n is the number of units in the batch. Another useful notion is the *workload process* $\{W_n\}_{n=1}^{\infty}$. It is described by the amount of work *W*$_n$ brought to the system by the *n*th arrival.

In the following a number of traffic models are described that can be used to generate traffic characterized by sequences of $\{N(t)\}_{t=0}^{\infty}, \{A_n\}_{n=1}^{\infty}, \{B_n\}_{n=1}^{\infty}$ or $\{W_n\}_{n=1}^{\infty}$.

In a **renewal process** $\{A_n\}_{n=1}^{\infty}$ are independent, identically distributed with general distribution [3.3.1], [3.3.2]. This model is simple but non-realistic in many cases because it is not able to capture the strong correlation structure present in most of the actual data traffic.

The **Poisson process** [3.3.1], [3.3.2] is a renewal process whose interarrival times $\{A_n\}_{n=1}^{\infty}$ are exponentially distributed with rate parameter λ. The definition can also be given by the counting process, where $\{N(t)\}_{t=0}^{\infty}$has independent and stationary increments with Poissonian marginals, i.e. $P\{N(t) = n\} = \exp(-\lambda t)(\lambda t)^n / n!$ Poisson processes are very frequently used in teletraffic theory due to their simplicity and several elegant properties. The voice call arrivals in telephony are typically modeled by Poisson processes.

Bernoulli processes [3.3.1], [3.3.2] are the discrete-time analogs of Poisson processes. In this model the probability of an arrival in any time-slot is *p*, independent of any other one. The number of arrivals in *k* time-slot is binomially distributed, i.e., $P\{N(t) = n\} = \binom{k}{n} p^n (1-p)^{k-n}$ and the times between arrivals are geometrically distributed, i.e., $P\{A_n = j\} = p(1-p)^j$

Phase-type renewal processes [3.3.1], [3.3.2] compose a special class of renewal processes having *phase-type distributed* interarrival times. It is an important class because these models are analytically tractable and, on the other hand, any distribution can be arbitrarily approximated by phase-type distributions.

Markov-based models [3.3.1], [3.3.2] introduce dependence into the random sequence *A*$_n$. The construction of the model is the following. Consider a Markov process $M = \{M(t)\}_{t=0}^{\infty}$ with a discrete state space. *M* behaves as follows: it stays in state *i* for an exponentially distributed holding time with parameter λ$_i$ which depends on *i* alone. Then it jumps to state *j* with probability *p*$_{ij}$. Each jump of this Markov process is interpreted as signaling an arrival so the interarrival times are exponential. This is the simplest *Markov traffic model*.

Markov renewal processes [3.3.1], [3.3.2] are more general than simple Markov processes but they can still be handled analytically. A Markov renewal process $R = \{(M_n, \tau_n)\}_{n=0}^{\infty}$ is defined by the Markov chain $\{M_n\}$ and its associated inter-jump times $\{\tau_n\}$, subject to the following constraints: the distribution of the pair (M_{n+1}, τ_{n+1}), of next state and inter-jump time, depends only on the current state M_n, but not on previous states nor on previous inter-jump times. In this model arrivals can also be interpreted when jumps occur.

Markov arrival processes (MAP) [3.3.1], [3.3.2], [3.3.3] constitute a broad subclass of Markov renewal processes. In MAP interarrival times are phase-type and arrivals occur at the absorption instants of the auxiliary Markov process. Moreover, the process is restarted with a distribution depending on the transient state from which the absorption had just occurred. MAP is still analytically usable and it is a very versatile process for modeling purposes.

In a **Markov-modulated process** a Markov process is evolving in time and its current state controls the probability law of traffic arrivals [3.3.1], [3.3.2]. Consider a continuous-time Markov process $M = \{M(t)\}_{t=0}^{\infty}$ with state space of *1,2,…m*. While *M* is in state *k*, the probability law of traffic arrivals is completely determined by *k*. When *M* goes to another state, say, state *j*, then a new probability law of traffic arrivals takes effect for the duration of state *j*, and so on. In other words, the probability law of traffic arrivals is modulated by the state of *M*. These stochastic processes sometimes also called *double stochastic processes*. The modulating process can also be much more complicated than a Markov process but such models are less tractable analytically.

The **Markov Modulated Poisson Process** (MMPP) [3.3.1], [3.3.2], [3.3.3] is the most commonly used Markov-modulated traffic model. In this model when the modulating Markov process is in state *k* of *M* then arrivals occur according to a Poisson process at rate λ_k . The simplest case of MMPP is the two-state MMPP model when one state is associated to an “ON” state with a specific Poisson rate, and the other is an “OFF” state with associated rate zero. This model is also called as *interrupted Poisson process*. Such models are used for modeling voice traffic sources with the ON state corresponds to talk spurt and OFF state corresponds to silence period.

In the **Markovian transition-modulated processes** [3.3.1], [3.3.2] the transition of the Markov process $M = \{M(t)\}_{t=0}^{\infty}$ is the modulating agent rather than the state of *M*. State transitions can be described by a pair of states: the one before transition and the one after it. The number of arrivals B_n in slot *n* is completely determined by the transition of the modulating chain given by $P\{B_n = k | M_n = i, M_{n+1} = j\} = t_{ij}(k)$, which is independent of any past state information.

In the **Generally Modulated Deterministic Process** (GMDP) [3.3.3] the source can be any of the possible *N* states. While it is in state *j* the traffic generated at a constant rate $\lambda_{j.}$ The time spent in state *j* can be described by a generally distribution but in most of the cases it is assumed geometric in order to keep analytical tractability. If you consider a two state GMDP where one of them has zero generation rate we have the slotted-time version of the ON/OFF model.

In the **fluid traffic modeling** technique the traffic is considered as a fluid instead of individual traffic units [3.3.1], [3.3.2], [3.3.3]. This is a good model where the individual traffic units (e.g. packets) are numerous relative to a chosen time scale. The advantage of this technique is the simplicity compared to traffic models that are aimed to capture the structures of individual traffic units. The simplest types of fluid models are assuming two states: an ON state when traffic arrives deterministically at a constant rate λ, and an OFF state when there is no traffic carried. In order to keep analytical tractability the durations of ON and OFF periods are typically assumed to be exponentially distributed and mutually independent. In other words, they form an alternating renewal process.

Autoregressive traffic models define the next variate in the sequence as an explicit function of the previous variates within a time window stretching from the present to the past [3.3.1], [3.3.2], [3.3.3]. Typical examples of these models are the *linear autoregressive (AR) processes*, the *moving average (MA) processes*, the *autoregressive moving average (ARMA) processes* and the *autoregressive integrated moving average (ARIMA) processes*. These models were found to be useful to characterize VBR video traffic.

The **Transform-Expand-Sample** (TES) [3.3.1], [3.3.2] approach aims to construct a model satisfying three requirements: marginal distributions should match its empirical counterpart, autocorrelation should approximate its empirical counterparts up to a reasonable lag and the sample paths generated by the model should "resemble" the empirical time series. TES models can be used e.g. for constructing MPEG video models.

Fractional Gaussian Noise (FGN) [3.3.1] is an exactly second-order self-similar process with self-similarity parameter *H*, provided *½<H<1*. It is a stationary Gaussian process, $X=\{X_k\}_{k=1}^{\infty}$, with autocorrelation function of the form $\rho_X(k)=\frac{1}{2}(|k+1|^{2H}-2|k|^{2H}+|k-1|^{2H}), k\geq 1$. FGN is also long-range dependent (LRD) with parameter *H*: $\rho_X(k)\approx H(2H-1)|k|^{2H-2}, k\to\infty$. FGN can be a candidate traffic model for characterizing aggregated LRD traffic at backbone links.

The **fractional ARIMA** (FARIMA) [3.3.1] model is based on the classical ARIMA(*p,q,d*) model but the parameter *d* used with the difference operator is allowed to take fractional values. FARIMA models are more flexible than FGN models to capture LRD traffic because they can also be tuned to capture the short-range dependent (SRD) characteristics as well.

The **M/Pareto model** is a Poisson process with rate λ of Pareto distributed overlapping bursts [3.3.4]. During the burst the arrival process is constant with rate *r*. The burst length period has a Pareto distributions with parameters $1<\gamma<2, \delta>0$: $P\{X>x\}=\begin{cases}\left(\frac{x}{\delta}\right)^{-\gamma}, x\geq\delta \\ 1 \quad ,\text{otherwise}\end{cases}$. This model generates LRD traffic with parameter $H=(3-\gamma)/2$, so it is also a good candidate to model fractal traffic.

Based on the various traffic models outlined above one can apply these models or combination of some of them to model specific application traffic. These models are the **applied traffic models to specific applications**. Here we present a guideline for possible modeling alternatives of some popular applications [3.3.5].

Application	Model	Distribution
TELNET	session interarrival times	exponential
	session duration	lognormal
	packet interarrival times	Pareto
	packet size	mostly 1 byte packets
FTP	session interarrival times	exponential
	number of items	empirical
	item size	log-Normal
CBR voice	session interarrival times	exponential
	session duration	exponential
	packet interarrival times	constant
	packet size	constant
VBR video teleconferencing	frame interarrival times	constant
	frame size	Gamma
MPEG video	frame interarrival times	constant
	scene length	Geometric
	frame size	lognormal
WWW	request interarrival times	exponential
	document size	Pareto

Table 3.3.1. Different models describing sevices

3.3.3. Teletraffic dimensioning of classical telephone networks

Teletraffic theory was fundamental to the design of classical telephone networks from the beginning. By assuming a stationary Poisson call arrival process the traffic and performance relationship can be expressed by the well-known Erlang

$$B = E(a,n) = \frac{a^n / n!}{\sum_{i=0}^{n} a^i / i!}$$

loss formula which gives the probability of call blocking *B*, when a certain volume of traffic, *a*, is offered to a given number of circuits, *n*:

It expresses that the blocking probability is a simple measure of the offered traffic. Note that blocking probabilities are insensitive to other details about the nature of traffic such as distribution of call holding time. (The formula is valid for an *M/G/n/n* queueing system.) This famous formula was intensively used during the history of teletraffic theory. Because telephone calls are initiated by independent individuals making independent decisions, random models assumed to be stationary within a busy hour are appropriate for engineering purposes. Since these calls are all point-to-point with a fixed bandwidth the Erlang formula was an excellent guide for network engineering.

Important refinements of this formula were also developed to different networking scenarios but the Erlang loss formula (and also the related Erlang delay formula) are still extensively used by teletraffic engineers in their daily work. No doubt that this formula got the highest success among all the results of teletraffic theory.

Besides the Erlang loss and delay formulae a number of techniques have been developed for telephone networks. These are, for example, the equivalent random method based on the work of Wilkinson, the different descriptions of traffic burstiness by peakedness and indices of dispersion, and models like the Engset model, etc. [3.3.11].

3.3.4. Teletraffic dimensioning of the Internet

We are just in the phase to see the birth of the teletraffic theory of the Internet. Currently network provisioning is based on some rules of thumb and teletraffic theory has no major impact in the design of the Internet. As we discussed in chapter 3 the nature of the data traffic is significantly different from the nature of voice traffic and no general laws can be found as it was in the case of voice traffic. New techniques and models are expected to develop in teletraffic theory of the Internet to cope with these challenges. In the following we review the most possible two alternatives of Internet teletraffic engineering. The one is called “big bandwidth philosophy”, the other is called “managed bandwidth philosophy”.

3.3.4.1 The big bandwidth philosophy

There is a certain belief that there is no real need for some advanced teletraffic engineering in the Internet because the overprovisioning of resources can solve the problems. This is the “*big bandwidth philosophy*”. People from this school say that in spite of the dramatic increase of Internet traffic volume in each year the capacity of links and also the switching and routing devices will be so cheap that overprovisioning of resources will be possible. It is worth investigating a little bit more deeply how realistic the “big bandwidth philosophy” is.

They expect that the transmission and information technology can follow the “Internet traffic doubling each year” trend [3.3.10] and can provide cheap solutions. From a technological point of view it seems that this expectation is not unrealistic at least in the near future. Indeed, if you imagine today’s Internet and you just increase the capacity of links you could have a network which supports even real-time communications without any QoS architectures. The current best effort type Internet could do it!

On the other hand, the locality of data in the future will also be dominant, which makes caching an important technical issue of future networks [3.3.10]. Even today if you want to transmit all the bits that are stored on hard drives it would take over 20 years. This trend probably gives a relative decrease in the total volume of transmitted information.

Another important factor is that the streaming traffic, which really requires some QoS support is not dominant in the Internet [3.3.10]. It was believed that it would become dominant but none of these expectations have been fulfilled so far, which can also be predicted for the future. The demand of this traffic type is not growing as fast as the capacity is increasing. Consider the following example: we have 1% streaming traffic so it needs some QoS support. We have two options. We can introduce some QoS architecture or we can increase the capacity by 5%. The people from the “big bandwidth philosophy” school argue that the second one is cheaper. They also say that multimedia applications will use *store-and-reply* technique instead of real-time streaming. They argue that the capacity of magnetic storage is increasing at about the same rate as transmission capacity. Moreover, due to transmission bottlenecks (e.g. wireless link) it makes sense to store information in local.

It is also interesting if we investigate the reason of capacity increase in the previous years. For example, we can see that people are not paying for cable modems or ADSL because their modem links could not bring them more data, but because when they click on a hyperlink they want that page on their screen immediately! So they need the big capacity not for downloading lots of bits but rather achieving a *low latency* when they initiate a file download. This is also the reason for the fact that the average utilization of LANs have been decreased by about a factor of 10 over the last decade: people wanted high bandwidth to provide low latency!

Will overprovisioning be the solution? Nobody knows at this time. It is rather difficult to predict what will happen mainly because this is not only a technical issue but rather depends on political and economic factors. However, as a modest prediction we might say that even if overprovisioning can be a solution for backbone networks it is less likely that it will happen also in access networks. For cases where overprovisioning cannot be applied we have a limited capacity which should be managed somehow. This leads us to the second alternative which is the *“managed bandwidth philosophy”*.

3.3.4.2 The managed bandwidth philosophy

In the case of limited network resources some kind of traffic control should be implemented to provide appropriate capacity and router memory for each traffic class or stream to fulfill its QoS requirements. Basically, there are three major groups of QoS requirements: transparency, accessibility and throughput [3.3.7]. *Transparency* expresses the time and semantic integrity of transferred data. As an example for data transfer semantic integrity is usually required but delay is not so important. *Accessibility* measures the probability of admission refusal and also the delay for set up in case of blocking. As an example the blocking probability is in this class, which is a well-known and frequently used measure in telephone networks. The *throughput* is the main QoS measure in data networks. As an example in today Internet a throughput of 100Kbit/s can ensure the transfer of most of the web pages quasi-instantaneously (less than one second).

Considering the traffic types by nature two main groups can be identified: stream traffic and elastic traffic [3.3.7]. The *stream traffic* is composed of flows characterized by their intrinsic duration and rate. Typical examples of stream traffic are the audio and video real-time applications: telephone, interactive video services, and videoconferencing. The time integrity of stream traffic must be preserved. The negligible loss, delay and jitter are the generally required QoS measures.

The *elastic traffic* usually consists of digital objects (documents) transferred from one place to another. The traffic is elastic because the flow rate can vary due to external causes (e.g. free capacity). Typical elastic applications are the web, e-mail or file transfers. In case of elastic traffic the semantic integrity must be preserved. Elastic traffic can be characterized by the arrival process of requests and the distribution of object sizes. The throughput and the response time are the typical QoS measures in this class.

In the following two subsections we overview the main principles of managing stream and elastic traffic, respectively.

3.3.4.3 The open-loop control of stream traffic

The stream traffic is usually controlled by an *open-loop preventive traffic control* based on the notion of traffic contract [3.3.7]. Traffic contract is a successful negotiation between the user and the network in which user requests are described by a set of traffic parameters and required QoS parameters. Based on these requests the network performs an admission control accepting the communication and the traffic contract only if QoS requirements can be satisfied.

The effectiveness of this control highly depends on how accurately the performance can be predicted based on the *traffic descriptors* [3.3.6]. From the practice it turned out that it is not simple to define practically useful traffic descriptors. It is because it should be *simple* (understandable by the user), *useful* (for resource allocation) and *controllable* (verifiable by the network). Results of intensive research on finding such traffic descriptors with all these properties showed that it is practically impossible. As an example the standardized *token bucket type descriptors* (both in ATM and Internet research bodies) are good controllable descriptors but they are less useful for resource allocation. The users are encouraged to use mechanisms (e.g. traffic shaping) to ensure declared traffic descriptors. Mechanisms can also be implemented at the network ingress to police traffic descriptors (traffic policing). Both shaping and policing are frequently based on the mentioned token bucket type mechanisms.

The major types of open-loop traffic control (admission control) strategies depend on whether statistical multiplexing gain is aimed to be utilized and to what extent [3.3.7]. The following table shows the main categories:

Approach	Buffer sharing	Bandwidth sharing
peak rate allocation	NO	NO
Rate envelope multiplexing	NO	YES
Rate sharing	YES	YES

If no multiplexing gain is targeted to achieve we have the simplest case and we can simply allocate the maximal rate (peak rate) of all the connections, which is called *peak rate allocation*. The advantage of this approach is that the only traffic descriptor is the peak rate of the connection. The admission control is very simple: it only has to check that the sum of required peak rates is over the total capacity or not. The main disadvantage of peak rate allocation is the waste of resources because statistically it is only a small fraction of the time when all the connections actually transmit traffic at the peak rate.

If we design to share the bandwidth but not to share the buffer among connections, we have the *rate envelope multiplexing* case [3.3.6], [3.3.7], [3.3.8]. This approach also called *bufferless multiplexing* because in the fluid modeling framework of this method no need for a buffer. Indeed, in rate envelope multiplexing the target is that the total input rate is maintained below the capacity. The events of exceeding the capacity should be preserved below a certain probability. i.e., $P(\Lambda_t > c) < \varepsilon$, where Λ_t is the input rate process, *c* is the link capacity and ε is the allowed probability of exceeding the capacity. In actual realizations buffers always needed to store packets which arrive simultaneously (cell scale congestion). All the excess traffic is lost, the overall loss rate is $E[(\Lambda_t - c)^+ / E(\Lambda_t)]$. The loss rate only depends on the stationary distribution of Λ_t and not on its time dependent properties. It is important because it means that the correlation structure has no effect on the loss rate. Therefore the very difficult task of capturing traffic correlations (e.g. long-range dependence) is not needed. The traffic structure can have impact on other performance measures but these can be neglected if the loss rate is small enough. For example, LRD traffic can yield to longer duration of the overloads than SRD traffic but using a small loss rate it can be neglected in practice. The main disadvantage of rate envelope multiplexing is that the utilization is still not very good.

If we want to further increase the link utilization we have to share the buffer as well, see Figure 3.3.1. This is the *rate sharing* method [3.3.6], [3.3.7], [3.3.8] or also called *buffered multiplexing*. The idea here is that by providing a buffer we can absorb some input rate excess. The excess of the queue length in the buffer at some level should be preserved below a certain probability, i.e., $P(Q > q) < \varepsilon$, where *q* is the targeted queue length level, *Q* is the actual queue length and ε is the allowed probability level of exceeding the targeted queue length. In this method much higher multiplexing gain and utilization can be achieved.

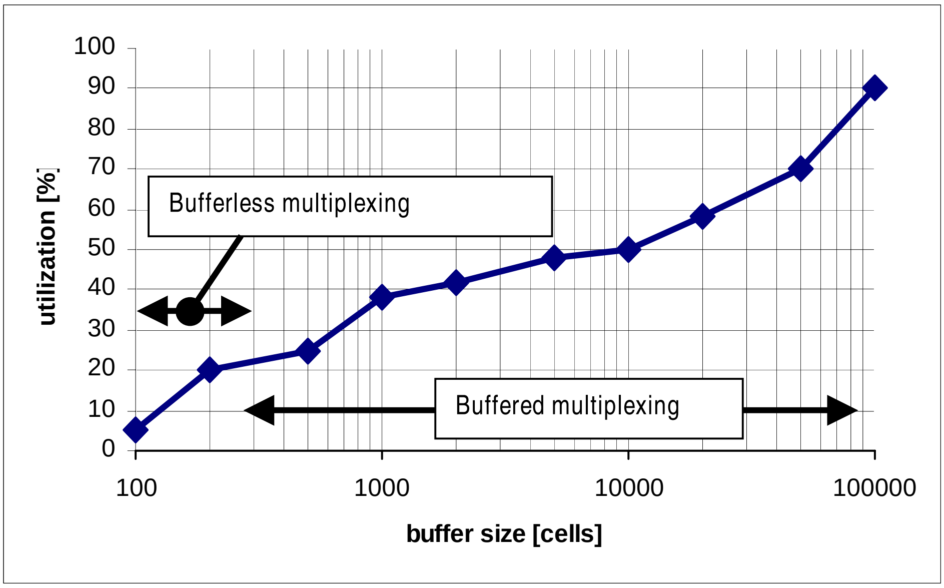

Figure 3.3.1: Alternatives for statistical multiplexing

The main problem in rate sharing is that the loss rate realized with a given buffer size and link capacity depends in a complicated way on the traffic characteristics including also the correlation structure. As an example the loss and delay characteristics are rather difficult to compute if the input traffic is LRD. This is the reason that the admission control methods are much more complicated for rate sharing than for rate envelope multiplexing [3.3.8]. Moreover, the disadvantage is not only the complex traffic control but the achievable utilization is also smaller in case of

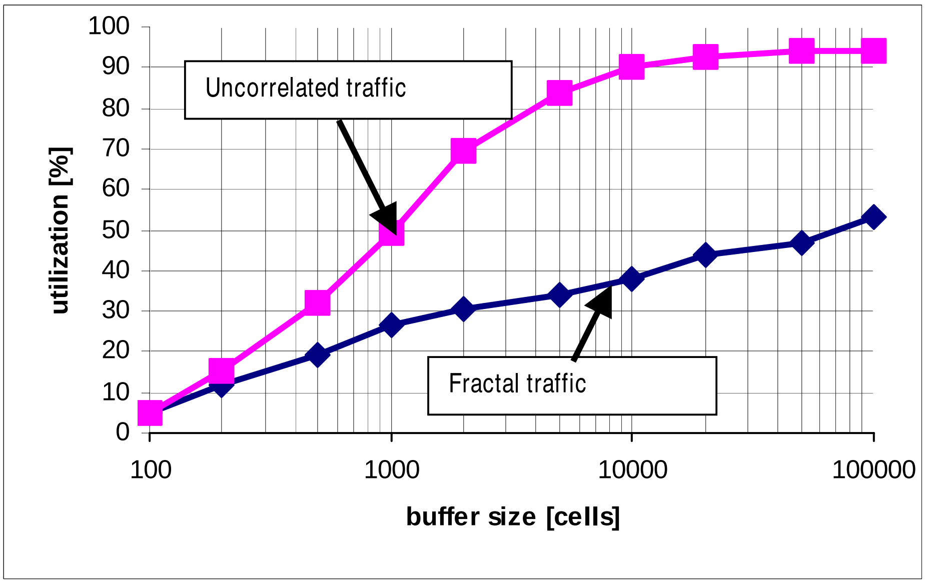

Figure 3.3.2: The impact of correlation structure

fractal traffic with strong SRD and LRD properties, see Figure 3.3.2.

A large number of admission control strategies have been developed for both rate envelope multiplexing and rate sharing [3.3.8]. It seems that the most powerful scheme is a kind of measurement-based admission control where the only traffic descriptor is the peak rate and the available rate is estimated in real-time.

3.3.4.4 The closed-loop control of elastic traffic

Elastic traffic is generally controlled by *reactive closed-loop traffic control* methods [3.3.6], [3.3.7]. This is the principle of the TCP in the Internet and the ABR in the ATM. These protocols target to fully exploit the available network bandwidth while keeping fair shares between contending traffic flows. Now we investigate the TCP as the general transfer protocol of the Internet. In TCP an additive increase, multiplicative decrease congestion avoidance algorithm has been implemented. If there is no packet loss the rate increases linearly but the packet transmission rate is halved whenever packet loss occurs. The algorithm tries to adjust its average rate to a value depending on the capacity and the current set of competing traffic flows on the links of its paths. The available bandwidth is shared in a roughly fair manner among the TCP flows.

A simple model of TCP [3.3.9], which also captures the fundamental behavior of the algorithm, is the well-known relationship between the flow throughput *B* and the packet loss rate *p*:

$$B(p) = \frac{c}{RTT\sqrt{p}},$$

where *RTT* is the TCP flow round-trip time and *c* is a constant. It should be noted that this simple formula is valid in case of a number of assumptions: *RTT* is constant, *p* is small (less than 1%) and the TCP source is greedy. The TCP mechanism is also assumed to be governed by the fast retransmit and recovery (no timeouts) and the slow-start phase is not modeled. More refined models were also developed but the square-root relationship between *B* and *p* seems to be a quite general rule of TCP.

Implementing admission control schemes for elastic traffic is a current and open research issue [3.3.6], [3.3.7]. In such a method the admittance threshold should be small enough to avoid flow rejection in normal load situations but large enough to ensure satisfactory throughput for admitted flows in overload.

3.3.5. Concluding remarks on teletraffic dimensioning

The importance of choosing a good traffic model determines how successful we are in capturing the most important traffic characteristics. The traffic model applied in the investigated teletraffic system, which is in most of the cases a queueing system, is the complex teletraffic model under investigation. The basic question is the fundamental relationship between the traffic characteristics, network resources and performance measures. Queueing models with some types of traffic models (e.g., Poisson, MMPP, MAP, etc.) are analytically tractable but others (e.g., ARIMA, TES, FGN, etc.) are not. It is a current research issue to develop new theoretical and applied tools to assist in solving teletraffic systems with emerging new and complex traffic models.

Our survey about the dimensioning methods shows that the teletraffic dimensioning of the Internet is not a fully solved problem and several issues are in the scope of current teletraffic research. As opposed to the dimensioning of telephone networks, which can be considered as a well understood and solved issue, the teletraffic theory of the Internet with dimensioning methods is mainly the topic of the future.

References

[3.3.1] D. L. Jagerman, B. Melamed, W. Willinger: Stochastic modeling of traffic processes, In J. Dshalalow, ed., Frontiers in Queueing: Models, Methods and Problems. CRC Press, 1997. pp. 271-320.

[3.3.2] V. S. Frost, B. Melamed: Traffic Models for Telecommunications Networks, IEEE Communications Magazine, March 1994. pp 70-81.

[3.3.3] G. D. Stamoulis, M. E. Anagnostou, A. D. Georganas: traffic sources models for ATM networks: a survey, Computer Communications, vol. 17, no. 6, June, 1994. pp. 428-438.

[3.3.4] R. G. Addie, M. Zukerman, T. D. Neame: Broadband Traffic Modeling: Simple Solutions to Hard Problems, IEEE Communications Magazine, August 1998. pp. 88-95.

[3.3.5] B. O. Lee, V. S. Frost, R. Jonkman: NetSpec 3.0 source Models for telnet, ftp, voice, video and WWW traffic, 1997.

[3.3.6] J. Roberts, Traffic Theory and the Internet, IEEE Communications Magazine, January 2000.

[3.3.7] J. Roberts, Engineering for Quality of Service, in the book of Self-Similar Network Traffic and Performance Evaluation, (eds. K. Park, W. Willinger), Wiley, 2000.

[3.3.8] J. Roberts, U. Mocci, J. Virtamo (eds.), Broadband Network teletraffic, Springer-Verlag, 1996.

[3.3.9] J. Padhye et al. Modeling TCP Throughput: A Simple Model and Its Empirical validation, Proc. SIGCOMM'88, ACM, 1998.

[3.3.10] A. Odlyzko: The history of communications and its applications for the Internet, available at http://www.research.att.com/~amo/doc/complete.html, 2000.

[3.3.11] H. Akimaru. K. Kawashima: Teletraffic, Theory and Applications, Springer-Verlag, 1999.