1.7. Teletraffic Theory

Sándor Molnár dr., author

László Jereb dr., reviewer

1.7.1. Introduction

Teletraffic theory [1.7.1], [1.7.7] is the basis for performance evolution and dimensioning of telecommunication networks. It was founded by Agner Krarup Erlang (1878-1929) [1.7.6], a Danish mathematician, at the beginning of the 20th century. The theory has been developed along with the developments of telephone networks [1.7.7] and it is an essential component in the design of traditional telecommunication networks.

Teletraffic theory has been developed together with the enormous developments of switching and networking technology in the last decades. It has been incorporating the recent advances of operation research and queueing theory. Integrating the results of different fields a continuous evolution of teletraffic theory can be observed.

Teletraffic theory deals with the application of mathematical modeling of the traffic demand, network capacity and realized performance relationships. The traffic demand is statistical in nature resulting in appropriate models derived from the theory of stochastic processes.

In this chapter first we present an introduction about the characteristics of network traffic. The nature of traffic had a strong impact on the developed teletraffic theory we have today. After we overview the basics of teletraffic theory including the notations, classification of systems and the fundamental teletraffic equations. The applications of basic teletraffic results outlined in this chapter can be found in Chapter 3.3 where teletraffic models and the teletraffic dimensioning methods are described.

1.7.2. The characteristics of network traffic

The nature of traffic in today's data networks (e.g. Internet) is completely different from classical telephone traffic and the characterization is not as simple as it was in the case of conventional POTS traffic [1.7.8], [1.7.9]. The main difference can be explained by the fact that in traditional telephony the traffic is highly *static* in nature. It was possible to find a typical user and behavior where averages simply describe the system performance adequately due to the *limited variability* of traffic characteristics.

The static nature of telephone traffic resulted in "universal laws" governing telephone networks like the *Poisson nature* of call arrivals [1.7.8], [1.7.9]. This law states that call arrivals are mutually independent and exponentially distributed with the same parameter. The Poisson call arrival model had a general popularity in the last fifty years. The great success of the Poissonian model is due to the parsimonious modeling, which is a highly desirable property in practice.

A similar "universal law" of the POTS traffic is that the call holding times follow more or less an *exponential distribution*. This model was also preferred due to its simplicity and analytical tractability in spite of the fact that the actual telephone call duration distribution sometimes deviates significantly from the exponential distribution. However, these deviations did not yield to major errors in network design thanks to the nice nature of Poisson arrival process. This is because several performance measures do not depend on the distribution but only of the average of holding time.

A dramatic change happened concerning the validity of these laws when telephone networks were used not only for voice conversations but also for FAX transmissions and Internet access. The statistical characteristics of these services are significantly different from voice calls. Especially, the call durations become much longer and more variable compared to classical voice calls. As the popularity of the Internet increased due to the success of Web, more and more people started to use the classical telephone networks for Internet access. These changes call for reviewing the old laws and present a challenge for today's teletraffic researchers.

The picture is completely different in case of data networks. All the expectations by finding similar universal laws for data traffic failed [1.7.8]. It is because data traffic is much more variable than voice traffic. Roughly speaking, it is impossible to find a general model because the individual connections of data communication can change from extremely short to extremely long and the data rate can also be in a huge range. There is no static and homogenous nature of data traffic as it was found in case of the voice traffic. This extremely bursty nature of data traffic is mainly caused by the fact that this traffic is generated by machine-to-machine communication in contrast to the human-to-human communication.

This high variability of data traffic in both *time* (traffic dependencies do not decay exponentially fast as it was the case in voice traffic but long-term dependencies are present, e.g. in the autocorrelation of the traffic) and in *space* (distributions of traffic related quantities do not have exponential tails as it was the case in the case of voice traffic but heavy tails are very common, e.g. in distributions of web item sizes) call for new models and techniques to be developed. Statistically, the long-term dependencies can be captured by *long-range dependence* (LRD), i.e., autocorrelations that exhibit power-law decay. The extreme spatial variability can be described by *heavy-tailed distributions* with infinite variance, which is typically expressed by the Pareto distributions. The power-law behavior in both time and space of some statistical descriptors often cause the corresponding traffic process to exhibit *fractal* characteristics [1.7.8].

The fractal properties often manifest themselves in *self-similarity*. It means that several statistical characteristics of the traffic are the same over a range of time scales. Self-similar traffic models seem to be successful parsimonious models to capture this complex fractal nature of network traffic in the previous decade. However, recent research indicates that the actual data traffic has a more refined burstiness structure, which is better captured by *multifractality* rather than only self-similarity, which is a special case of *monofractality*. Multifractal traffic models have also been developed [1.7.8].

Besides the very variable characteristics of data traffic there are other factors that make predictions about data traffic characteristics more unreliable. The Internet traffic is doubling each year. This extreme traffic increase with the possible so-called “killer applications” could disrupt any predictions. However, from the history of the Internet we can identify only three “killer applications” that dramatically changed the traffic mix of the Internet (the e-mail, the web and the recently emerging Napster-like applications) but nobody knows when we can face a popular application which will take the major role of the Internet traffic characteristics. The picture is even more complicated if we think of Quality of Service (QoS) requirements of data services which can be very different from one application to the other. Different QoS requirents generate different traffic characteristics.

To describe these different traffic characteristics in case of both stream and elastic traffic flows a number of traffic models and traffic characterization techniques have been developed. Based on a successful traffic modeling one can also hope to find successful traffic dimensioning methods for resource allocation. The most important traffic models and dimensioning methods are described in Chapter 3.3.

1.7.3. Basic concepts of teletraffic theory

In this subsection the most important teletraffic concepts are overviewed [1.7.1].

1.7.3.1. Basic notions

A demand for a connection in a network is defined as a *call,* which is activated by a *customer*. The call duration is defined as *holding time* or *service time*. The *traffic load* is the total holding time per unit time. The unit of traffic load is called *erlang* (erl) after the father of teletraffic theory.

The traffic load has the following important properties:

1. The traffic load (offered traffic) a is given by $a=ch$ (erl) where c is the number of calls originating per unit time and h is the mean holding time.
2. The traffic load (offered traffic) is equal to the number of calls originating in the mean holding time.
3. The traffic load (carried traffic) carried by a single trunk is equivalent to the probability (fraction of time) that the trunk is used (busy).
4. The traffic load (carried traffic) carried by a group of trunks is equivalent to the mean (expected) number of busy trunks in the group.

1.7.3.2. Classification of teletraffic systems

The *switching system* is defined as a system connecting between inlets and outlets. A system is called a *full availability system* if any inlet can be connected to any idle outlet. *Congestion* is a state of the system when a connection cannot be made because of busy outlets or internal paths. The system is called a *waiting* or *delay system* if an incoming call can wait for a connection in case of congestion. If no waiting is possible in congestion state the call is blocked and the system is called as *loss system* or *non-delay system*.

A full availability system can be described by the following [1.7.1]:

1. Input process: This describes the way of call arrival process.

2. Service mechanism: This describes the number of outlets, service time distributions, etc.

3. Queue discipline: This specifies ways of call handling during congestion. In delay systems the most typical queueing disciplines are the first-in first-out (FIFO), last-in first-out (LIFO), priority systems, processor sharing, etc.

The *Kendall notation* is used [1.7.1], [1.7.3], [1.7.4] for classification of full availability systems named after David A. Kendall, a British statistician:

A/B/C/D/E-F

where *A* represents the interarrival time distribution, *B* service time distribution, *C* number of parallel servers, *D* system capacity, *E* finite customer population, and *F* is the queueing discipline. The following notations are used:

M: Exponential (Markov)

E_k: Phase k Erlangian

H_n: Order n hyper-exponential

D: Deterministic

G: General

GI: General independent

MMPP: Markov modulated Poisson process

MAP: Markov arrival process

As an example $M/M/1/\infty//\infty$-FCFS represents a queueing system with Poisson arrivals and exponentially distributed service times. The system has only one server, an infinite waiting queue. The customer population is infinite and the customers are served on a first come first served basis.

1.7.3.3. Fundamental relations

PASTA

For a Poisson arrival process (exponential interarrival times) in steady state the distribution of existing calls at an arbitrary instant is equal to the distribution of calls just prior to call arrival epochs. This relationship is called *PASTA* (Poisson arrivals see time averages) [1.7.1] because this probability is equal to the average time fraction of calls existing when observed over a sufficiently long period.

Markov property

If the interarrival time is exponentially distributed, the residual time seen at an arbitrary time instant is also exponential with the same parameter. A model with interarrival time and service time both exponentially distributed is called *Markovian model* [1.7.1], otherwise it is called *non-Markovian model*.

Little Formula

The formula $N=\lambda W$ is called the *Little formula* [1.7.1], [1.7.3], [1.7.4] where N is the mean number of customers in the system, λ is the mean arrival rate and W is the mean waiting time in the system. Note that the Little formula applies to any stationary system where customers are not created or lost in the system.

Loss Formula

The probability of an arbitrary customer being lost [1.7.5] is

$$P_{loss} = 1 - \frac{1-\phi}{\rho},$$

where ρ is the offered load and ϕ is the probability that the server is idle. The formula is called the *loss formula* and is also valid for multiserver systems with ρ being interpreted to be the mean load per server and ϕ is the probability that arbitrarily chosen server is idle.

Unfinished work vs. number in the system

In a constant service time single server system we have the following relationship between the unfinished work V_t in the system and the number of costumers X_t in the system: $X_t = \lceil V_t \rceil$. Based on this we have the following identity for the complementary distributions [1.7.5]:

$P(X_t > n) = P(V_t > n)$, for *n* an integer.

Packet loss probability vs. queue length tail probability

Consider a discrete time *G/D/1* system with fixed length packet arrivals (cells). An upper bound for the cell loss probability can be given [1.7.5] by

$\rho P_{loss} \leq P(X_t^{\infty} > K)$,

where ρ is the load, X_t^{∞} is the queue length in a hypothetical infinite capacity queue.

The generalized Beneŝ formula

Consider a service system with unlimited buffer. Assume that the system is stationary so that 0 represents an arbitrary time instant. The server capacity is 1 unit work per unit of time. The complementary distribution of the amount of work in the system at time 0 can be computed [1.7.5] by

$$P(V_0 > x) = \int_{u>0} P(\xi(u) \geq x > \xi(u+du) \qquad and \qquad V_{-u} = 0),$$

where $\xi(t)$ is defined by $\xi(t) = A(t) - t$, $t \geq 0$, and *A(t)* is the amount of work arriving to the system in the interval [-*t*,0). The result covers all realizable queueing systems and found to be very useful in teletraffic theory.

1.7.4. The M/G/1 queue

The queueing system with Poisson arrivals, general service time distributions and a single server (*M/G/1*) is a very important category in teletraffic theory. In this subsection we overview the major results related to this queueing system [1.7.1], [1.7.2], [1.7.3], [1.7.4].

The following notations are used:

W: waiting time in the queue

T: response time in the system

N_q: number of customers in the queue

N: number of customers in the system

S: service time

The average waiting time and the number of customers in the queue for the *M/G/1* queueing system are given by the following equations [1.7.2]:

$$E(W) = \frac{\rho E(S)(1+c_S^2)}{2(1-\rho)}, \; E(N_q) = \frac{\rho^2(1+c_S^2)}{2(1-\rho)}$$

where ρ is the load of the queue and c_S^2 is the squared coefficient of variation of the service time, i.e. Var(S)/E^2(S).

The distribution of the number of customers in the system can be computed from the Pollaczek-Khinchin transform equation [1.7.2]:

$$G_N(z) = L_S(\lambda(1-z))\frac{(1-\rho)(1-z)}{L_S(\lambda(1-z))-z},$$

where $G_N(z) = E[z^N]$ the probability generating function for *N,* $L_X(s) = E[e^{-sX}]$ the Laplace transform for *X* and λ is the Poisson arrival rate. Based on this key equation the obtained queue length distributions for the most frequently used *M/G/1* systems are summarized below.

Queue	Queue length distribution P(*N*=*n*)
M/M/1	$(1-\rho)\rho^n$
M/H₂/1	$q(1-\alpha_1){\alpha_1}^n + (1-q)(1-\alpha_2)\alpha_2^n$
M/D/1	$(1-\rho)\sum_{k=0}^{n} e^{k\rho}(-1)^{n-k}\frac{(k\rho+n-k)(k\rho)^{n-k-1}}{(n-k)!}$
M/E$_k$/1	$(1-\rho)\sum_{j=0}^{n}(-1)^{n-j}\frac{\alpha^{n-j-1}}{(1-\alpha)^{kj}}\left[\binom{kj}{n-j}\alpha + \binom{kj}{n-j-1}\right]$

Here H_2 means the *hyperexponential* distribution given by parameters α_1, α_2 and *q*. E_k refers to the *k-Erlangian* distribution given by parameters α and *k*. These two distributions are important because using them we can approximate *M/G/1* systems where the squared coefficient of variation of the service time less than or equal to 1 (*M/E$_k$/1* queues) and greater than or equal to 1 (*M/H$_2$/1* queues).

1.7.5. General queueing systems

General queueing systems (*G/G/n* queues) are usually difficult to solve but there are subclasses that can be handled more easily than others. For example, the *G/M/1* queueing system is less useful than the *M/G/1* in data networks but the analysis is simpler than its dual pair. A remarkable result of the *G/G/1* systems is the *Lindley's integral equation* [1.7.1], [1.7.3] which gives the stationary waiting time distribution:

$$F_W(t) = \int_{-\infty}^{t} F_W(t-v)dF_U(v)\,,$$

where the random variable *U =S-A* with *A* denoting the time between the arrivals of two consecutive customers.

1.7.6. Teletraffic techniques

Beyond the classical queueing methods there are numerous approximations, bounds, techniques to handle teletraffic systems. In this subsection we overview the most significant methods.

The *fluid flow approximation* [1.7.3] is a useful technique when in the time scale under investigation we have lots of traffic units (packets). In this case we can treat it as a continuous flow like fluid entering a piping system. We can define *A(t)* and *D(t)* to be the random variables describing the number of arrivals and departures respectively in (0,*t*). The number of customers in the system at time *t* is *N(t)=A(t)-D(t)*, assuming that initially the system is empty. By the weak law of large numbers, when *A(t)* gets large it gets close to its mean and this is the same for *D(t)*. The fluid flow approximation simply replaces *A(t)* and *D(t)* by the their means, which are continuous deterministic processes. Fluid flow models are frequently used in teletraffic systems modeling.

The fluid flow approximation uses mean values and the variability in the arrival and departure processes is not taken into account. The *diffusion approximation* [1.7.3] extends this model by modeling this variability (motivated by the central limit theorem) by normal distribution around the mean. Diffusion approximations are also applied to solve difficult queueing systems. For example, the in the complex *G/G/1* system the queue length distribution can be obtained by diffusion methods.

An approach based on the information theory called the *maximum entropy method* [1.7.3] is often useful in solving teletraffic systems. The basis is Bernoulli's principle of insufficient reasons which states that all events over a sample space should have the same probability unless there is evidence to the contrary. The entropy of a random variable is minimum (zero) when its value is certain. The entropy is maximum when its value is uniformly distributed because the outcome of an event has maximum uncertainty. The idea is that the entropy be maximized subject to any additional evidence. The method is successfully used for example in queueing theory.

A number of other methods have also been developed like queueing networks with several solving techniques, fixed point methods, decomposition techniques, etc. Interested readers should refer to the reference list of this chapter.

References

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

[1.7.2] R. Nelson: Probability, Stochastic Processes, and queueing Theory, Springer-Verlag, 1995.

[1.7.3] P. G. Harrison, N. M. Patel: Performance Modelling of Communication Networks and Computer Architectures, Addison-Wesley, 1993.

[1.7.4] R. Jain: The Art of Computer Systems Performance Anaysis, Wiley, 1991.

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

[1.7.6] E. Brockmeyer, F. L. Halstrom, A. Jensen: The Life and Works of A. K. Erlang, Acta Polytechnica Scandinavica, 1960.

[1.7.7] R. Syski, Introduction to Congestion Theory in Telephone Systems, Oliver and Boyd Ltd. 1960.

[1.7.8] W. Willinger, V. Paxson: Where Mathematics Meets the Internet, Notices of the American Mathematical Society, vol.45, no.8, Aug. 1998, pp. 961-970.

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