2

Classical Traffic Theory

There is nothing more practical than a good theory.

Immanuel Kant (18th century)

2.1 Introduction to Traffic Theory

2.1.1 Basic Examples

To illustrate the complexity of traffic theory we begin this chapter with three basic examples. Each example refers to certain aspects of traffic theory briefly and motivates the broad variety of models we will introduce in later sections of this chapter.

PAR Protocols

We already noted in section 1.1.4 that protocols following positive acknowledgment with retransmit (PAR protocols) are basic for reliable data transfers as implemented with TCP. A sender is transmitting data packets to a receiver. If the packets are received without error according to the underlying protocol, those packets are acknowledged by a signal, called the ACK (positive acknowledgment). If the ACKs are not received within a certain timeout interval, the sender retransmits the respective packets. This procedure is repeated until the sender receives an ACK for each transmitted packet or a given number of retransmits is reached and an error message is delivered to the respective application.

We denote by η_N the transmission time and by τ the delay, caused by a single retransmission cycle. We want to compute the virtual transmission time η_ν. This is the actual time for the successful transmission. We have to impose a certain probability for the failure transmission, which we denote by p_η. So we have by assuming the independence of the repeated transmission:

- with a probability of $1 - p_\eta$ the virtual transmission time $\eta_\nu = \eta$ and
- with probability p_η an additional transmission time of $\eta + \tau$ for η_ν.

Considering the packet arrival time, we get a classical traffic model, where the serving time is the virtual transmission time. As we will see in section 2.8.2, this is a so called one level GI/G/1 system, assuming general distributed arrival and serving time. As expected the analysis will provide e.g. queueing time and traffic load depending on the error probability p_η.

Data Transmission

To give another example for the application of the classical traffic theory, we consider the basic concept of data transmission using the packet switched technique. As described in section 3.4.3 we apply the most basic model, given in the figure 2.1.

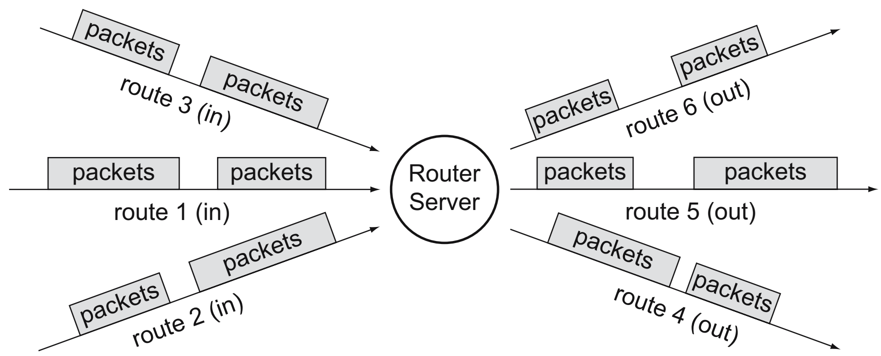

Fig. 2.1. Scheme of incoming and outgoing data transmission at a router on the packet level

It corresponds to the packet or burst level, which we discussed in section 1.2.1 and explained by the figure 1.5. We can consider each of the packets as *demands*, which have to directed in the router resp. server to further outgoing lines. These packets arrive with a certain rate, depending e.g. on the user behavior. This is determined e.g. by the dial in of all users into the Internet or intranet and indicates the interarrival times of the packets. Thus, the incoming stream is modeled as the *arrival process*. The router resp. server plays the rôle of the *serving part* of the traffic model. Here, depending of the future structure of the net and the destination of each of the transmitted data, these packets are distributed or served. Hence, we will call this the *serving process*. Since the incoming streams may surpass the outgoing capacity, we will observe an storage of data prior to the further transmission. Thus, these packets have to wait and will go into the queueing. We will denote this as the *waiting room* and the corresponding process the *queueing process*.

Call Attempts in a Cellular Mobile System

Let's consider a cell in a mobile communication system, more precisely in the most prominent global system for mobile communication, the *Global System*

for Mobile Communications (GSM). This system is the most successful cellular communication system which was introduced in 1991, mainly in Europe and with adjustments in the USA and Japan. It has been based from the outset on the already existing large variety of different cellular systems. In GSM several new technologies entered the mobile communication, where, next to the unique standardization, we encounter the first time a digital transmission technique on the air interface. For our present purpose we have to know that the area covered by the mobile communication is split into a beehive-like structure, i.e. into several cells, each equipped with one base station or *Base Transceiver Station* (BTS). Every mobile station can get in contact with this fixed station, where data (mostly the encoded speech) is transferred via fixed line, finally to the core network for further handling.
The basic transmission technique on the air interface is a combination of time and frequency multiplex, since each user has to be separated from the other. To avoid interference between neighbouring cells, network management has to apply a careful frequency planning for each cell, a fact which can be neglected for the *Universal Mobile Telecommunications System* (UMTS) technique. Guard bands have to be implied as well.

- Time multiplex: Each frequency band is splitted in 8 time slots, where for each second frequency multiplex channel one channel is used for service transmission.
- Frequency multiplex: For each user there exits a pair of frequencies – the upper frequency for download the lower one for upload. In GSM900 the first and mostly used standard the band 935 to 960 MHz are receiver frequency and 890 to 915 MHz is the upload frequency band. With a distant of 200 kHz a maximum of 124 channel pairs can be provided.

The basic quality of service in GSM is that a certain blocking of incoming call attempts are not allowed to be surpassed by a fixed given probability. For the modeling of the GSM cell we use:

- Serving process: In each cell we have n given frequency pairs for the voice transmission. They build the serving units. The duration of a call is describe as serving time by a random variable η. An accepted call will occupy a pair of channel, i.e. a serving unit. If all serving units resp. channel pairs are occupied, the next incoming call is blocked and thus, rejected. We encounter a pure loss system.
- Arrival process: In each cell we find a finite number of m users, which are 'silent', 'active' or 'idle'. We can determine three kinds of arrival processes.
  - Arrival process with finite sources: we have m users, i.e. a finite number. The aggregated traffic is the compound of all users in the state 'silent' (see section 2.4.3).
  - Arrival process with infinite sources: suppose the number m of users is sufficient large. Then we can approximate the system by a model with infinite sources. This model is easier to handle and leads to the well known Erlang formula (see section 2.4.1).

– Model with call repetition: in reality an overload situation can be detected, which emerges faster, since usually rejected call in a blocking situation try to repeat the call in decreasing time intervals (so called snow ball effect) (see section 2.4.3).

2.1.2 Basic Processes and Kendall Notation

The classical model in the traffic theory can be divided into three basic processes: arrival, serving and state processes. We briefly summarize the main characteristics of these three models.

Arrival Process

The incoming demands are modeled as discrete time spots. They can include call attempts in the telecommunication, data packets or data units as observed in high speed networks like ATM. We can divide this class of arrival processes into single or group arrivals. The time between two succeeding arrivals is called interarrival time and represents a central issue in modeling (fig. 2.2).

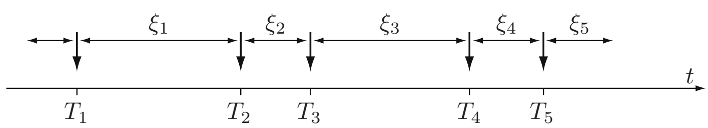

Fig. 2.2. Arrival process: arrival times T_i and interarrival times ξ_i

Serving Process

After its arrival the demand is handed over to the serving process. This leads to a number of demands in the serving process (a discrete process) and to the serving time (a time continuous process). The interarrival times as the serving time (serving process) are often modeled as Markov processes, i.e. we consider both processes as memoryless. It is assumed that the remaining time has the same distribution as the whole interarrival or serving time. Here, the remaining time is the time left in the serving process (fig. 2.3).

State Process

Often two kinds of state processes are considered:

- Number of demands in the system: Here, we have a discrete process. For example every incoming call increases the number of demands, where each of these calls could be served or fail.

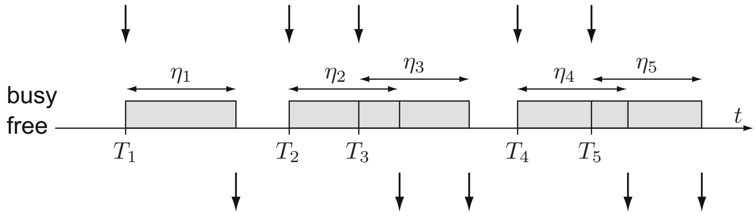

Fig. 2.3. Serving process: arrival times T_i and service times η_i

- Residual work in the system: Every incoming demand increases the remaining amount of work. The residual work is not identical with the serving time, since we have to consider the waiting time in addition. The residual time will decrease continuously and thus, we deal with a continuous process (fig. 2.4).

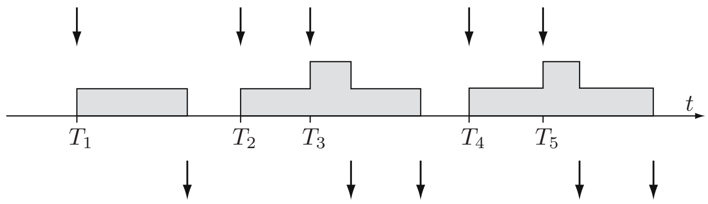

Fig. 2.4. State process: upper arrows mark arrivals, lower arrows departures of demands; gray shaded areas the number of demands in the system

We can represent arrival and serving processes in a short notation, the so called *Kendall notation*, e.g. M/M/$n-S$. Here, the first symbol reflects the arrival process, the second one the serving process, with n we indicate the number of possible simultaneous serving processes and S is the number of waiting places. In the above example the symbol M means that we have chosen a 'Markov' process as arrival and serving process. If we do not assume a Markov process – and this happens especially in the IP serving processes – then we use the notation M/G/$n-S$, GI/M/$n-S$ or GI/G/$n-S$. Here, the symbol G means 'generally' distributed resp. GI 'generally independent' distributed. Hence, the arrivals (resp. calls or initializations of Internet connections) are stochasticly independent. The distribution of arrival resp. serving processes are arbitrary.

2.1.3 Basic Properties of Exponential Distributions

We call a distribution $F_\xi(t) = 1 - e^{-\alpha t}$ with density $f_\alpha(x) = \alpha e^{-\alpha x}$, the *exponential distribution* for α, where the parameter α is called *intensity.* If we choose independent RV $\xi_1, \ldots, \xi_k$ for a $k \in \mathbb{N}$ exponential distributed with

http://www.springer.com/978-3-540-70603-8

IP-Traffic Theory and Performance
Grimm, C.; Schlüchtermann, G.
2008, XIV, 487 p., Hardcover
ISBN: 978-3-540-70603-8