Traffic intensity

$$a = \lambda \cdot T$$

where

λ = number of carried connections per time unit (arrival rate, call rate)

T = mean duration of a connection or holding time

- Traffic intensity is a bare number, but in order to emphasize the context, one often writes as its ”unit” erlang (E, erl)
- A.K. Erlang (1878-1929) was the pioneer of traffic theory, which he applied to study telephone systems
- Traffic intensity describes the mean number of simultaneous call in progress
- Instead of a "connection" we may consider reservation of any resource (trunk, modem, capacity unit)

Example

- In a local switch the number of calls in an hour is 1800
- The mean holding time of a call is 3 min

$$a = 1800 \times 3 / 60 = 90 \text{ erlang}$$

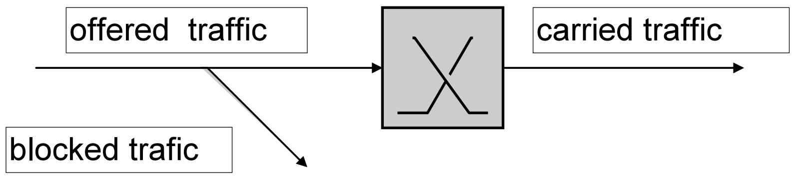
Typical traffic intensities per a single source are (fraction of time they are being used)

– private subscriber	0.01 - 0.04 erlang
– business subscriber	0.03 - 0.06 erlang
– mobile phone	0.03 erlang
– PBX	0.1 - 0.6 erlang
– coin operated phone	0.07 erlang

A load of 90 erlang is created by a population of some 2250 - 9000 private subscribers.

Traffic flows

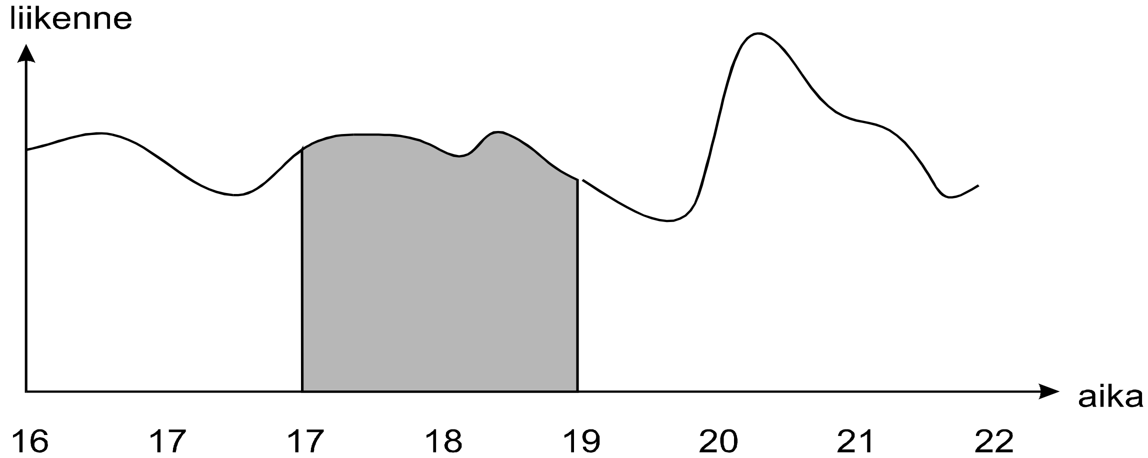
We distinguish between three components:



- Offered traffic a_o
  - traffic, which would be carried were there no constraints in the system
  - a theoretical concept
- Carried traffic a_c
  - traffic that is actually being carried
- Blocked (lost) traffic a_l
  - difference between the offered and carried traffics

Traffic volume

The amount of traffic carried during a given period of time is called the *volume of the traffic*



- The unit of traffic volume is e.g.
  - erlang · hour
  - call · minute
- Traffic volume in a period divided by the length of the period is the average traffic intensity in that period

Traffic variations

Traffic fluctuates over several time scales

- *trend* (> year)
  - the overall traffic growth: number of users, changes in the usage
  - traffic predictions give the basis for network planning
- *seasonal variations* (months)
  - changes related to different seasons (e.g. vacation period)
- *weekly variations* (days)
  - different activities on different days
- *daily profile* (hours)
  - variations related to different daily routines
- *random fluctuations* (seconds -- minutes)
  - fluctuations in the number of independent active users (Poisson process)

- The last component is purely *stochastic*
- The other variations by large follow a given *profile*, around which the traffic randomly fluctuates (each day, week, month... is different)

Traffic variations (continued)[1]

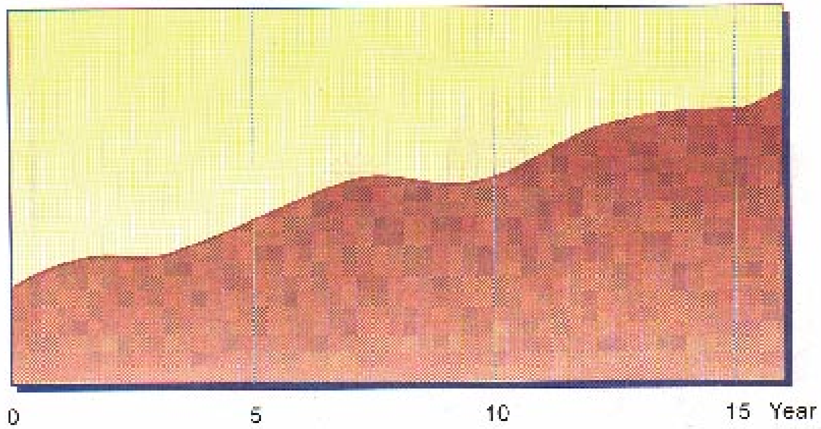


Figure 7 Many-year traffic trend with market fluctuations

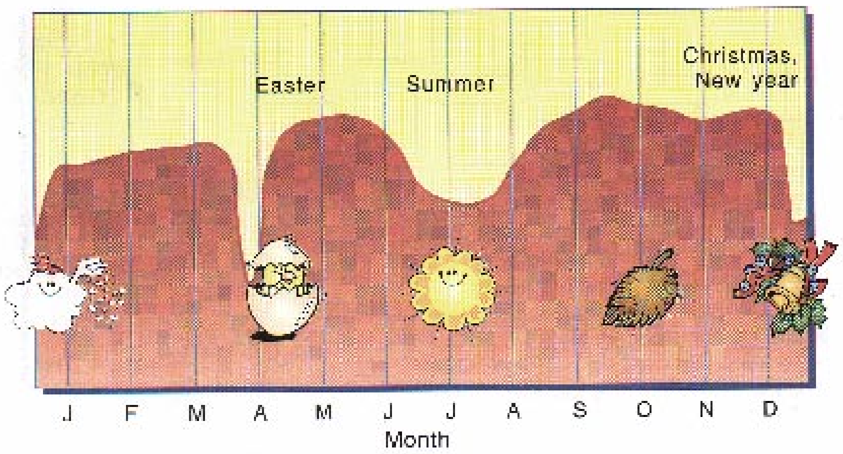


Figure 6 Typical busy hour variation over one year

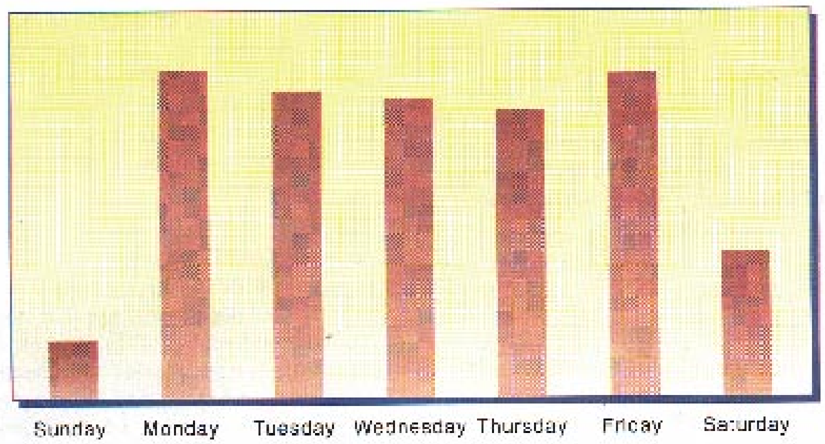


Figure 5 Typical busy hour traffic variation over seven weekdays

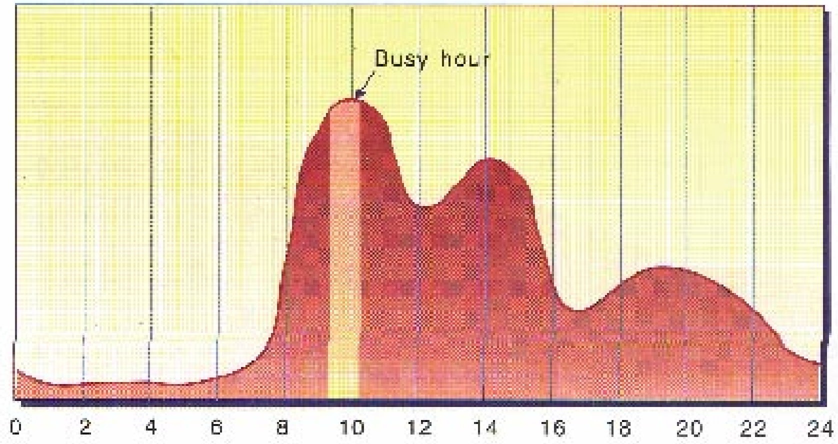
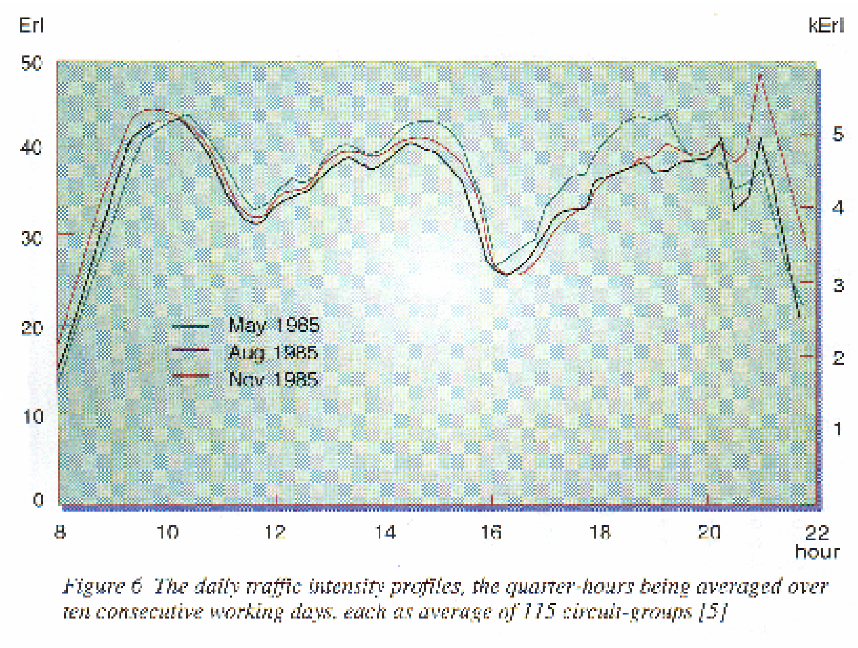


Figure 4 Typical 24-hour traffic variation

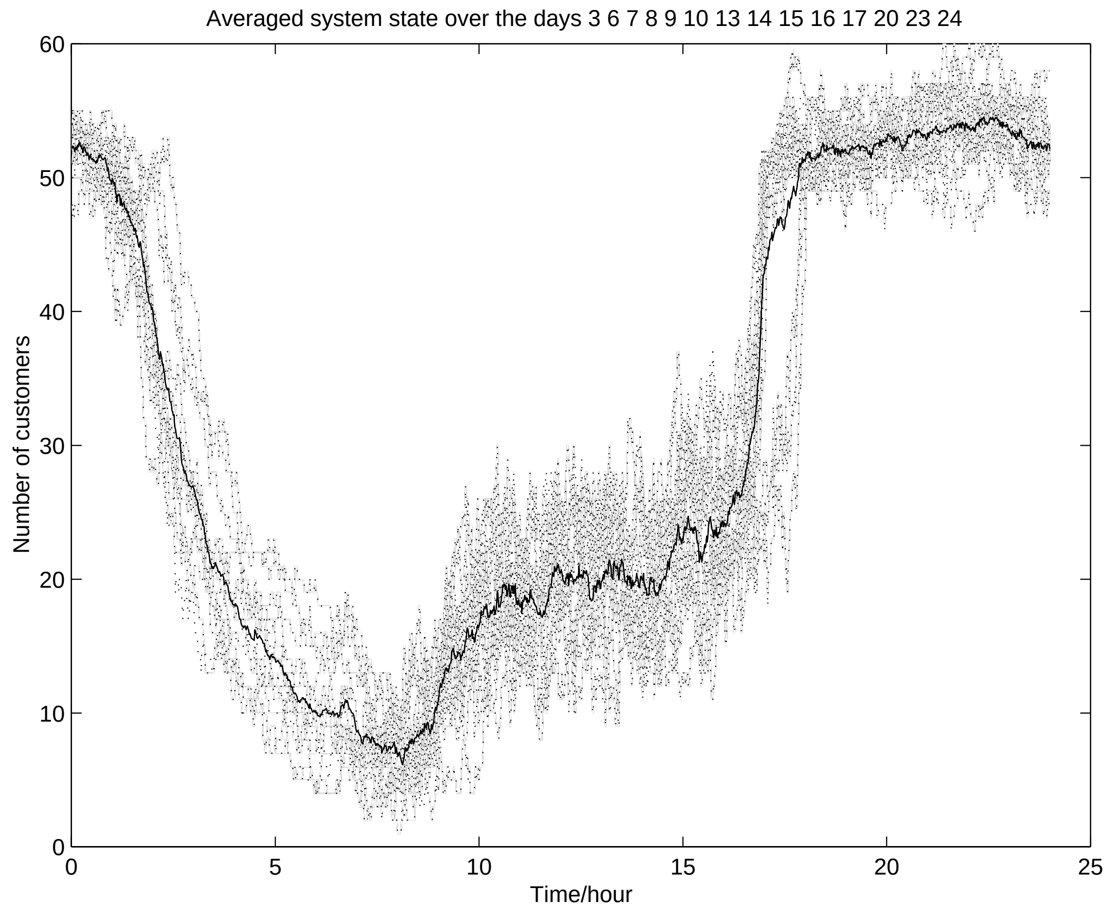
[1] From A. Myskja, *An introduction to teletraffic*, Telektronikk 2/3 (1995), pp 3-40

Traffic variations (continued)

Some daily traffic profiles measured in Finland



Daily profile of the telephone traffic in 1985 in three different months[1].



Traffic of the student modem pool at HUT; working days in October 1997[2].

[1] A. Parviala, *Observed traffic variations…*, Telektronikk 2/3 (1995), pp 69-78.

[2] J. Lakkakorpi, Erikoistyö (1998)

Traffic variations (continued)

- On a short time scale, the call arrival process can be considered a Poisson process with a given intensity λ.
- Due to changes in the activities of the users, the intensity λ is not constant but depends on time

$$\boxed{\lambda = \lambda(t)}$$

- By the terminology of stochastic processes, we have
  - an inhomogeneous Poisson process, if $\lambda(t)$ is a deterministic function of time
  - a double stochastic Poisson process, if $\lambda(t)$ itself constitutes a stochastic process
- In either case, the intensity of the Poisson process varies as a function of time: $\lambda(t)$
- The actual arrivals occur at random instants of time according to the Poisson process
  - the probability for an arrival in the interval $(t,t+dt)$ is $\lambda(t)\, dt$

Traffic variations (continued)

- In reality, $\lambda(t)$ is a stochastic process
- There is, however, a very strong deterministic component, related to the known predictable variations
  - regular daily profile
  - regular weekly profile
  - regular annual profile
  - evolution over a long time, trend
- Note. At no time scale does the word "regular" imply fully deterministic behaviour. $\lambda(t)$ fluctuates around the average profile.
- In addition, the arrival process may exhibit variations induced by some external event
  - predictable / unpredictable
  - regular / irregular

Busy hour

- It is not practical to dimension a network for the largest traffic peak that may ever occur. For pragmatic dimensioning work, one has developed a computational quantity which tries to adequately describe the peak load, but where singular peaks have been averaged out.

The period of duration of one hour where the volume of the traffic is the greatest

- Due to several random factors, the traffic fluctuates around its average
- In order to determine an appropriate dimensioning load, the recommendations define how the busy hour traffic shall be measured
- In fact, there are several definitions (ITU E.600)
  - an operator may choose the most appropriate one

Busy hour (continued)

- ADPH (Average Daily Peak Hour)
  - one determines the busiest hour separately for each day (different time for different days), and then averages over e.g. 10 days
  - the resolution of the start time of the busy hour may be either a full hour (ADPH-F) or a quarter of an hour (ADPH-Q)

- TCBH (Time Consistent Busy Hour)
  - a period of one hour, the same for each day, which gives the greatest average traffic over e.g. 10 days

- FDMH (Fixed Daily Measurement Hour)
  - a predetermined, fixed measurement hour (e.g. 9.30-10.30); the measured traffic is averaged over e.g. 10 days

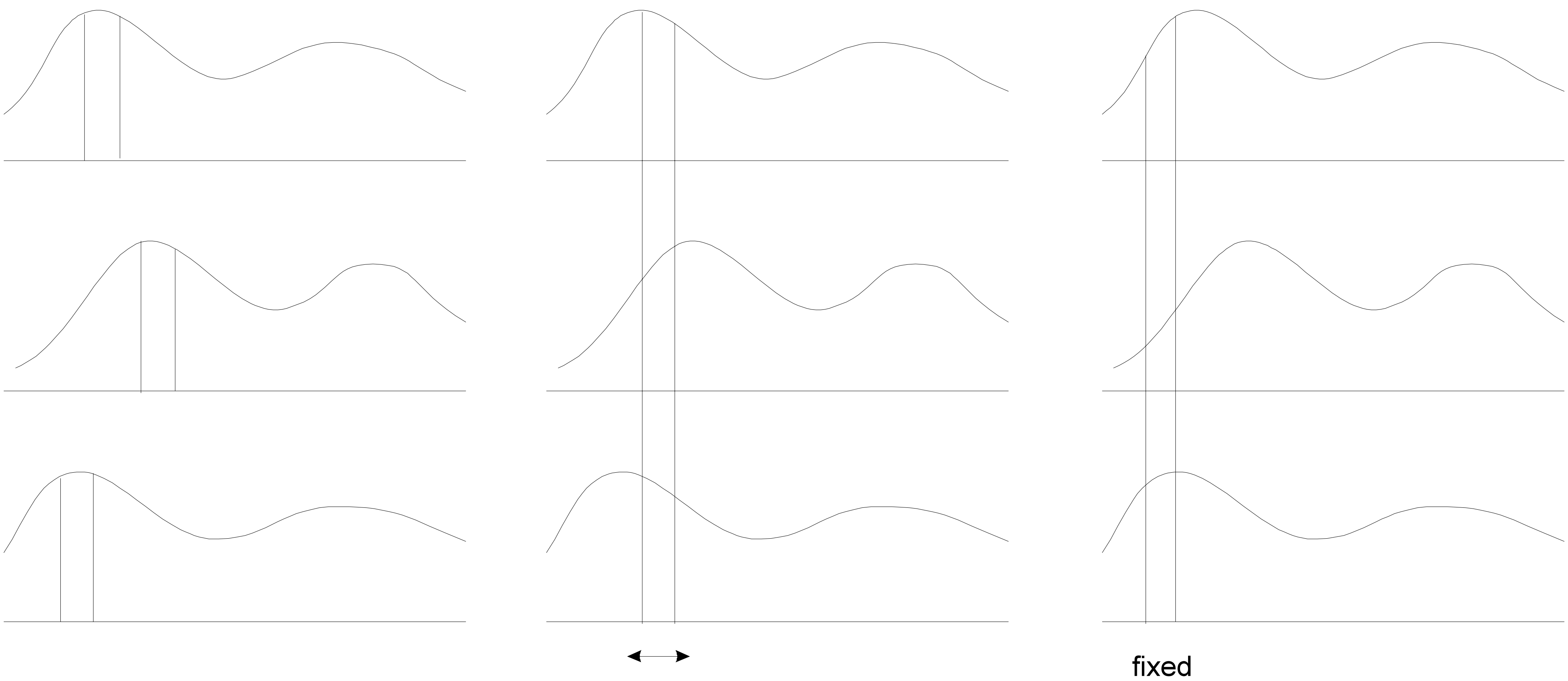
$$a_{FDMH} \le a_{TCBH} \le a_{ADPH}$$

Busy hour

- The busy hour definitions are further divided according to the used time resolution. For instance,
  - ADPH-F resolution of an hour
  - ADPH-Q resolution of an quarter of an hour

$$a_{\text{ADPH-F}} \leq a_{\text{ADPH-Q}}$$

Busy hour



ADPH, Average Daily Peak Hour

TCBH, Time Consistent Busy Hour

FDMH, Fixed Daily Measurement hour

Quality of service

- A network cannot be dimensioned for the worst case peaks. Then, occasionally the requested service is not available or the quality of the service is reduced.

- The dimensioning has to made according to the stated (statistical) criteria for the quality of service
  - grade of service (GoS): quality at the call level (e.g. telephone network)
  - quality of service (QoS): quality during a connection or session (e.g. ATM network)

- In a telephone network a call that cannot be immediately carried
  - may be blocked: loss system
  - may have to wait (ringing tone): waiting system

- The GoS requirement
  - loss system: P(call is blocked) < x %
  - waiting system: P(waiting time > z seconds) < x %

Quality of service (continued)

- Loss system
  - typically a blocking may occur during the busy hour
  - this happens with a certain probability, which depends on the traffic intensity during the busy hour and the dimensioning of the network as described by Erlang's formula(so called B formula)
  - the blocking probabilities in different parts of the network can summed to approximately estimate the end-to-end blocking
- Waiting system
  - if connecting the call is not immediately possible, the call may be put in a waiting state
  - a small waiting time does not matter, a user may not notice it at all
  - long waiting times are unacceptable for the users
  - one sets an upper limit to the waiting time, after which the call is blocked
  - the behaviour of a waiting system is described by so called Erlang's C formula
- There may be reattempts after unsuccessful calls

Quality of service (continued)

- It is not reasonable to dimension the network for a very small blocking probability, since the call may be unsuccessful due to other reasons with a much higher probability:
  - B subscriber does not answer
  - B subscriber is busy
  - one has dialled a wrong number
- Often the set limit for the blocking probability is 1 %

Quality of service (continued)

- In other networks than the traditional POTS, the quality of service is described by many other quantities, in place of or in addition to the blocking probability

- In ATM networks and in packet networks, e.g. the Internet, the following may be important
  - packet / cell delays
  - delay variation (jitter)
  - the proportion of lost packets / cells
  - the proportion of erroneous packets / cells
  - throughput

Erlang's formula

Assumptions

- A loss system: a blocked call is cleared (no reattempts)
- There are n trunks; any free trunk can be used
- The arrivals constitute a Poisson process
  - the arrivals occur at average rate λ
  - otherwise, the arrivals are completely random
  - this is good model when the calls originate from a large population of independent users
- traffic intensity $A = \lambda \cdot s$, where s is the mean holding time

$$\mathrm{E}(n, A) = \frac{\dfrac{A^n}{n!}}{1 + A + \dfrac{A^2}{2!} + \cdots + \dfrac{A^n}{n!}}$$

Relates the system (n), the traffic (A) and quality of service (E)

Erlang's formula (continued)

Example

- In a modem pool there are n = 4 modems and the offered traffic intensity is A = 2 erlang. What is the probability that a call attempt fails?
  - Consult precomputed graphs / tables or compute directly from the formula

$$\mathrm{E}(4,2) = \frac{\frac{2^4}{4!}}{1 + 2 + \frac{2^2}{2!} + \frac{2^3}{3!} + \frac{2^4}{4!}} \approx 9.5\%$$

- What is the blocking probability, if the number of modems in increased to 6?

$$\mathrm{E}(6,2) = \frac{\frac{2^6}{6!}}{1 + 2 + \frac{2^2}{2!} + \frac{2^3}{3!} + \frac{2^4}{4!} + \frac{2^5}{5!} + \frac{2^6}{6!}} \approx 1.2\%$$

Erlang's formula (continued)

The required number of trunks for different traffic intensities when the maximum allowed blocking probability is 1 %:

A (erlang)	n	n / A
3	8	2.7
10	18	1.8
30	42	1.4
100	117	1.17
300	324	1.08
1000	1029	1.03

- When A is small, the number of trunks is manyfold in comparison with the mean load
  - the load factor (utilization) is small
- For large A, the need for overprovisioning is very small
  - then it is more important to focus on the correctness of A, on which the dimensioning is based

Time and call blockings

One has to be careful to make a distinction between

- *Time blocking*
  - the fraction of time when all the resources are occupied
- *Call blocking*
  - the fraction of all calls that are blocked

- In general, these are two different things
  - from the point of view of the applications, one is often more interested in the call blocking; this defines the quality of service as experienced by the customers
  - time blocking, however, is often easier to compute

- Fortunately, with the assumptions of Erlang's formula (Poisson arrivals), time blocking and call blocking are the same

Quality of Service (QoS) in the Internet

- One of the hot topics in the development of the Internet
- In the Internet as it is today, the only service offered is so called Best Effort service
  - no guarantees for the quality
  - no mechanism for differentiating the service between different customers (applications) according to their needs (neither differentiation in charging)
  - there are no resource reservation for the flows
  - relies on the flow control of TCP, where the sources upon detecting the network in a congested state (by means of lost packets) voluntarily reduce their sending rate according to a given algorithm

The service architectures of the Internet

- Integrated Services (IntServ)
  - a reservation based approach: resources are reserved for each flow
  - reservation protocol RSVP (recipient initiated)
  - guaranteed service, controlled load service and best effort service
  - admission control: if the service cannot be provided, the request is denied
  - soft state (reservations are torn down unless they are updated)
  - necessitates keeping per flow state information; a hard task for the core routers
  - the approach is not considered to be scalable
  - may best suit to the access part of the network

The service architectures of the Internet (continued)

- Differentiated Services (DiffServ)
  - based on differentiating the service between flow aggregates (not individual flows)
  - traffic classification and marking is performed at the edge routers (the TOS field of an IP packet)
  - in association with this, the traffic is also measured, controlled and possibly shaped
  - the so called service profile of the customer defined bounds for allowed traffic

    the packets are classified: in-profile, out-of-profile (quality is assured for the in-profile traffic)
  - different aggregates are treated differently in the routers of the network:

    sharing the bandwidth and buffer resources (so called per hop behaviours, PHBs)
  - Expedited Forwarding (EF), Assured Forwarding (AF)

    within each class, the packets may have different loss priorities
  - no per flow state information, no signalling
  - the most complex classification and other operations take place at the edge routers
  - suits well for the core network
  - enables the ISPs to offer different service for different customer groups