Teletraffic theory
(for beginners)

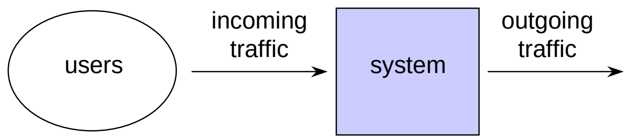
Samuli Aalto
samuli.aalto@hut.fi

Contents

- Purpose of Teletraffic Theory
- Network level: switching principles
- Telephone traffic models
- Data traffic models

Traffic point of view

- Telecommunication system from the **traffic point of view**:



- Ideas:
  - the **system serves** the incoming **traffic**
  - the traffic is generated by the **users** of the system

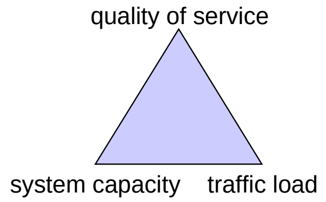
Teletraffic theory (for beginners) Samuli Aalto

Interesting questions

- Given the system and incoming traffic,
  what is the quality of service experienced by the user?
- Given the incoming traffic and required quality of service,
  how should the system be dimensioned?
- Given the system and required quality of service,
  what is the maximum traffic load?

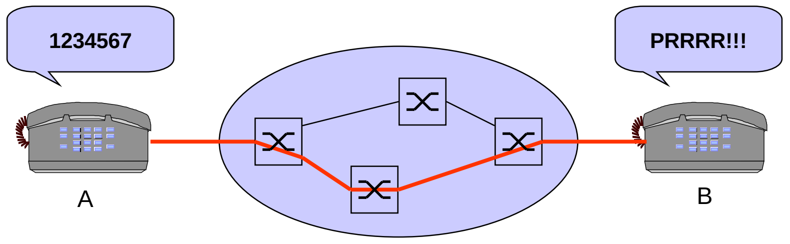
General purpose

- Determine **relationships** between the following three factors:
  - quality of service
  - traffic load
  - system capacity



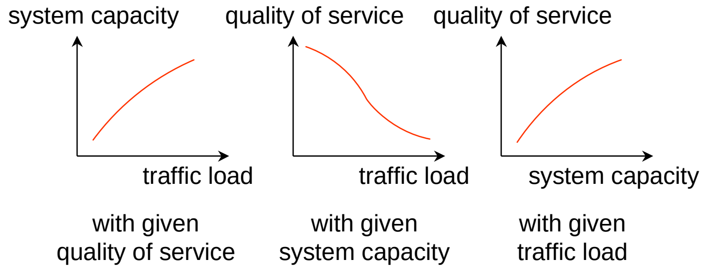
Example

- Telephone traffic
  - system = telephone network
  - traffic = telephone calls by everybody
  - quality of service = probability that the connection can be set up, i.e., “the line is not busy”



Relationships between the three factors

- Qualitatively, the relationships are as follows:



- To describe the relationships quantitatively, **mathematical models** are needed

Teletraffic models

- Teletraffic models are **stochastic** (= **probabilistic**)
  - systems themselves are usually deterministic but traffic is typically stochastic
  - “you never know, who calls you and when”
- It follows that the variables in these models are **random variables**, e.g.
  - number of ongoing calls
  - number of packets in a buffer
- Random variable is described by its **distribution**, e.g.
  - probability that there are n ongoing calls
  - probability that there are n packets in the buffer
- **Stochastic process** describes the temporal development of a random variable

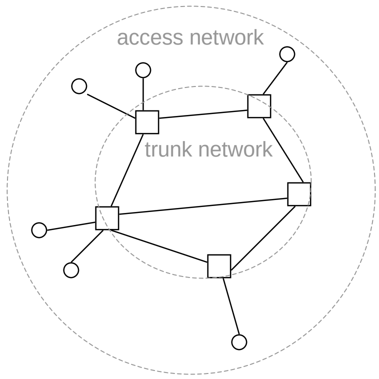
Practical goals

- Network planning
  - dimensioning
  - optimization
  - performance analysis
- Network management and control
  - efficient operating
  - fault recovery
  - traffic management
  - routing
  - accounting

Contents

Telecommunication network

- A simple model of a telecommunication network consists of
  - **nodes**
    - terminals ○
    - network nodes □
  - **links** between nodes
- **Access network**
  - connects the terminals to the network nodes
- **Trunk network**
  - connects the network nodes to each other

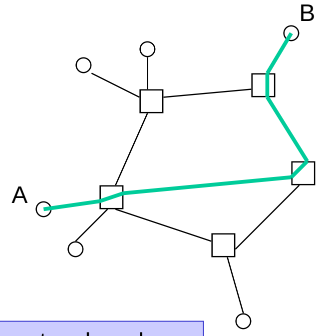


Switching modes

- **Circuit switching**
  - telephone networks
  - mobile telephone networks, e.g. GSM
- **Packet switching**
  - data networks
  - two possibilities
    - **connection oriented**: e.g. X.25, Frame Relay
    - **connectionless**: e.g. Internet (IP), SS7 (MTP)
- **Cell switching**
  - fast (connection oriented) packet switching with fixed length packets (called **cells**), e.g. ATM
  - integration of different traffic types (voice, data, video) $\Rightarrow$ multiservice networks

Circuit switching (1)

- **Connection oriented**:
  - connections **set up** end-to-end before information transfer
  - resources **reserved** for the whole duration of connection
  - e.g. telephone call reserves one (two-way) **channel** from each link along its route (time division multiplexing)
- Information transfer as **continuous stream**



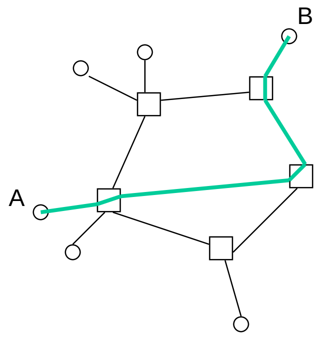
network node = (telephone) switch

Teletraffic theory (for beginners) Samuli Aalto

Circuit switching (2)

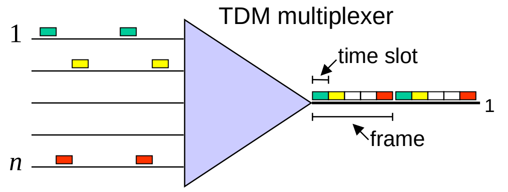
- Before information transfer
  - delay (to set up the connection)
- During information transfer
  - no overhead
  - no extra delays (besides the propagation delay)
- Efficient only if

connection holding time >> connection set up time



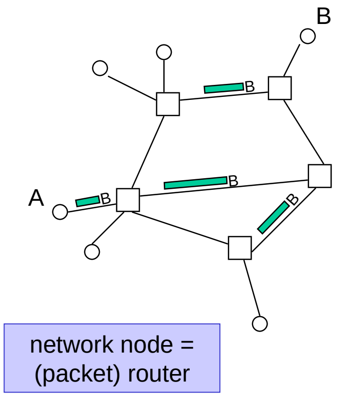
Time division multiplexing (TDM)

- Used in digital circuit switched systems
  - information conveyed on a link transferred in **frames** of fixed length
  - fixed portion (time slot) of each frame reserved for each channel
  - location of the time slot within the frame identifies the connection
- TDM multiplexer
  - input: n 1-channel physical connections
  - output: 1 n-channel physical connection



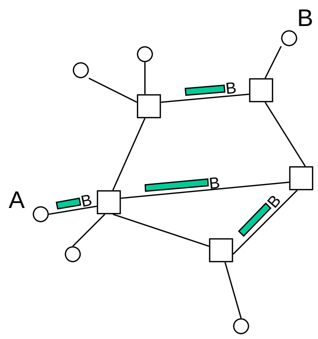
Connectionless packet switching (1)

- **Connectionless**:
  - no connection set-up
  - no resource reservation
- Information transfer as **discrete packets**
  - varying length
  - including header with global address (of the destination)
  - packets compete dynamically for processing capacity of nodes (next hop from routing table) and transmission capacity of links (**statistical multiplexing**)



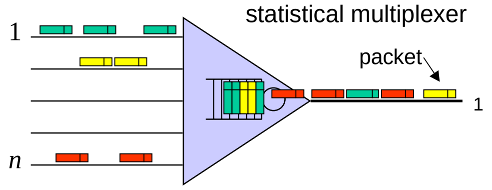
Connectionless packet switching (2)

- Before information transfer
  - no delays
- During information transfer
  - overhead (header bytes)
  - packet processing delays
  - packet transmission delays
  - queueing delays (since packets compete for joint resources)



Statistical multiplexing

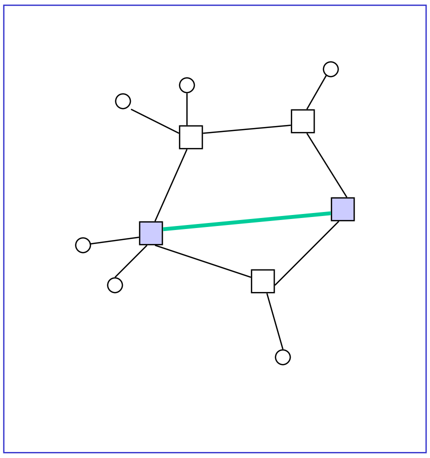
- Used in digital packet/cell switched systems, e.g. Internet, ATM
- Statistical multiplexer combines the packet flows of n incoming links to a joint outgoing link
  - capacity of the outgoing link reserved dynamically as packets arrive asynchronously and randomly $\Rightarrow$ need for buffering



Contents

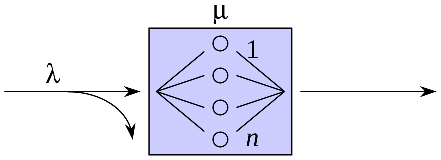
Classical model for telephone traffic (1)

- Loss models have traditionally been used to describe (circuit-switched) telephone networks
  - pioneering work made by Danish mathematician *A.K. Erlang* (1878-1929)
- Consider a link between two telephone exchanges
  - traffic consists of the ongoing telephone calls on the link

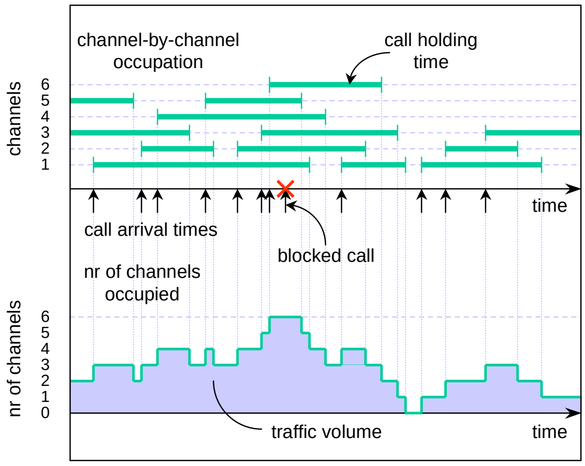


Classical model for telephone traffic (2)

- Erlang modelled this as a **loss system** with n servers
  - customer = (telephone) call
    - λ = call arrival rate
  - service time = (call) holding time
    - h = average holding time
  - server = channel on the link
    - n = number of parallel channels on the link



Traffic process



Traffic intensity

- In telephone networks:

Traffic ↔ Calls

- The amount of traffic is described by traffic intensity a
- By definition, **traffic intensity** a is
  the product of the arrival rate λ and the mean holding time h:

$$a = \lambda h$$

  - Note that the traffic intensity is a **dimensionless** quantity
  - Anyway, the unit of traffic intensity a is called **erlang**

Example

- Consider a local exchange. Assume that,
  - on the average, there are 1800 new calls in an hour, and
  - the mean holding time is 3 minutes
- It follows that the traffic intensity is

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

- If the mean holding time increases from 3 minutes to 10 minutes, then

$$a = 1800 * 10 / 60 = 300 \text{ erlang}$$

Blocking

- In a loss system some calls are lost
  - a call is lost if all n channels are occupied when the call arrives
  - the term **blocking** refers to this event
- There are (at least) two different types of blocking quantities:
  - **Call blocking** B_c = probability that an arriving call finds all n channels occupied = the fraction of calls that are lost
  - **Time blocking** B_t = probability that all n channels are occupied at an arbitrary time = the fraction of time that all n channels are occupied
- The two blocking quantities are not necessarily equal
  - If calls arrive according to a Poisson process, then $B_c = B_t$
- Call blocking is a better measure for the quality of service experienced by the subscribers but, typically, time blocking is easier to calculate

Teletraffic analysis

- System capacity
  - n = number of channels on the link
- Traffic load
  - a = (offered) traffic intensity
- Quality of service (from the subscribers' point of view)
  - B_c = probability that an arriving call finds all n channels occupied
- If we assume an **M/G/*n*/*n* loss system**, that is
  - calls arrive according to a **Poisson process** (with rate λ)
  - call holding times are independently and identically distributed according to **any distribution** with mean h
- Then the quantitive relation between the three factors is given by the **Erlang's blocking formula**

Erlang's blocking formula

$$B_c = \mathrm{Erl}(n,a) = \frac{\frac{a^n}{n!}}{\sum_{i=0}^{n} \frac{a^i}{i!}}$$

- Note: $n! = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1$
- Other names:
  - Erlang's formula
  - Erlang's B-formula
  - Erlang's loss formula
  - Erlang's first formula

Example

- Assume that there are n = 4 channels on a link and the offered traffic is a = 2.0 erlang. Then the call blocking probability B_c is

$$B_c = \mathrm{Erl}(4,2) = \frac{\frac{2^4}{4!}}{1+2+\frac{2^2}{2!}+\frac{2^3}{3!}+\frac{2^4}{4!}} = \frac{\frac{16}{24}}{1+2+\frac{4}{2}+\frac{8}{6}+\frac{16}{24}} = \frac{2}{21} \approx 9.5\%$$

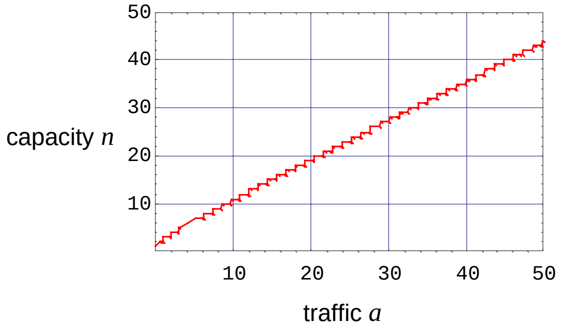
- If the link capacity is raised to n = 6 channels, B_c reduces to

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

Required capacity vs. traffic

- Given the quality of service requirement that $B_c < 20\%$, required capacity n depends on traffic intensity a as follows:

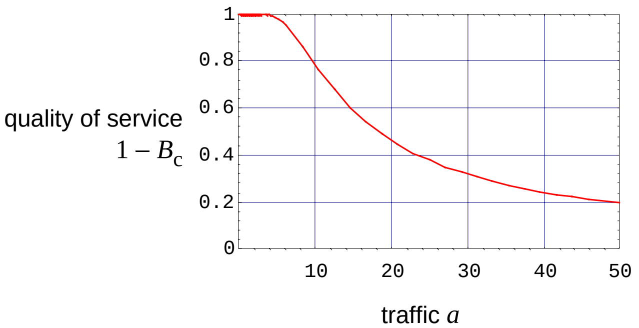
$$n(a) = \min\{N = 1,2,\ldots \mid \mathrm{Erl}(N,a) < 0.2\}$$



Required quality of service vs. traffic

- Given the capacity $n = 10$ channels, required quality of service $1 - B_c$ depends on traffic intensity a as follows:

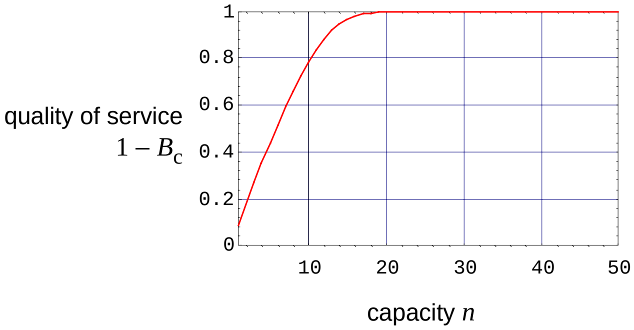
$$1 - B_c(a) = 1 - \mathrm{Erl}(10, a)$$



Required quality of service vs. capacity

- Given the traffic intensity $a = 10.0$ erlang, required quality of service $1 - B_c$ depends on capacity n as follows:

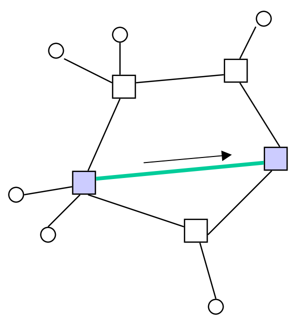
$$1 - B_c(n) = 1 - \mathrm{Erl}(n, 10.0)$$



Contents

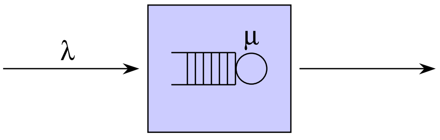
Classical model for data traffic (1)

- Queueing models are suitable for describing (packet-switched) data networks
  - pioneering work made by ARPANET researchers in 60's and 70's (e.g. *L. Kleinrock*)
- Consider a link between two packet routers
  - traffic consists of data packets transmitted on the link

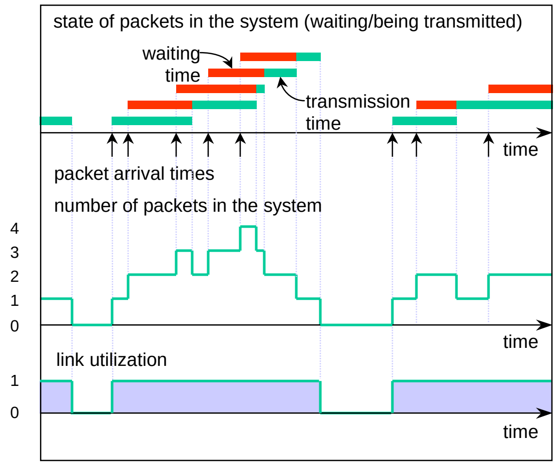


Classical model for data traffic (2)

- This can be modelled as a **waiting system** with a single server and an infinite buffer
  - customer = packet
    - λ = packet arrival rate
    - L = average packet length (data units)
  - server = link, waiting places = buffer
    - R = link's speed (data units per time unit)
  - service time = packet transmission time
    - $1/\mu = L/R$ = average packet transmission time



Traffic process



Traffic load

- In packet-switched data networks:

Traffic ↔ Packets

- The amount of traffic is described by traffic load ρ
- By definition, **traffic load** ρ is the quotient between the arrival rate λ and the service rate μ = R/L:

$$\rho = \frac{\lambda}{\mu} = \frac{\lambda L}{R}$$

  – Note that the traffic load is a **dimensionless** quantity
  – It can also be interpreted as the probability that the server is busy. So, it tells the **utilization factor** of the server

Example

- Consider a link between two packet routers. Assume that,
  - on the average, 10 new packets arrive in a second,
  - the mean packet length is 400 bytes, and
  - the link speed is 64 kbps.
- It follows that the traffic load is

$$\rho = 10 * 400 * 8 / 64{,}000 = 0.5 = 50\%$$

- If the link speed is increased up to 150 Mbps, the load is just

$$\rho = 10 * 400 * 8 / 150{,}000{,}000 = 0.0002 = 0.02\%$$

  - 1 byte = 8 bits
  - 1 kbps = 1 kbit/s = 1,000 bits per second
  - 1 Mbps = 1 Mbit/s = 1,000,000 bits per second

Teletraffic analysis

- System capacity
  - R = link speed in kbps
- Traffic load
  - λ = packet arrival rate in packet/s (considered here as a variable)
  - L = average packet length in kbits (assumed here that L = 1 kbit)
- Quality of service (from the users' point of view)
  - P_z = probability that a packet has to wait "too long", i.e., longer than a given reference value z (assumed here that P_z = 0.1 s)
- If we assume an **M/M/1 queueing system**, that is
  - packets arrive according to a Poisson process (with rate λ)
  - packet lengths are independent and identically distributed according to **exponential** distribution with mean L
- Then the quantitive relation between the three factors is given by the following waiting time formula

Waiting time formula for an M/M/1 queue

$$P_z = \mathrm{Wait}(R,\lambda;L,z) = \begin{cases} \frac{\lambda L}{R}\exp(-(\frac{R}{L}-\lambda)z), & \text{if } \lambda L < R \ (\rho < 1) \\ 1, & \text{if } \lambda L \geq R \ (\rho \geq 1) \end{cases}$$

- Note:
  - The system is **stable** only in the former case ($\rho < 1$). Otherwise the queue builds up without limits.

Example

- Assume that packets arrive at rate $\lambda = 50$ packet/s and the link speed is $R = 64$ kbps. Then the probability P_z that an arriving packet has to wait too long (i.e., longer than $z = 0.1$ s) is

$$P_z = \mathrm{Wait}(64,50;1,0.1) = \frac{50}{64}\exp(-1.4)) \approx 19\%$$

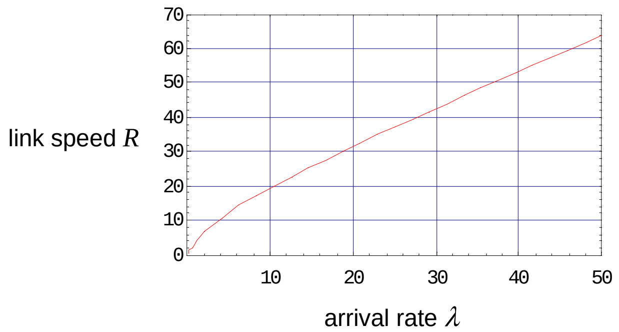
- Note that the system is stable, since

$$\rho = \frac{\lambda L}{R} = \frac{50}{64} < 1$$

Required link speed vs. arrival rate

- Given the quality of service requirement that $P_z < 20\%$, required link speed R depends on arrival rate λ as follows:

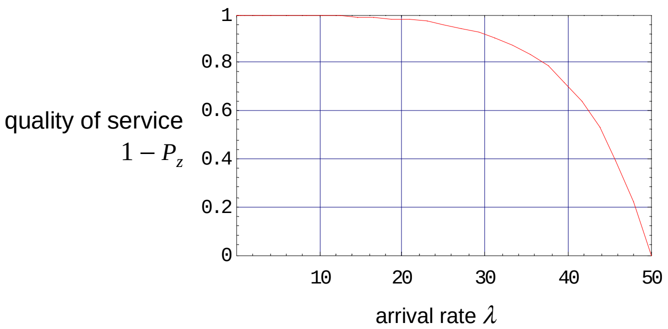
$$R(\lambda) = \min\{r > \lambda L \mid \mathrm{Wait}(r,\lambda;1,0.1) < 0.2\}$$



Required quality of service vs. arrival rate

- Given the link speed R = 50 kbps, required quality of service $1 - P_z$ depends on arrival rate λ as follows:

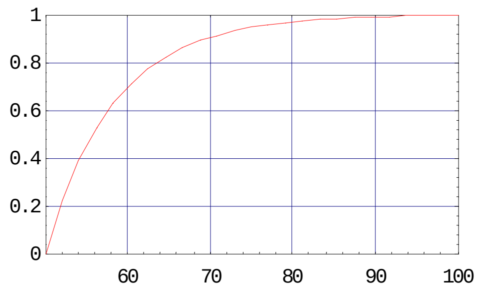
$$1 - P_z(\lambda) = 1 - \mathrm{Wait}(50,\lambda;1,0.1)$$



Required quality of service vs. link speed

- Given the arrival rate $\lambda = 50$ packet/s, required quality of service $1 - P_z$ depends on link speed R as follows:

$$1 - P_z(R) = 1 - \text{Wait}(R, 50; 1, 0.1)$$



THE END