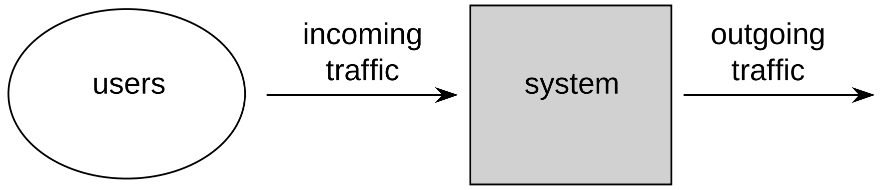
1. Introduction

Contents

- Purpose of Teletraffic Theory
- Teletraffic models
- Classical model for telephone traffic
- Classical model for data traffic

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

Interesting questions

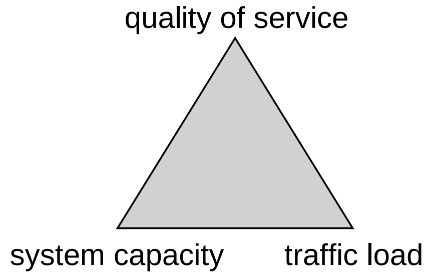
- Given the system and incoming traffic,
  what is the quality of service experienced by the user?

- 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?

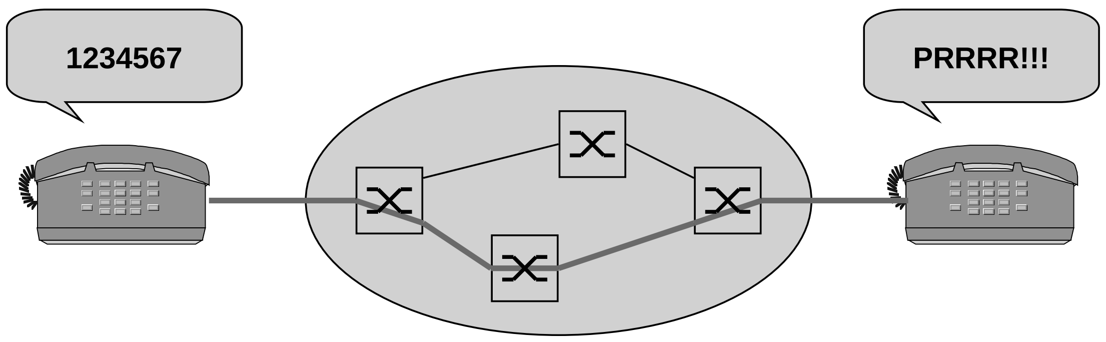
General purpose

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



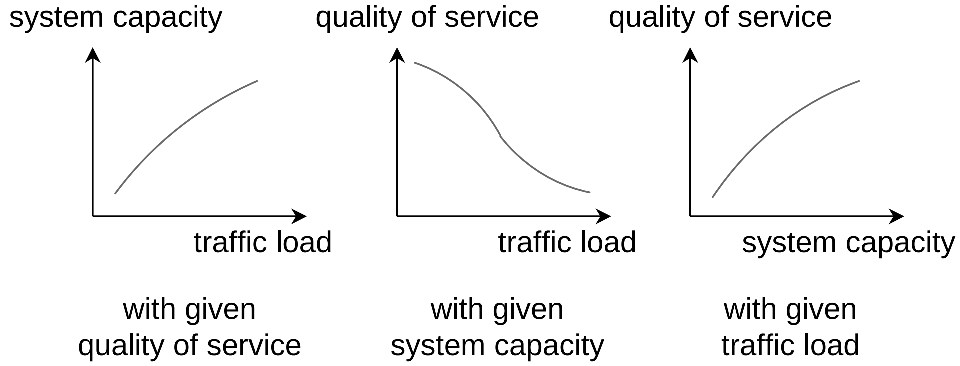
Example

- Telephone call
  - traffic = telephone calls by everybody
  - system = telephone network
  - quality of service = probability that the phone rings at the destination



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

Related fields

- Probability Theory
- Stochastic Processes
- Queueing Theory
- Statistical Analysis (traffic measurements)
- Operations Research
- Optimization Theory
- Decision Theory (Markov decision processes)
- Simulation Techniques (object oriented programming)

Difference between the real system and the model

- Typically,
  - the model describes just one part or property of the real system under consideration and even from one point of view
  - the description is not very accurate but rather approximative
- Thus,
  - caution is needed when conclusions are drawn

Practical goals

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

Literature

- Teletraffic Theory
  - Teletronikk (1995) Vol. 91, Nr. 2/3, Special Issue on “Teletraffic”
  - S-38.118 course book: “Understanding Telecommunications 1”, Ch. 10
  - COST 242, Final report (1996) “Broadband Network Teletraffic”, Eds. J. Roberts, U. Mocci, J. Virtamo, Springer
  - J.M. Pitts and J.A. Schormans (1996) “Introduction to ATM Design and Performance”, Wiley
- Queueing Theory
  - L. Kleinrock (1975) “Queueing Systems, Volume I: Theory”, Wiley
  - L. Kleinrock (1976) “Queueing Systems, Volume II: Computer Applications”, Wiley
  - D. Bertsekas and R. Gallager (1992) “Data Networks”, 2nd ed., Prentice-Hall
  - P.G. Harrison and N.M. Patel (1993) “Performance Modelling of Communication Networks and Computer Architectures”, Addison-Wesley

Contents

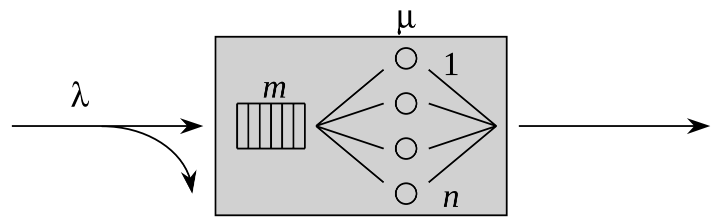
- Purpose of the Teletraffic Theory
- Teletraffic models
- Classical model for telephone traffic
- Classical model for data traffic

Teletraffic models

- Two phases in modelling:
  - modelling of the incoming traffic ⇒ **traffic model**
  - modelling of the system itself ⇒ **system model**
- Two types of system models:
  - loss systems
  - waiting/queueing systems
- These models can be combined to create models for whole telecommunication networks
  - loss network models
  - queueing network models
- Next we will present a simple teletraffic model
  - describing a single resource

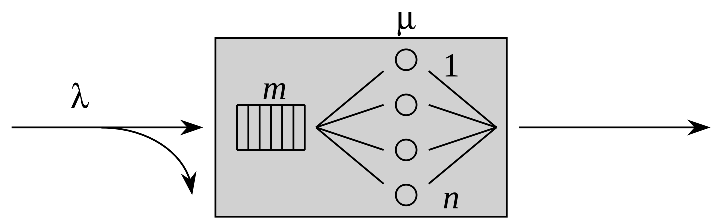
Simple teletraffic model

- **Customers arrive** at rate λ (customers per time unit)
  - $1/\lambda$ = average inter-arrival time
- Customers are **served** by n parallel **servers**
- When busy, a server serves at rate μ (customers per time unit)
  - $1/\mu$ = average service time of a customer
- There are m **waiting** places
- It is assumed that blocked customers (arriving in a full system) are lost



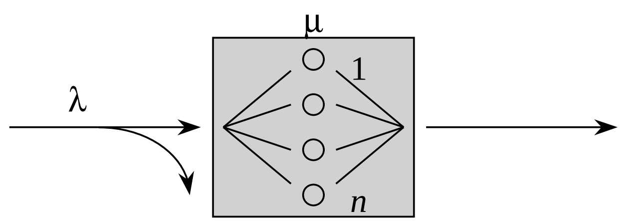
Exercise

- Consider the simple teletraffic model presented above
  - What is the traffic model?
  - What is the system model?



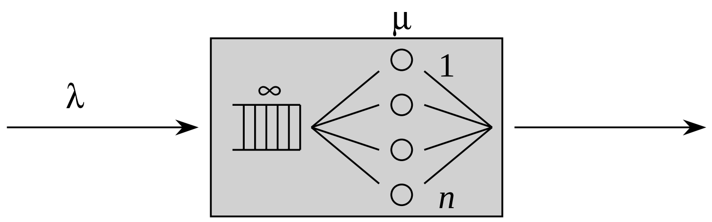
Pure loss system

- No waiting places ($m = 0$)
  - If the system is full (with all n servers occupied) when a customer arrives, she is not served at all but lost
  - Some customers are lost
- From the customer's point of view, it is interesting to know e.g.
  - What is the probability that the system is full when she arrives?
- From the system's point of view, it is interesting to know e.g.
  - What is the utilization factor of the servers?



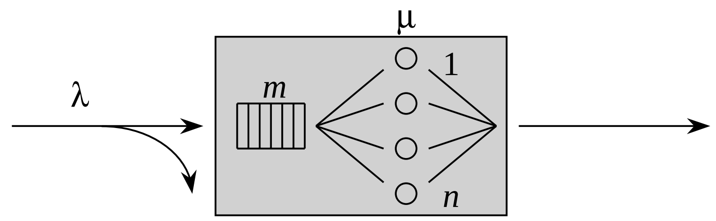
Pure waiting system

- Infinite number of waiting places ($m = \infty$)
  - If all n servers are occupied when a customer arrives, she occupies one of the waiting places
  - No customers are lost but some of them have to wait before getting served
- From the customer's point of view, it is interesting to know e.g.
  - what is the probability that she has to wait "too long"?
- From the system's point of view, it is interesting to know e.g.
  - what is the utilization factor of the servers?



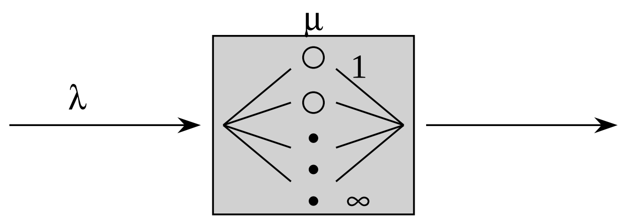
Mixed system

- Finite number of waiting places ($0 < m < \infty$)
  - If all n servers are occupied but there are free waiting places when a customer arrives, she occupies one of the waiting places
  - If all n servers and all m waiting places are occupied when a customer arrives, she is not served at all but lost
  - Some customers are lost and some customers have to wait before getting served



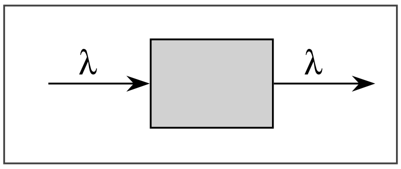
Infinite system

- Infinite number of servers ($n = \infty$)
  - No customers are lost or even have to wait before getting served
- Sometimes,
  - this hypothetical model can be used to get some approximate results for a real system (with finite system capacity)
- Always,
  - it gives bounds for the performance of a real system (with finite system capacity)
  - it is much easier to analyze than the corresponding finite capacity models



Little's formula

- Consider a system where
  - new customers arrive at rate λ
- Assume **stability**:
  - Every now and then, the system is empty
- Consequence:
  - Customers depart from the system at rate λ



- Let

$$\overline{N} = \text{average nr of customers in the system}$$

$$\overline{T} = \text{average time a customer spends in the system}$$

- **Little's formula**:

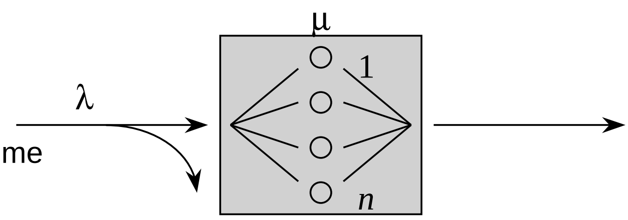
$$\overline{N} = \lambda \overline{T}$$

Contents

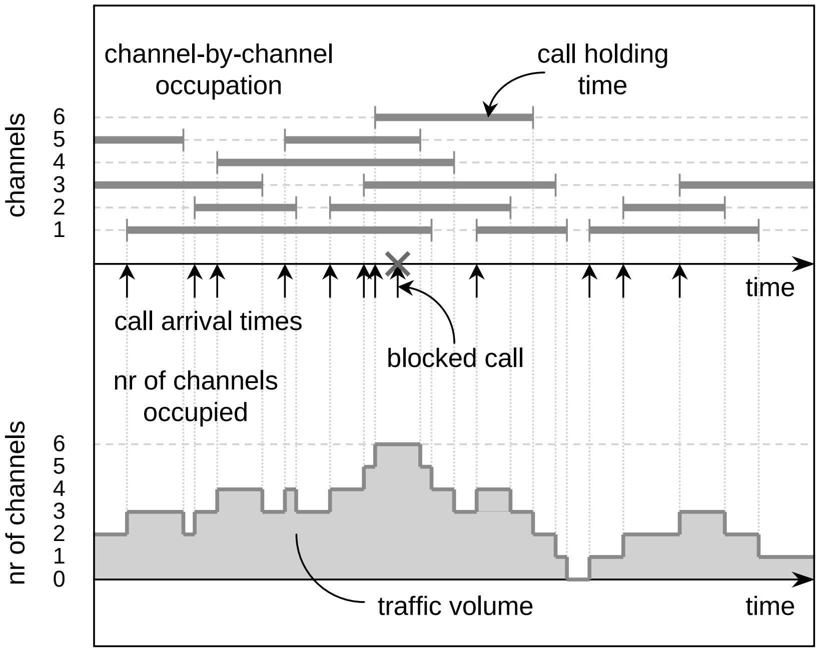
- Purpose of the Teletraffic Theory
- Teletraffic models
- Classical model for telephone traffic
- Classical model for data traffic

Classical model for telephone traffic

- 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
- Erlang modelled this as a **pure loss system** ($m = 0$)
  - customer = call
    - λ = call arrival rate
  - service time = (call) holding time
    - $h = 1/\mu$ = average holding time
  - server = channel on the link
    - n = nr of channels on the link



Traffic process



Traffic intensity

- In telephone networks:

Traffic ↔ Calls

- The amount of traffic is described by the **traffic intensity** a
- By definition, the 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 the traffic intensity a is called **erlang** (**erl**)
  - traffic of one erlang means that, on the average, one channel is occupied

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

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

Characteristic traffic

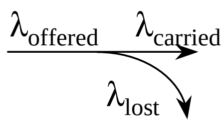
- Here are typical characteristic traffics for some subscriber categories (of ordinary telephone users):
  - private subscriber: 0.01 - 0.04 erlang
  - business subscriber: 0.03 - 0.06 erlang
  - private branch exchange (PBX): 0.10 - 0.60 erlang
  - pay phone: 0.07 erlang
- This means that, for example,
  - a typical private subscriber uses from 1% to 4% of her time in the telephone (during so called “busy hour”)
- Referring to the previous example, note that
  - it takes between 2250 - 9000 private subscribers to generate 90 erlang traffic

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

Call rates

- In a loss system each call is either **lost** or **carried**
- Thus, there are three types of call rates:
  - λ_{offered} = arrival rate of all call attempts
  - λ_{carried} = arrival rate of carried calls
  - λ_{lost} = arrival rate of lost calls
- Note:



$$\lambda_{\text{offered}} = \lambda_{\text{carried}} + \lambda_{\text{lost}} = \lambda$$
$$\lambda_{\text{carried}} = \lambda(1 - B_{\text{c}})$$
$$\lambda_{\text{lost}} = \lambda B_{\text{c}}$$

Traffic streams

- The three call rates lead to the following three traffic concepts:
  - **Traffic offered** $a_{\text{offered}} = \lambda_{\text{offered}} h$
  - **Traffic carried** $a_{\text{carried}} = \lambda_{\text{carried}} h$
  - **Traffic lost** $a_{\text{lost}} = \lambda_{\text{lost}} h$
- Note:

$$a_{\text{offered}} = a_{\text{carried}} + a_{\text{lost}} = a$$
$$a_{\text{carried}} = a(1 - B_{\text{c}})$$
$$a_{\text{lost}} = a B_{\text{c}}$$

- Traffic offered and traffic lost are hypothetical quantities, but traffic carried is **measurable** (key: Little’s formula):
  - Traffic carried = the average number of occupied channels on the link

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

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

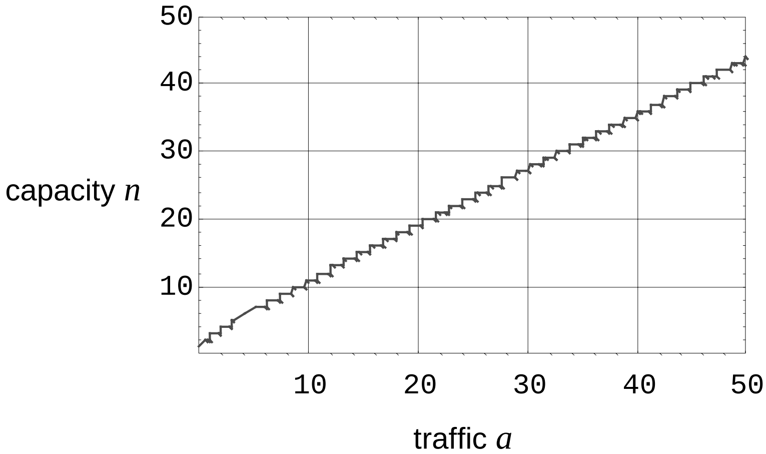
- If the link capacity is raised to n = 6 channels, then B_c reduces to

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

Required capacity vs. traffic

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

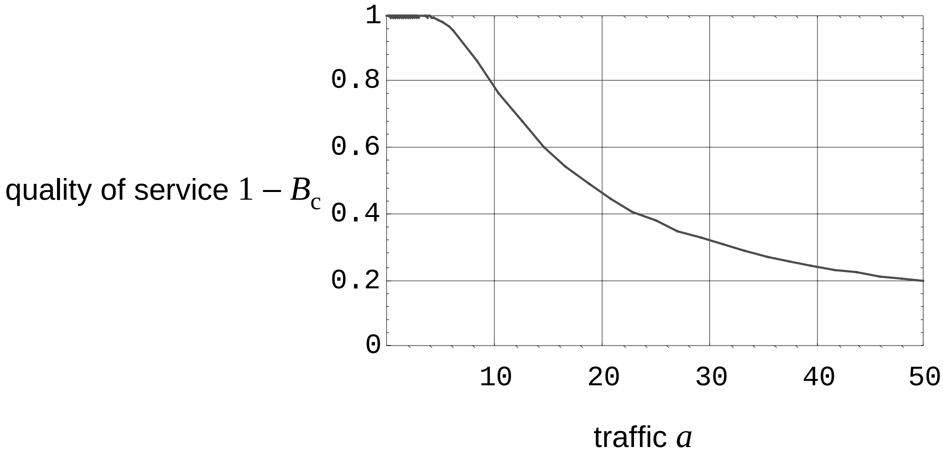
$$n(a) = \min\{N = 1,2,\ldots \mid \text{Erl}(N,a) < 0.2\}$$



Required quality of service vs. traffic

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

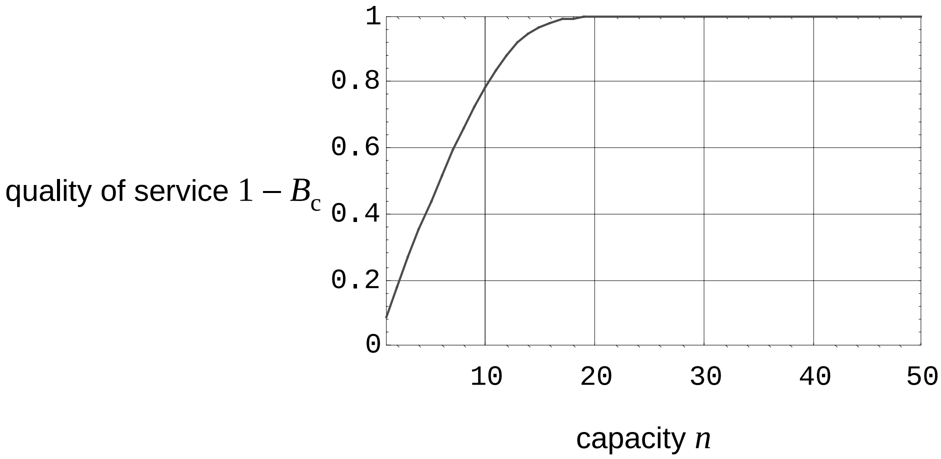
$$1 - B_c(a) = 1 - \mathrm{Erl}(10, a)$$



Required quality of service vs. capacity

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

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

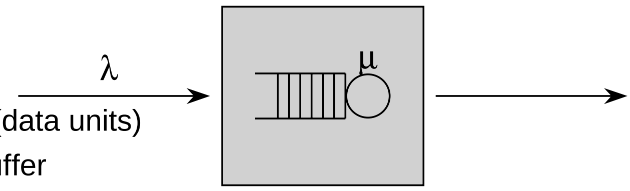


Contents

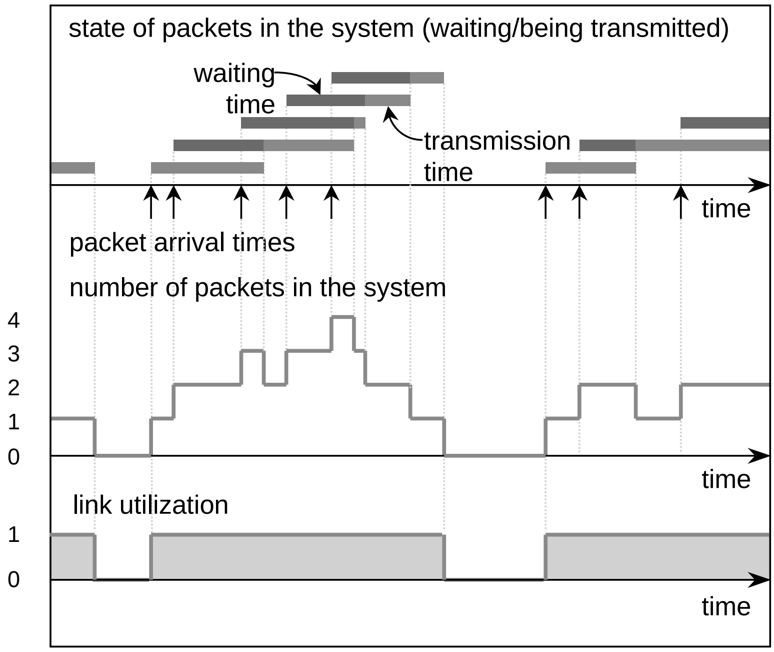
- Purpose of the Teletraffic Theory
- Teletraffic models
- Classical model for telephone traffic
- Classical model for data traffic

Classical model for data traffic

- Queueing models are suitable for describing (packet-switched) data networks
  - Pioneering work made by many people in 60’s and 70’s (ARPANET)
- Consider a link between two packet routers
  - traffic consists of data packets transmitted on the link
- This can be modelled as a **pure waiting system** with a single server (n = 1) and an infinite buffer ($m = \infty$)
  - 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 the **traffic load** ρ
- By definition, the 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 called the **traffic intensity** (as in loss systems)
  - By Little’s formula, 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 to 150 Mbps, then the load is just

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

- Note:
  - 1 byte = 8 bits
  - 1 kbps = 1 kbit/s = 1 kbit per second = 1,000 bits per second
  - 1 Mbps = 1 Mbit/s = 1 Mbit per second = 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 to be constant 1 kbit)
- Quality of service (from the users' point of view)
  - P_z = probability that a packet has to wait “too long”, that is longer than a given reference value z (assumed here to be constant 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 ($\rho < 1$)

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

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

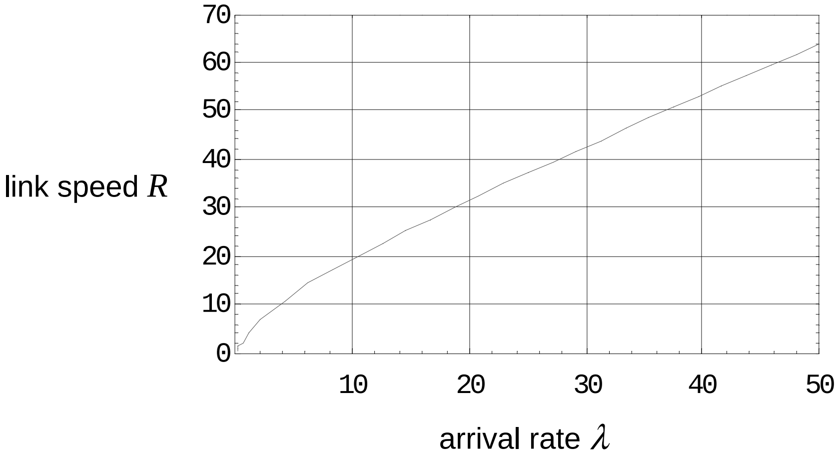
- Note that the system is stable, since

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

Required link speed vs. arrival rate

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

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

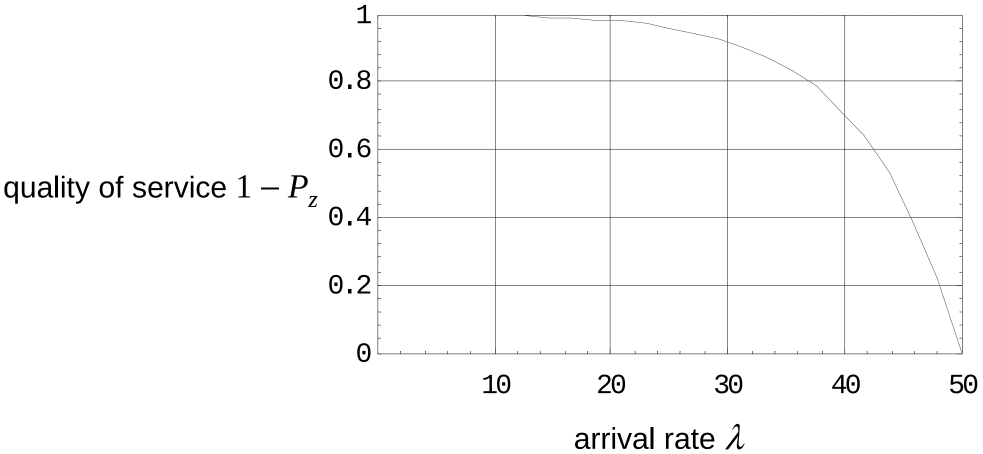


1. Introduction

Required quality of service vs. arrival rate

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

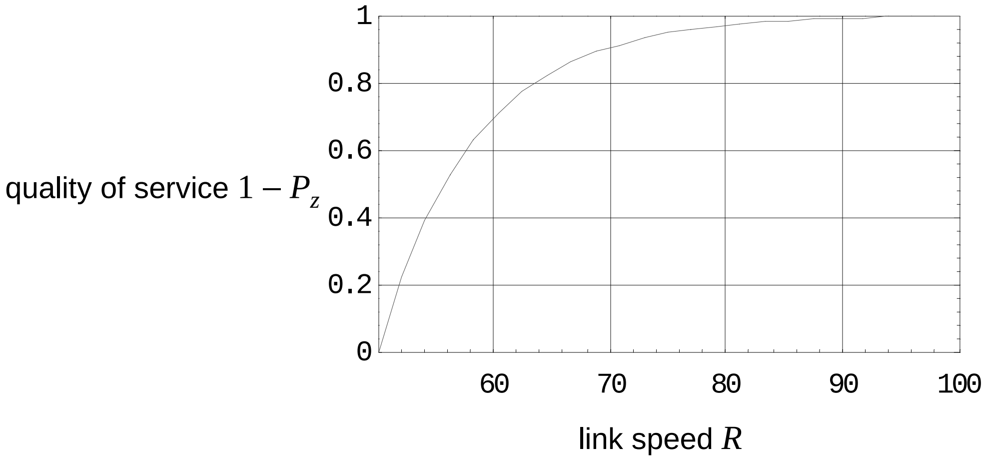
$$1 - P_Z(\lambda) = 1 - \mathrm{Wait}(50, \lambda; 1, 0.1)$$



Required quality of service vs. link speed

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

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



THE END