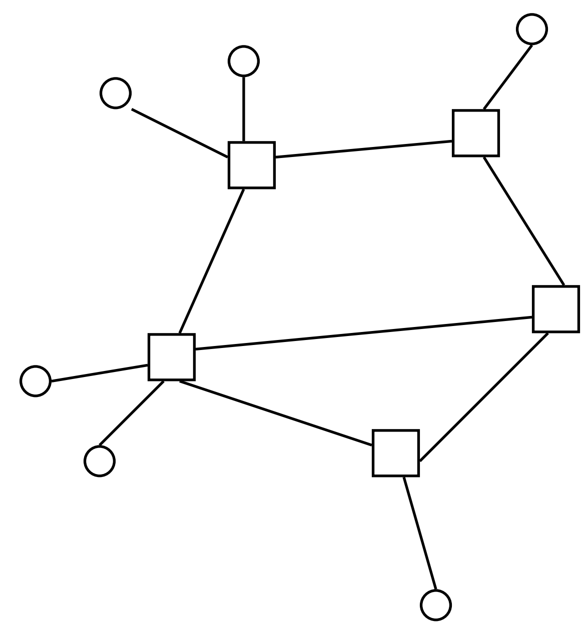
1. Introduction

Contents

- Telecommunication networks and switching modes
- Purpose of Teletraffic Theory
- Teletraffic models
- Little’s formula

Telecommunication network

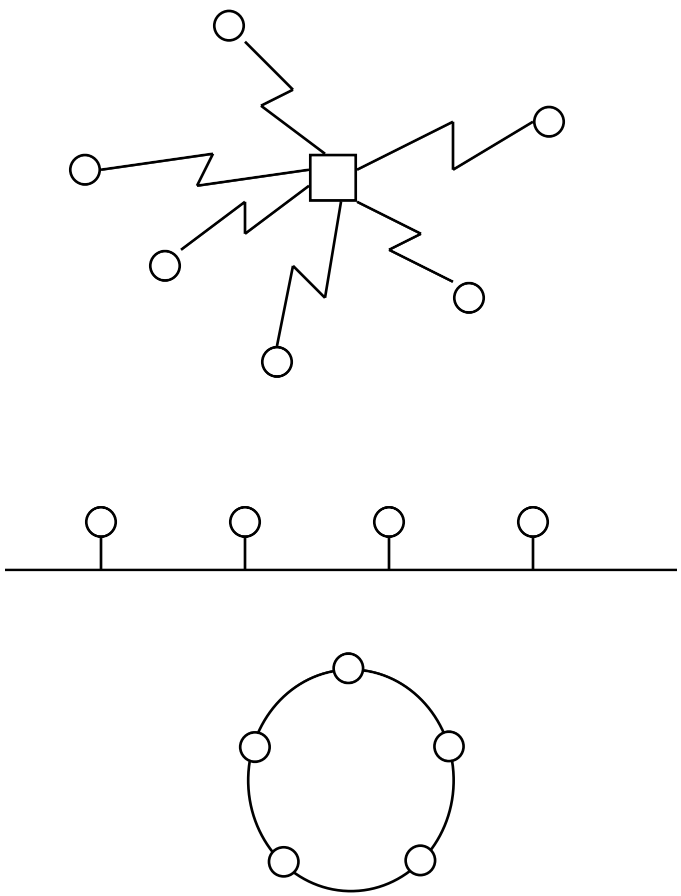
- A simple model of a telecommunication network consists of
  - **nodes**
    - terminals
    - network nodes
  - **links** between nodes
- **Access network**
  - connects the terminals to the network nodes
- **Trunk network**
  - connects the network nodes to each other



Shared medium as an access network

- In the previous model,
  - connections between terminals and network nodes are **point-to-point** type ($\Rightarrow$ no resource sharing within the access netw.)
- In some cases, such as
  - mobile telephone network
  - local area network (LAN) connecting computers

  the access network consists of **shared medium**:
  - users have to **compete** for the resources of this shared medium
  - **multiple access** (MA) techniques are needed

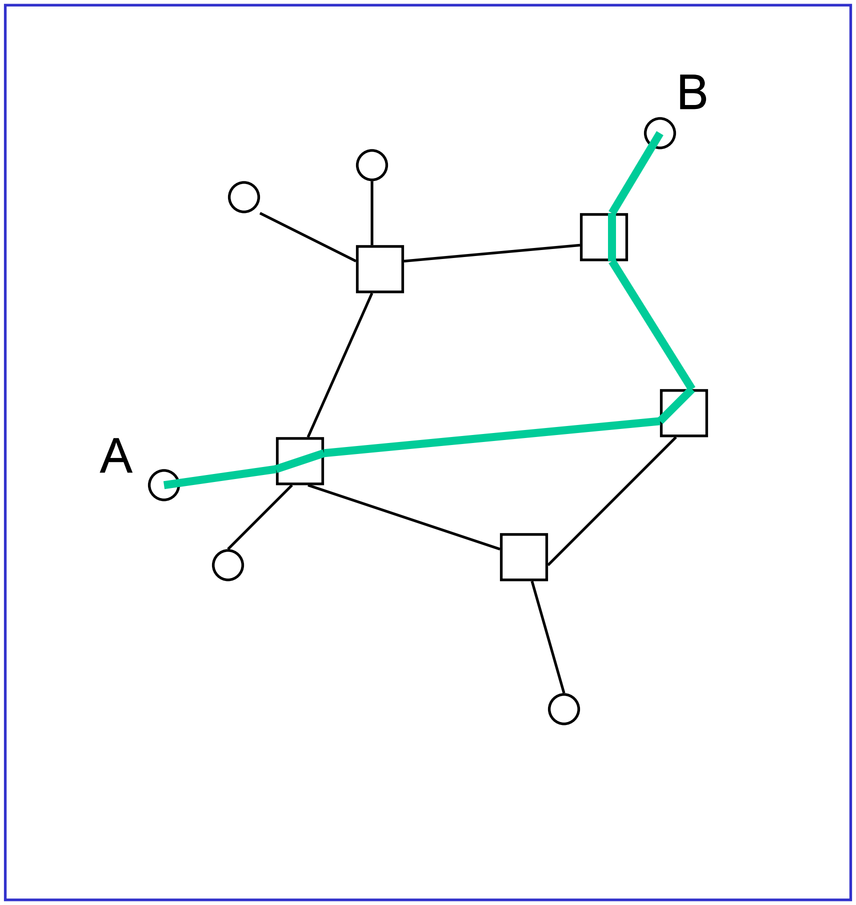


Switching modes

- **Circuit switching**
  - telephone networks
  - mobile telephone networks
  - optical networks
- **Packet switching**
  - data networks
  - two possibilities
    - **connection oriented**: e.g. X.25, Frame Relay
    - **connectionless**: e.g. Internet (IP), SS7 (MTP)
- **Cell switching**
  - ATM networks
  - connection oriented
  - fast packet switching with fixed length packets (cells)

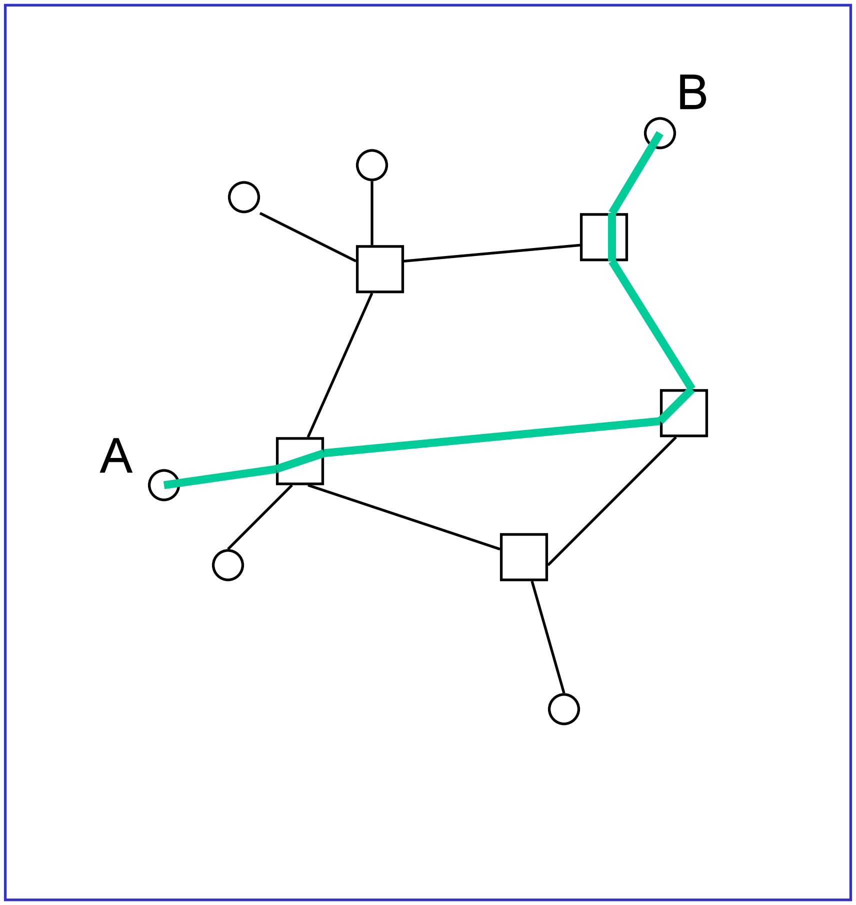
Circuit switching (1)

- **Connection oriented**:
  - connections **set up** end-to-end before information transfer
  - resources **reserved** for the whole duration of connection
  - if resources are not available, the call is blocked and lost
- Information transfer as **continuous stream**



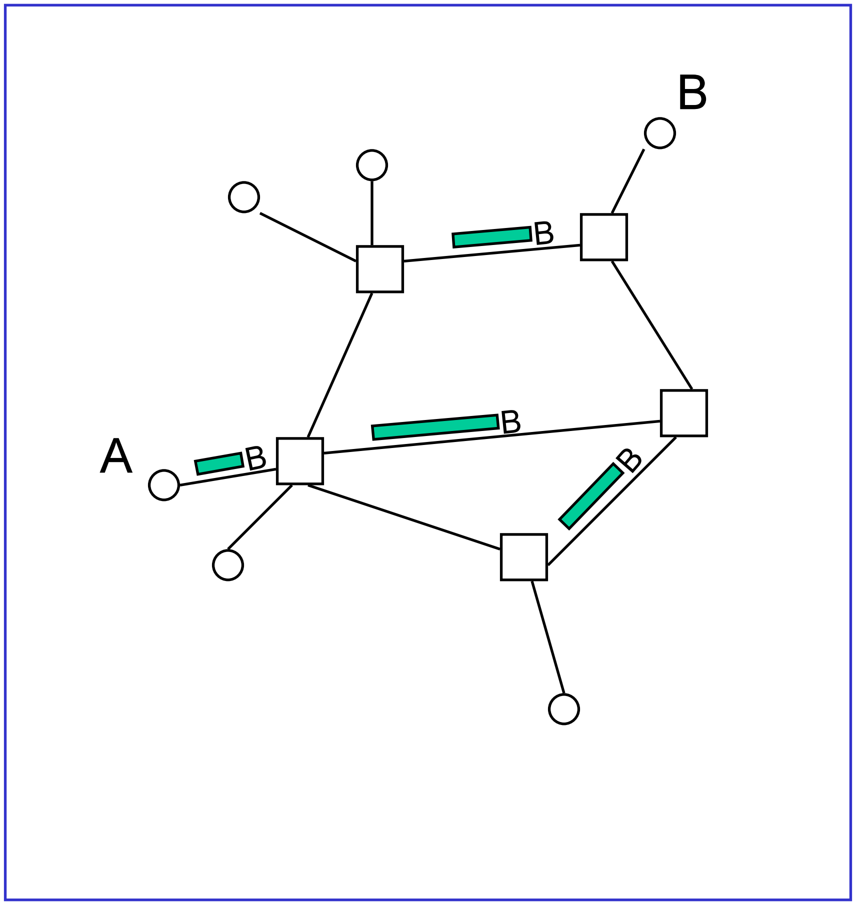
Circuit switching (2)

- Before information transfer
  - Set-up delay
- During information transfer
  - signal propagation delay
  - no overhead
  - no extra delays
- Example: telephone network



Connectionless packet switching (1)

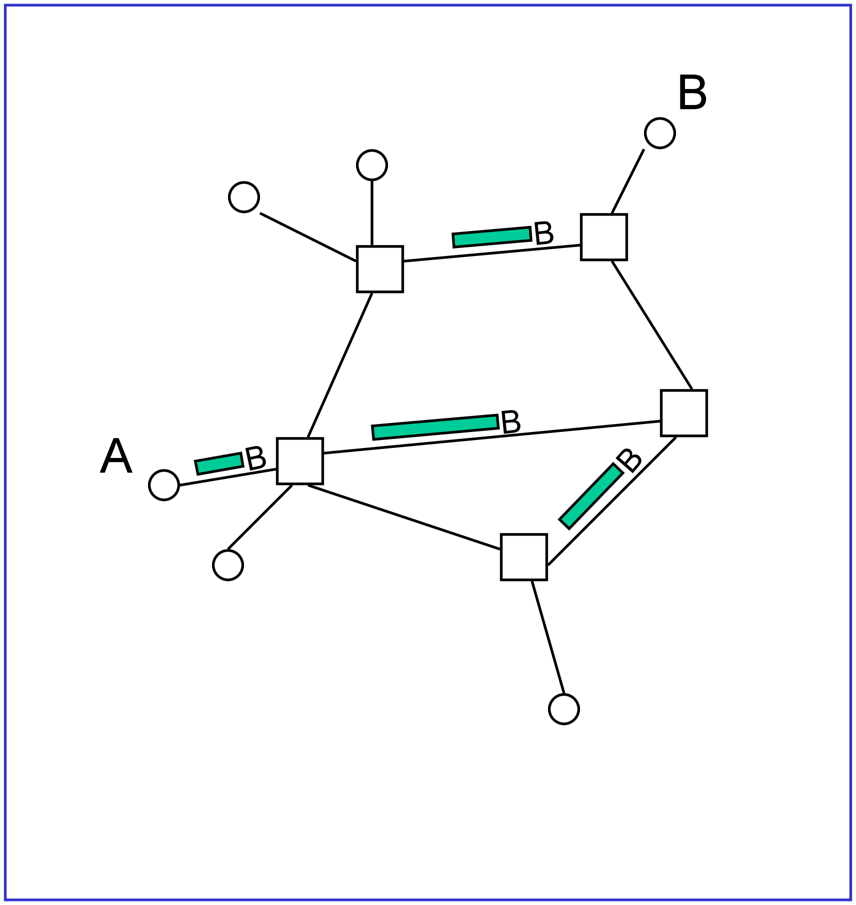
- **Connectionless**:
  - no connection set-up
  - no resource reservation
  - no blocking
- Information transfer as **discrete packets**
  - varying length
  - global address (of the destination)



Connectionless packet switching (2)

- Before information transfer
  - no delays
- During information transfer
  - overhead (header bytes)
  - packet processing delays
  - queueing delays (since packets compete for joint resources)
  - transmission delays (due to finite capacity links)
  - signal propagation delay
  - packet losses (due to finite buffers)

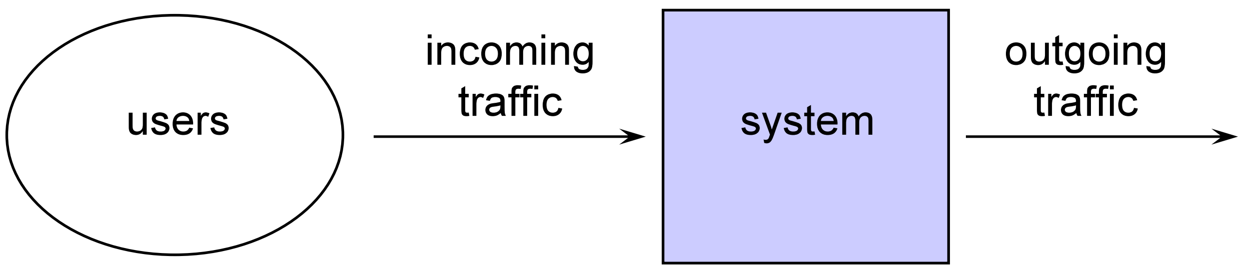
- Example: Internet (IP-layer)



Contents

Traffic point of view

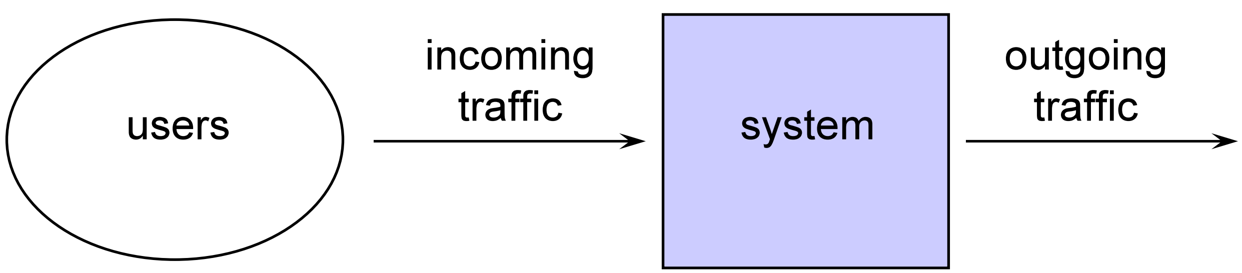
- Telecommunication system from the **traffic point of view**:



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

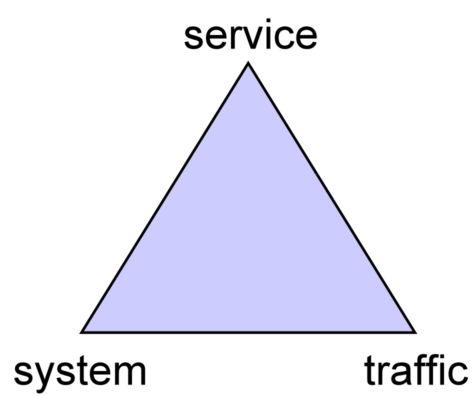
Interesting questions

- 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 (1)

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

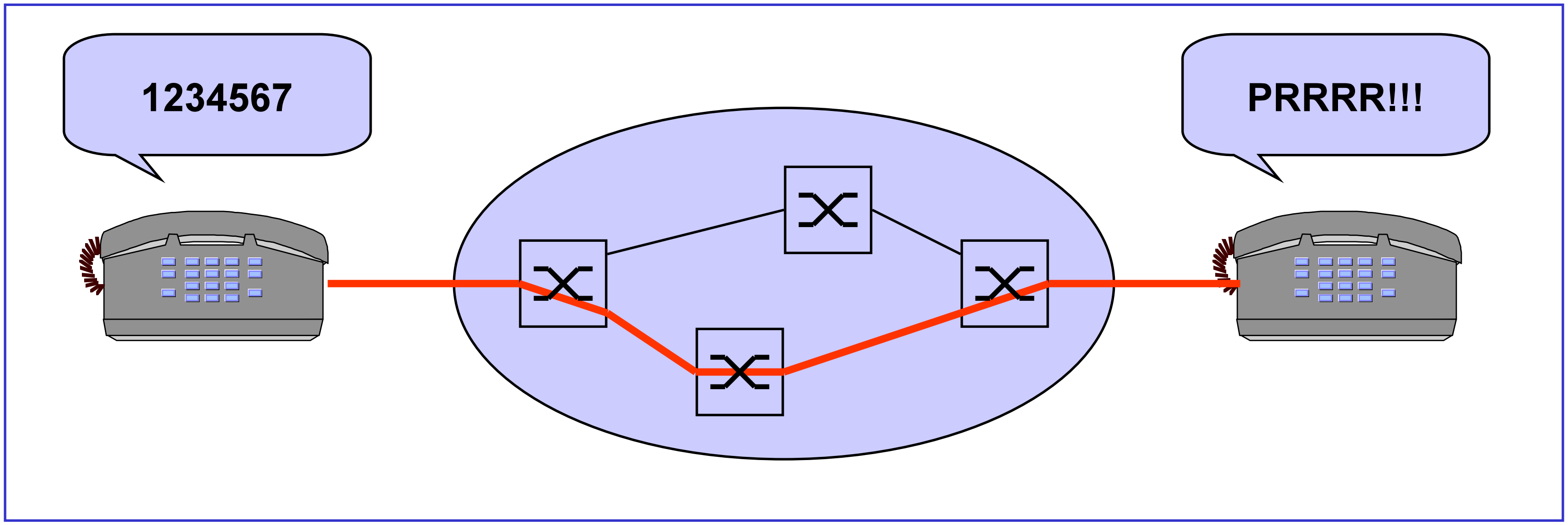


General purpose (2)

- System can be
  - a single device (e.g. link between two telephone exchanges, link in an IP network, packet processor in a data network, router’s transmission buffer, or statistical multiplexer in an ATM network)
  - the whole network (e.g. telephone or data network) or some part of it
- Traffic consists of
  - bits, packets, bursts, flows, connections, calls, …
  - depending on the system and time scale considered
- Quality of service can be described from the point of view of
  - the customer (e.g. call blocking, packet loss, packet delay, or throughput)
  - the system, in which case we use the term **performance** (e.g. processor or link utilization, or maximum network load)

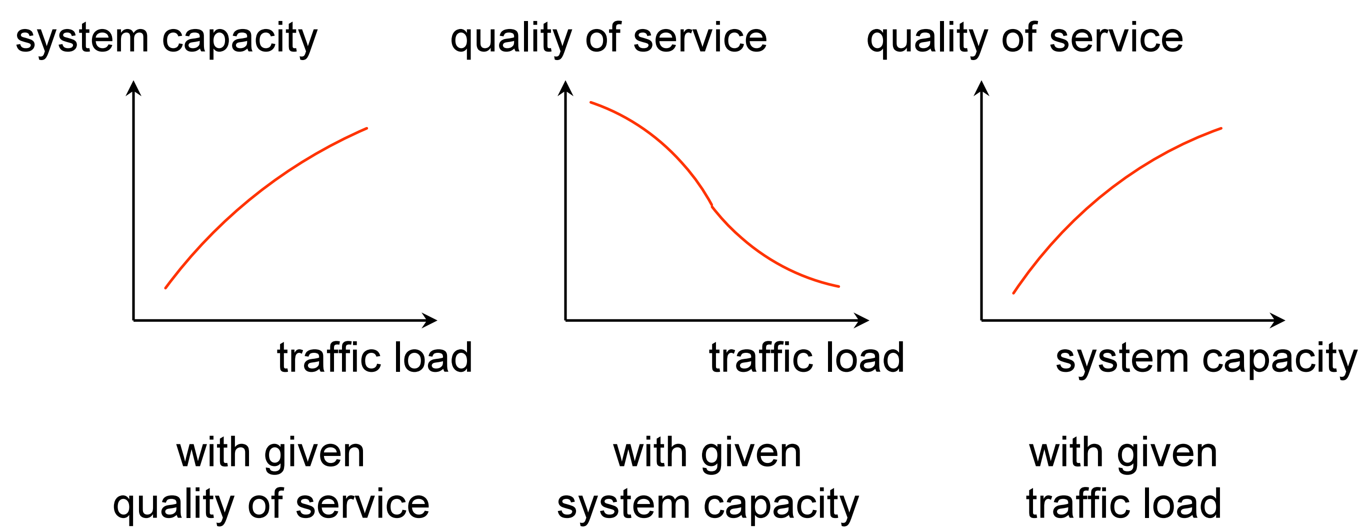
Example

- 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

Real system vs. 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* Vol. 91, Nr. 2/3, Special Issue on “Teletraffic”, 1995
  - V. B. Iversen, *Teletraffic Engineering Handbook*,
    http://www.tele.dtu.dk/teletraffic/handbook/telehook.pdf
  - J. Roberts, *Traffic Theory and the Internet*,
    IEEE Communications Magazine, Jan. 2001, pp. 94-99
    http://perso.rd.francetelecom.fr/roberts/Pub/Rob01.pdf
- Queueing Theory
  - L. Kleinrock, *Queueing Systems, Vol. I: Theory*, Wiley, 1975
  - L. Kleinrock, *Queueing Systems, Vol. II: Computer Applications*, Wiley, 1976
  - D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed., Prentice-Hall, 1992
  - Myron Hlynka's Queueing Theory Page
    http://www2.uwindsor.ca/~hlynka/queue.html

Contents

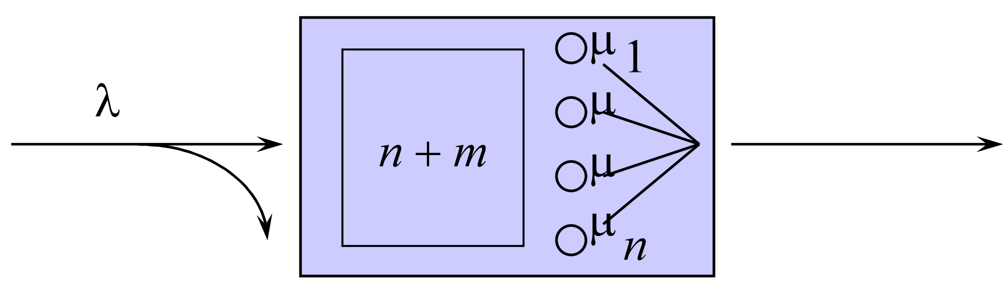
- Telecommunication networks and switching modes
- Purpose of Teletraffic Theory
- Teletraffic models
- Little’s formula

Teletraffic model types

- Three types of system models:
  - **loss systems**
  - **queueing systems**
  - **sharing systems**
- Next we will present simple teletraffic models
  - describing a single resource
- These models can be combined to create models for whole telecommunication networks
  - loss networks
  - queueing networks
  - sharing networks

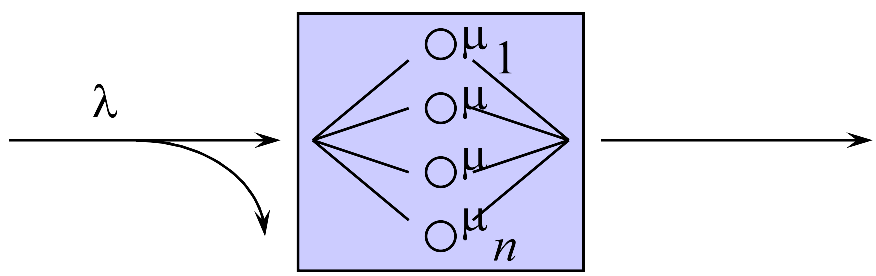
Simple teletraffic model

- **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 $n + m$ **customer places** in the system
  - at least n **service places** and at most m **waiting places**
- It is assumed that blocked customers (arriving in a full system) are lost



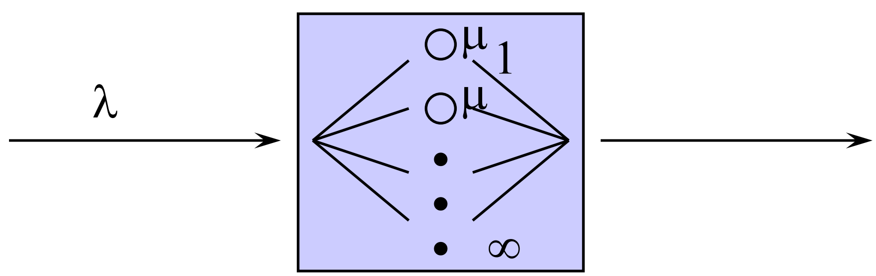
Pure loss system

- Finite number of servers ($n < \infty$), n service places, no waiting places ($m = 0$)
  - If the system is full (with all n servers occupied) when a customer arrives, it is not served at all but lost
  - Some customers may be 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 it arrives?



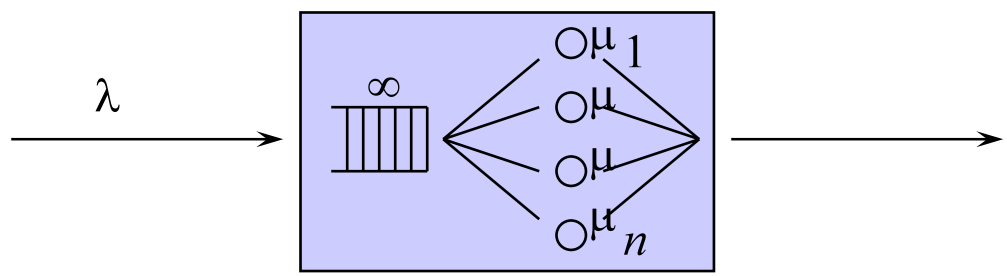
Infinite system

- Infinite number of servers ($n = \infty$), no waiting places ($m = 0$)
  - 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



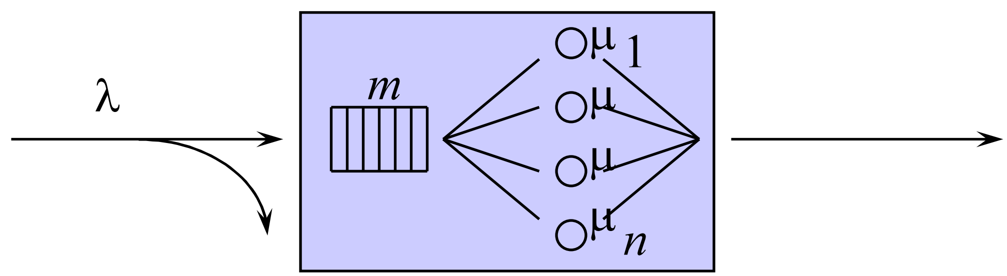
Pure queueing system

- Finite number of servers ($n < \infty$), n service places, infinite number of waiting places ($m = \infty$)
  - If all n servers are occupied when a customer arrives, it 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 it has to wait "too long"?



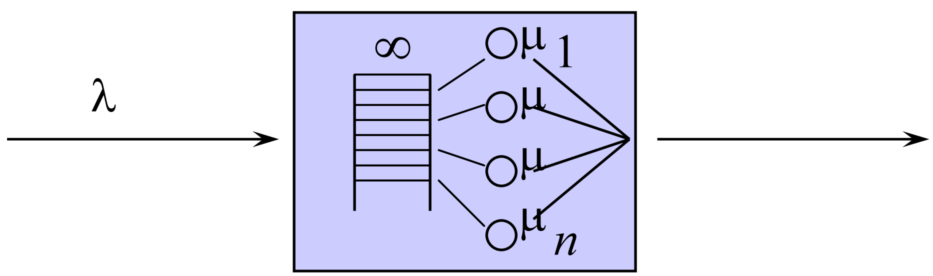
Lossy queueing system

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



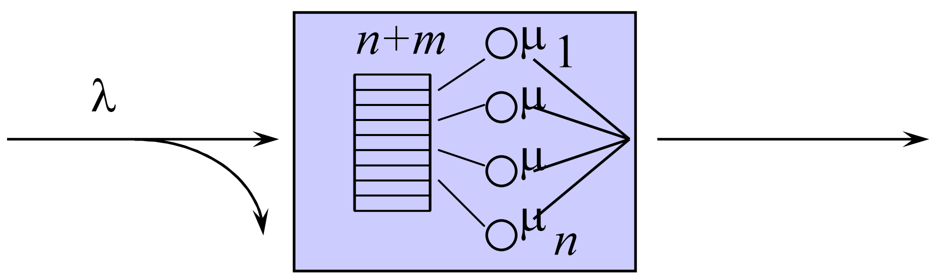
Pure sharing system

- Finite number of servers ($n < \infty$), infinite number of service places ($n + m = \infty$), no waiting places
  - If there are at most n customers in the system ($x \leq n$), each customer has its own server. Otherwise ($x > n$), the total service rate ($n\mu$) is shared fairly among all customers.
  - Thus, the rate at which a customer is served equals $\min\{\mu, n\mu/x\}$
  - No customers are lost, and no one needs to wait before the service.
  - But the delay is the greater, the more there are customers in the system. Thus, delay is an interesing measure from the customer's point of view.



Lossy sharing system

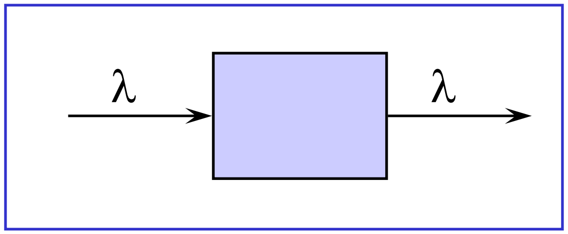
- Finite number of servers ($n < \infty$), finite number of service places ($n + m < \infty$), no waiting places
  - If there are at most n customers in the system ($x \leq n$), each customer has its own server. Otherwise ($x > n$), the total service rate ($n\mu$) is shared fairly among all customers.
  - Thus, the rate at which a customer is served equals $\min\{\mu, n\mu/x\}$
  - Some customers are lost, but no one needs to wait before the service.



Contents

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
  - $\overline{N}$ = average number of customers in the system
  - $\overline{T}$ = average time a customer spends in the system = average delay
- **Little's formula**:



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

Proof (1)

- Let
  - $N(t)$ = the number of customers in the system at time t
  - $A(t)$ = the number of customers arrived in the system by time t
  - $B(t)$ = the number of customers departed from the system by time t
  - T_i = the time customer i spends in the system = its delay
- As $t \to \infty$,

$$\frac{1}{t}\int_0^t N(s)ds \to \overline{N}, \quad \frac{1}{A(t)}\sum_{i=1}^{A(t)} T_i \to \overline{T}, \quad \frac{1}{B(t)}\sum_{i=1}^{B(t)} T_i \to \overline{T} \quad (1)$$

- In addition (due to the stability assumption),

$$\tfrac{1}{t}A(t) \to \lambda, \quad \tfrac{1}{t}B(t) \to \lambda \quad (2)$$

Proof (2)

- We may assume that
  - the system is empty at time $t = 0$,
  - the customers depart from the system in their arrival order (FIFO)
- Then (see the figure in the following slide)

$$\sum_{i=1}^{B(t)} T_i \le \int_0^t N(s)ds \le \sum_{i=1}^{A(t)} T_i$$

- Thus,

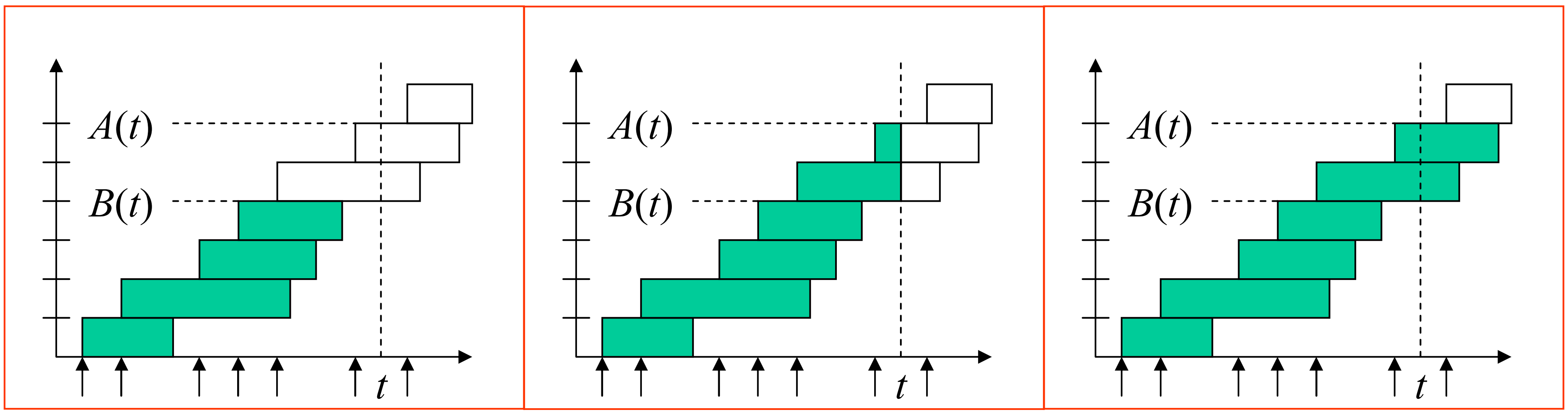
$$\frac{B(t)}{t}\frac{1}{B(t)}\sum_{i=1}^{B(t)} T_i \le \frac{1}{t}\int_0^t N(s)ds \le \frac{A(t)}{t}\frac{1}{A(t)}\sum_{i=1}^{A(t)} T_i$$

- As $t \to \infty$, we have, by (1) and (2),

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

- Q.E.D.

Proof (3)



$\sum_{i=1}^{B(t)} T_i$

$\int_0^t N(s)ds$

$\sum_{i=1}^{A(t)} T_i$

THE END