EP2200
Queueing theory and teletraffic systems

Viktoria Fodor
Laboratory of Communication Networks
School of Electrical Engineering

Lecture 1

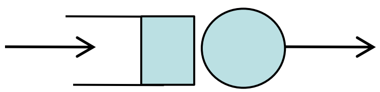
"If you want to model networks
Or a complex data flow
A queue's the key to help you see
All the things you need to know."

(Leonard Kleinrock, Ode to a Queue from IETF RFC 1121)

What is queuing theory?
What are teletraffic systems?

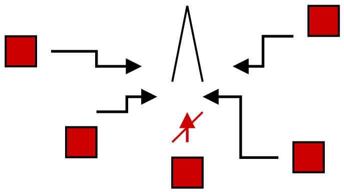
Queuing theory

- Mathematical tool to describe resource sharing systems, e.g., telecommunication networks, computer systems
  - Requests arrive and are served dynamically
  - Request may form a queue to wait for service
- Applied probability theory



Teletraffic systems

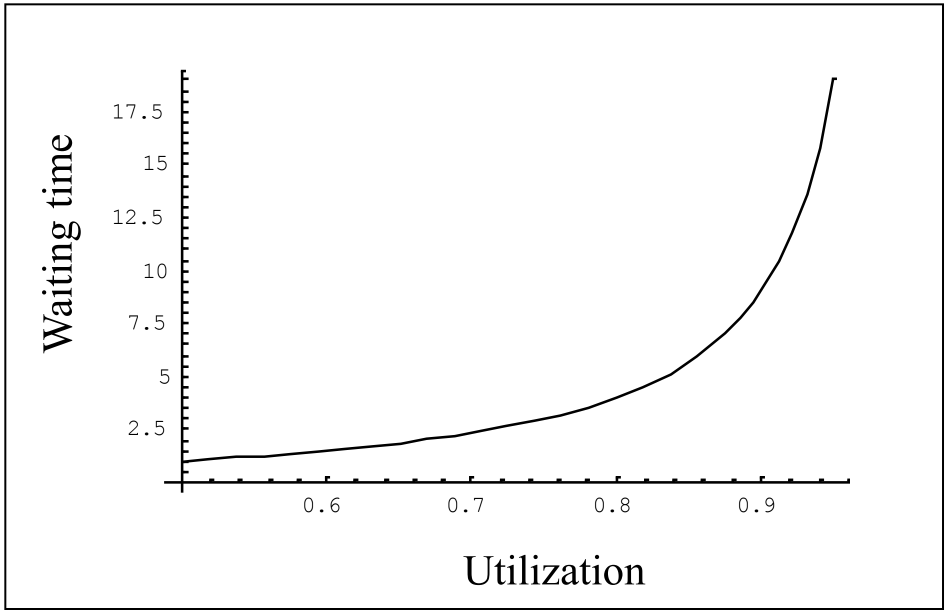
- Systems with telecommunication traffic (data networks, telephone networks)
- Are designed and evaluated using queuing theory
  - Traffic control (congestion and error control)
  - Traffic engineering (routing, virtual networks)
  - Network dimensioning (capacity planning, topology design,)



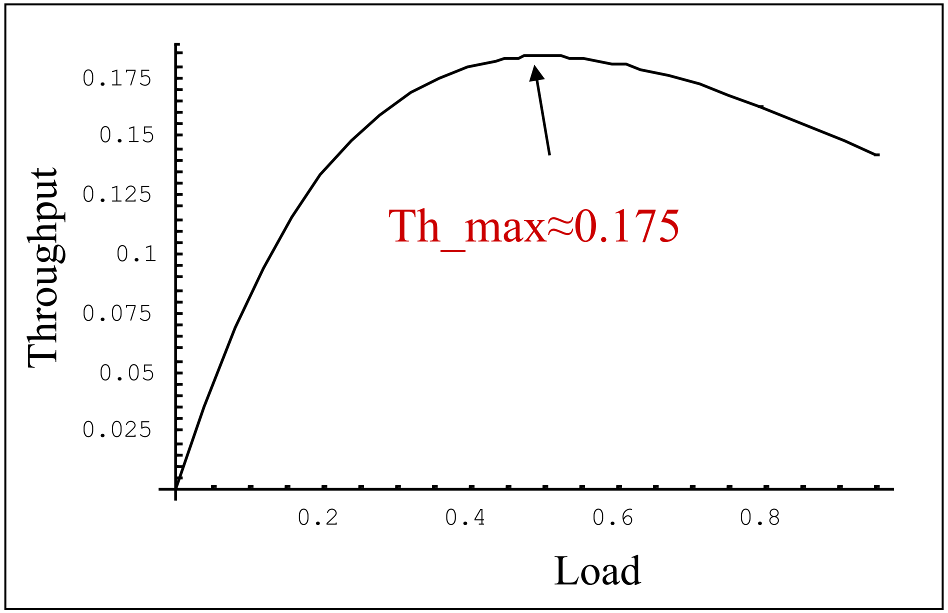
Why do we need a whole theory for that?

Teletraffic systems are non-linear

Waiting time at a router vs. utilization

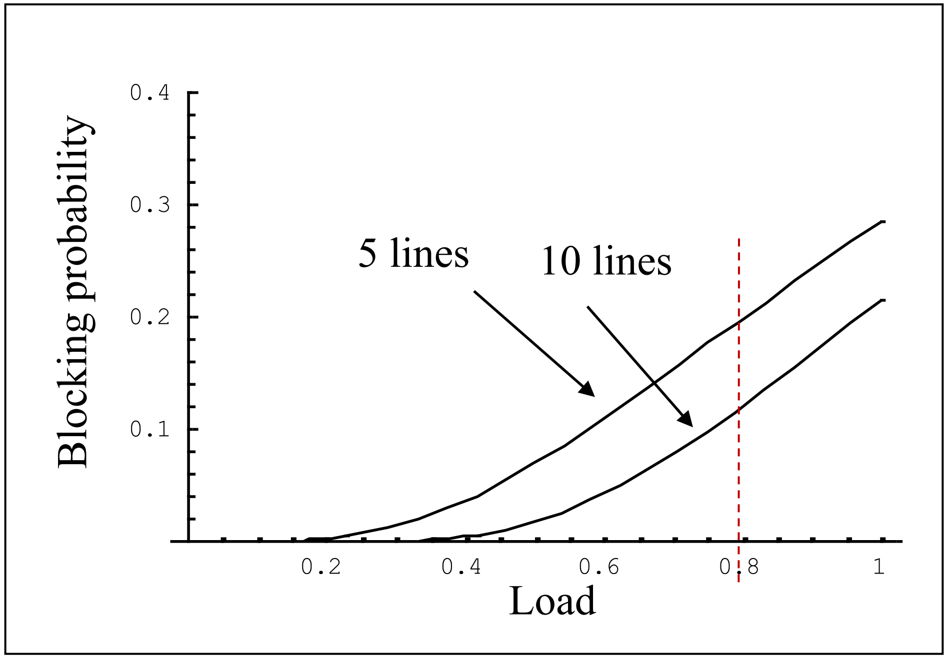


Throughput in a WLAN vs. load

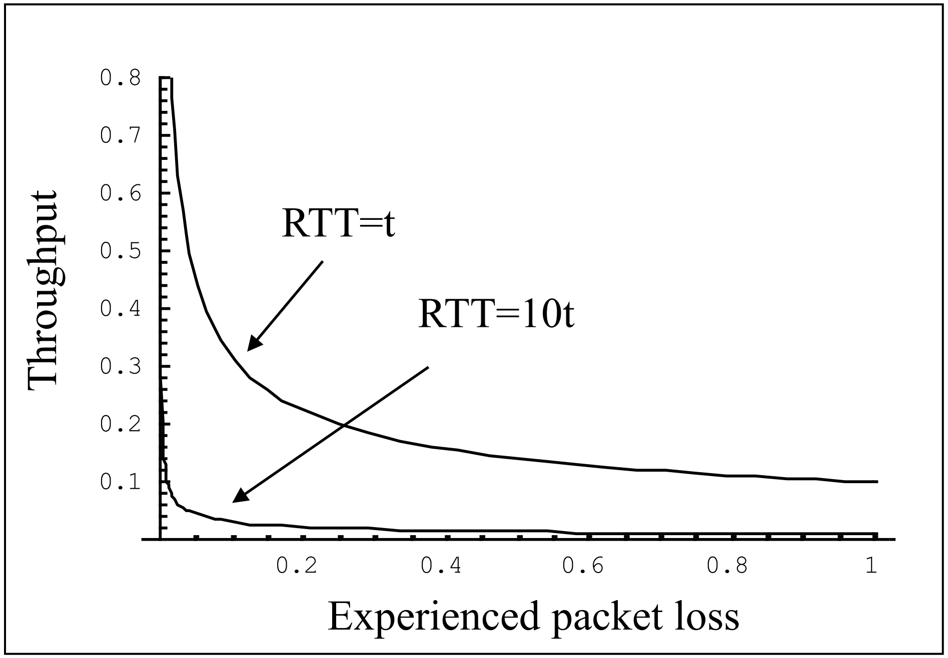


Teletraffic systems are non-linear

Call blocking probability
in a telephone network vs. load



TCP throughput vs. packet loss



Course objectives

- Basic theory
  - understand the theoretical background of queuing systems, apply the theory for systems not considered in class
- Applications
  - find appropriate queuing models of simple problems, derive performance metrics
- Basis for modeling more complex problems
  - advanced courses on performance evaluation
  - master thesis project
  - industry (telecommunication engineer)
- Prerequisites
  - mathematics, statistics, probability theory, stochastic systems
  - communication networks, computer systems

Course organization

- **Course responsible**
  - Viktoria Fodor <vfodor@s3.kth.se>
- **Lectures**
  - Viktoria Fodor
- **Recitations**
  - Ioannis (John) Glarapoulos <ioannisg@s3.kth.se>
  - Liping Wang
- **Course web page**
  - http://www.kth.se/student/kurser/kurs/EP2200?l=en_UK
  - Home assignments, messages, updated schedule and course information
    - Your responsibility to stay up to date!
  - Useful resources: Erlang and Engset calculators, Java Aplets
  - Useful links: on-line books
  - Links to probability theory basics

Course material

- **Course binder**
  - Lecture notes by Jorma Virtamo, HUT, and Philippe Nain, INRIA
    - Used with their permission
  - Excerpts from L. Kleinrock, *Queueing Systems*
  - Problem set with outlines of solutions
  - Old exam problems with outlines of solutions
  - Erlang tables (get more from course web, if needed)
  - Formula sheet

  For sale at STEX, Q2 building. Costs 100 SEK.
- **No text book needed!**
  - If you would like a book, then you can get one on your own
    - Ng Chee Hock, *Queueing Modeling Fundamentals*, Wiley, 1998. (simple)
    - L. Kleinrock, *Queueing Systems, Volume 1: Theory*, Wiley, 1975 (well known, engineers)
    - D. Gross, C. M. Harris, *Fundamentals of Queueing Theory*, Wiley, 1998 (difficult)
  - Beware, the notations might differ

Course organization

- 12 lectures – cover the theoretical part
- 12 recitations – applications of queuing models
- Home assignments and project (1.5 ECTS, compulsory, pass/fail)
- Home assignment
  - problems and (later) solutions on the web
  - individual submission, only handwritten version
  - you need 75% satisfactory solution to pass this moment
  - submission on Nov. 21, submit at the STEX office
- Small project
  - computer exercise
  - details later
  - submission deadline: Jan 13.

Exam

- There is a written exam to pass the course, 5 hours
  - Consists of five problems of 10 points each
    - Passing grade usually 20 ± 3 points
  - Allowed aid is the Beta mathematical handbook (or similar) and simple calculator. Probability theory and queuing theory books are not allowed!
  - The sheet of queuing theory formulas will be provided, also Erlang tables and Laplace transforms, if needed

- Possibility to complementary oral exam if you miss E by 2-3 points (Fx)
  - Complement to E

- Registration is mandatory for all the exams
  - At least two weeks prior to the exam

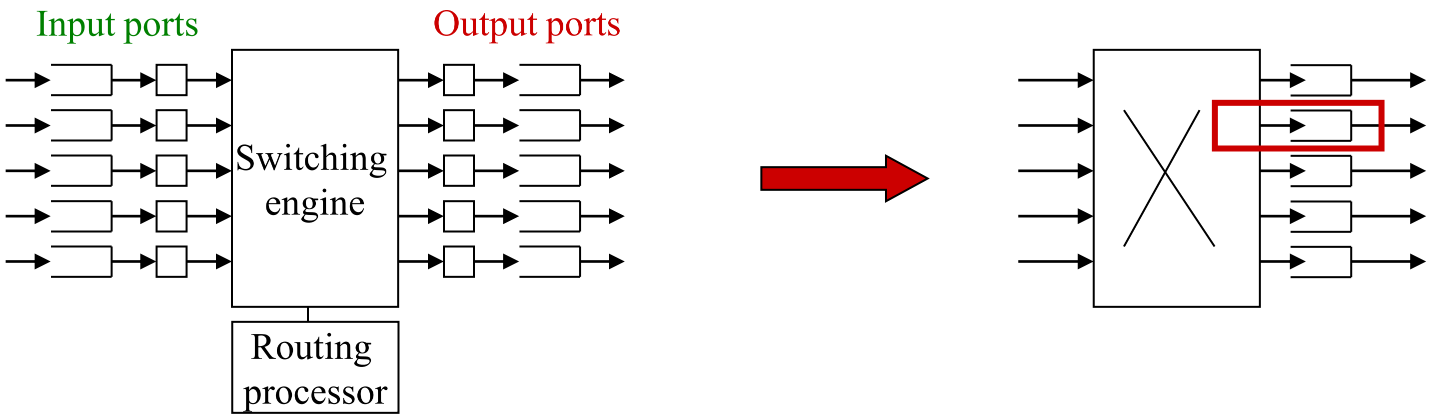
- Students from previous years: contact me or STEX (stex@ee.kth.se) if you are not sure what to do

Lecture 1
Queuing systems - introduction

- Teletraffic examples and the performance triangle
- The queuing model
  - System parameters
  - Performance measures
- Stochastic processes recall

Example

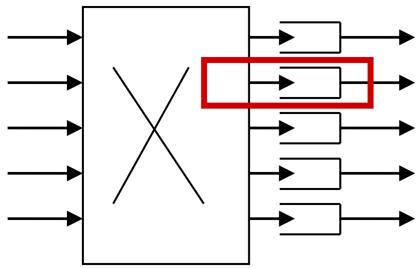
- Packet transmission at a large IP router



- We simplify modeling
  - typically the switching engine is very fast
  - the transmission at the output buffers limits the packet forwarding performance
  - we do not model the switching engine, only the output buffers

Example

- Packet transmission at the output link of a large IP router – packets wait for free output link

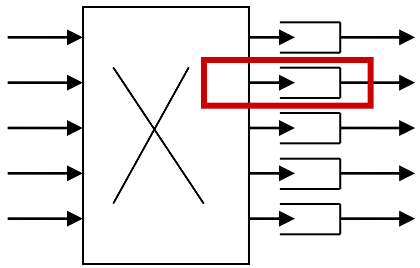


- Performance:
  - Waiting time in the buffer
  - Number of packets waiting

- Depends on:
  - How many packets arrive in a time period (packet/sec)
  - How long is the transmission time (packet out of the buffer)
    - Link capacity (bit/s)
    - Packet size (bits)

Example

- Packet transmission at the output link of a large IP router - packets wait for free output link



- Performance:
  - Number of packets waiting
  - Waiting time in the buffer

- Depends on:
  - How many packets arrive
  - Packet size

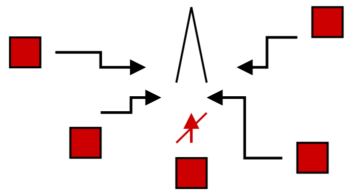
  $\rightarrow$ Service demand

  - Link capacity

  $\rightarrow$ Server capacity

Example

- Voice calls in a GSM cell – call blocked if all channels busy



- Performance
  - Utilization of the channels
  - Probability of blocking a call

- Depends on:
  - How many calls arrive
  - Length of a conversation

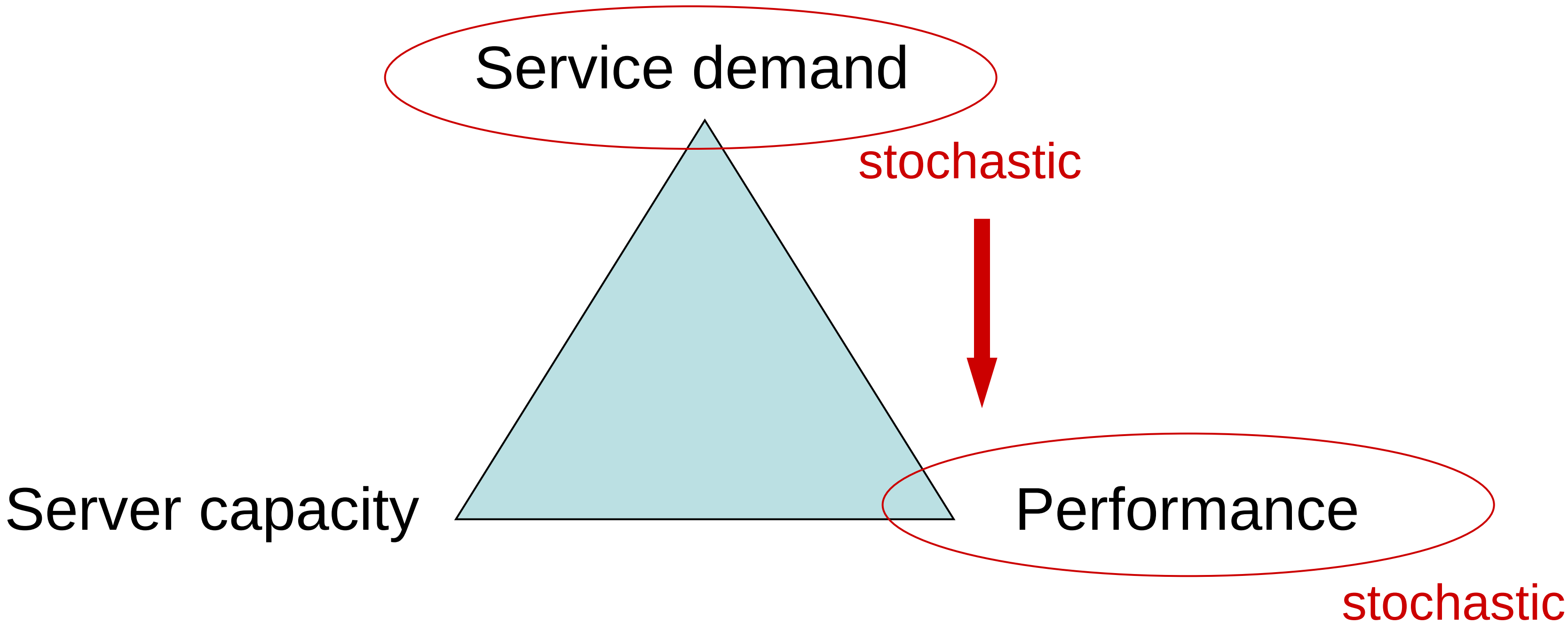
  $\rightarrow$ Service demand

  - Cell capacity (number of voice channels)

  $\rightarrow$ Server capacity

Performance of queuing systems

- The triangular relationship in queuing



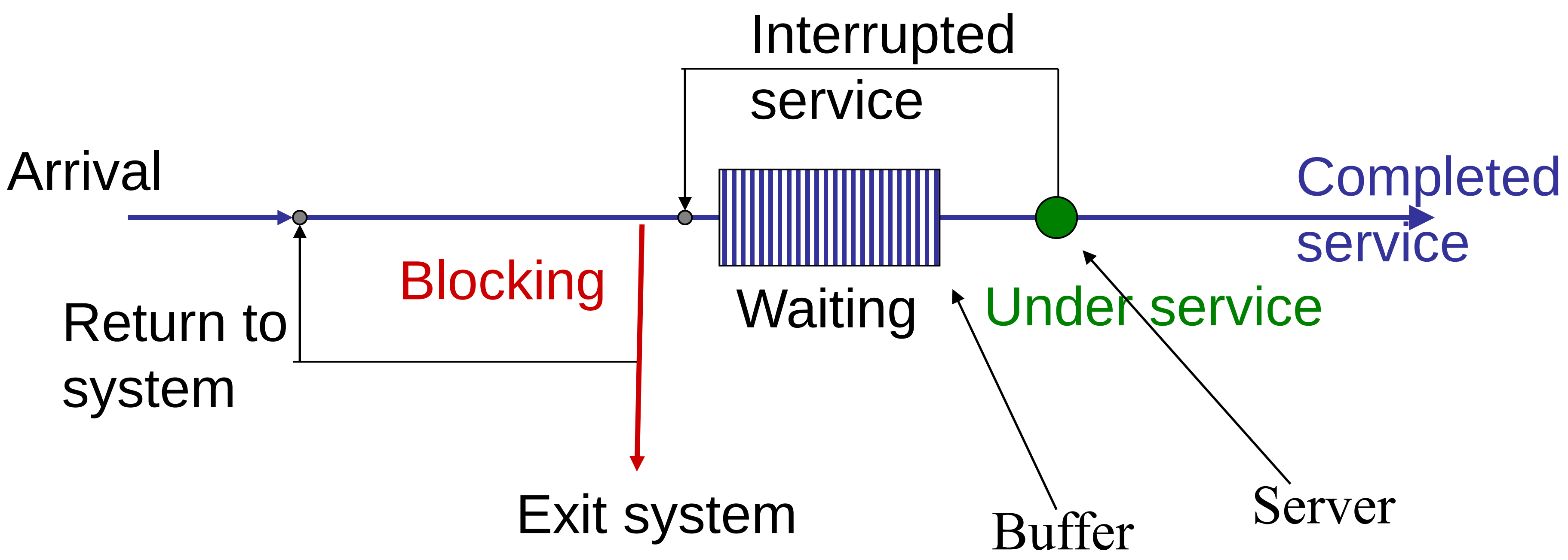
- Works in 3 directions
  - Given service demand and server capacity → achievable performance
  - Given server capacity and required performance → acceptable demand
  - Given demand and required performance → required server capacity

Lecture 1
Queuing systems - introduction

- Teletraffic examples and the performance triangle
- The queuing model
  - System parameters
  - Performance measures
- Stochastic processes recall

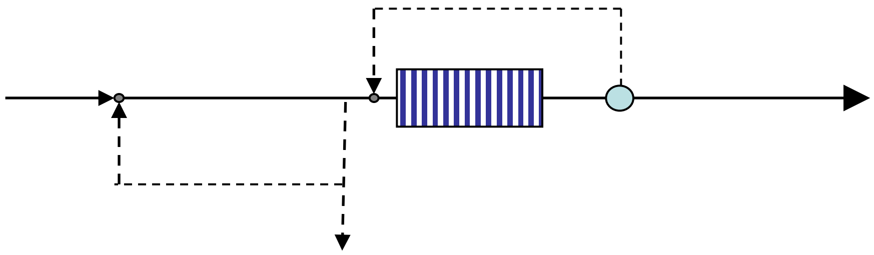
Block diagram of a queuing system

- Queuing system: abstract model of a teletraffic system
  - buffer and server(s)
- Customers arrive, wait, get served and leave the queuing system
  - customers can get blocked, service can be interrupted



Description of queuing systems

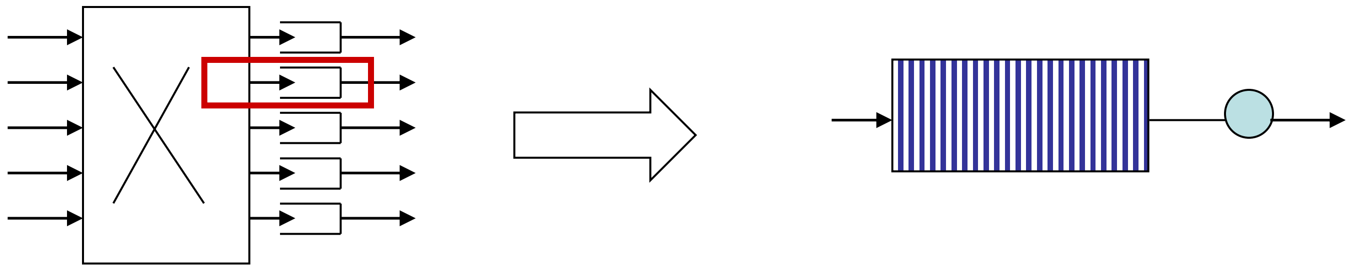
- System parameters
  - Number of servers (customers served in parallel)
  - Buffer capacity
    - Infinite: enough waiting room for all customers
    - Finite: customers might be blocked
  - Order of service (FIFO, random, priority)
- Service demand (stochastic, given by probability distributions)
  - Arrival process: How do the customers arrive to the system
  - Service process: How much service does a customer demand



Customer:
- IP packet
- Phone call

Examples in details

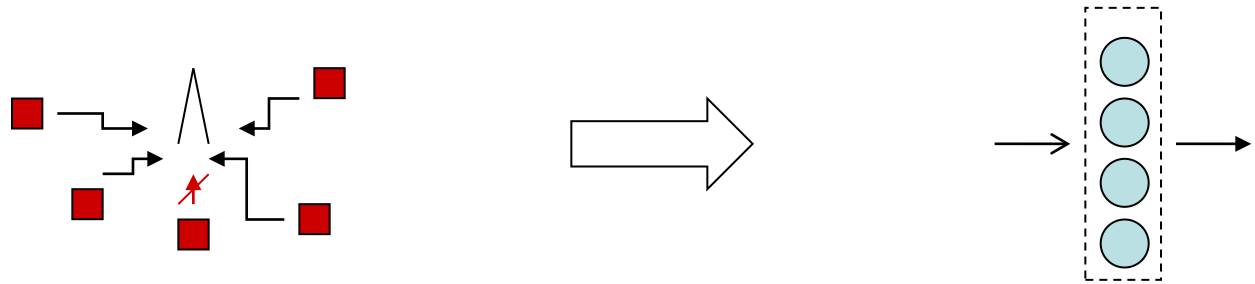
- Packet transmission at the output link of a large IP router



- Number of servers: 1
- Buffer capacity: max. number of IP packets
- Order of service: FIFO
- Arrival process: IP packet multiplexed at the output buffer
- Service process: transmission of one IP packet
  (service time = transmission time = packet length / link transmission rate)

Examples in details

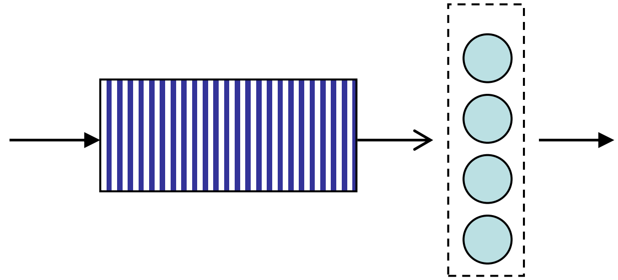
- Voice calls in a GSM cell
  - channels for parallel calls, each call occupies a channel
  - if all channels are busy the call is blocked



- Number of servers: number of parallel channels
- Buffer capacity: no buffer
- Order of service: does not apply
- Arrival process: calls attempts in the GSM cell
- Service process: the phone call (service time = length of the phone call)

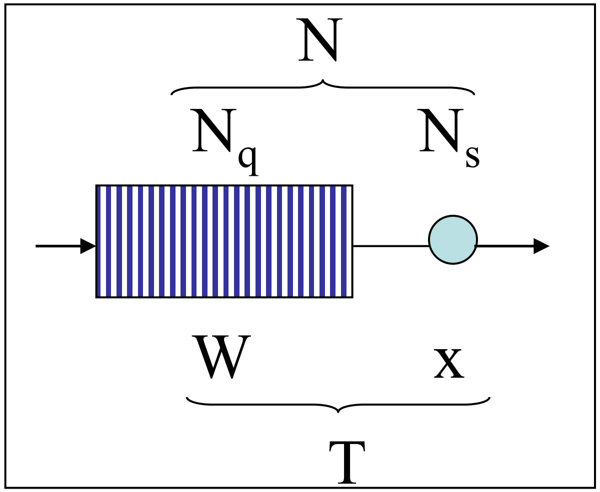
Group work

- Service at a bank, with “queue numbers” and several clerks
- Draw the block diagram of the queuing systems



- Arrival process: customers arriving to the bank
- Service process: the service of a customer at one of the clerks
- Number of servers: number of clerks working in the bank
- Buffer capacity: infinite/finite
- Order of service: FIFO (queue numbers)

Performance measures



- Number of customers in the system
  - Number of customers in the queue
  - Number of customers in the server
- System time
  - Waiting time of a customer
  - Service time of a customer
- Probability of blocking (blocked customers / all arrivals)
- Utilization of the server (time server occupied / all considered time)

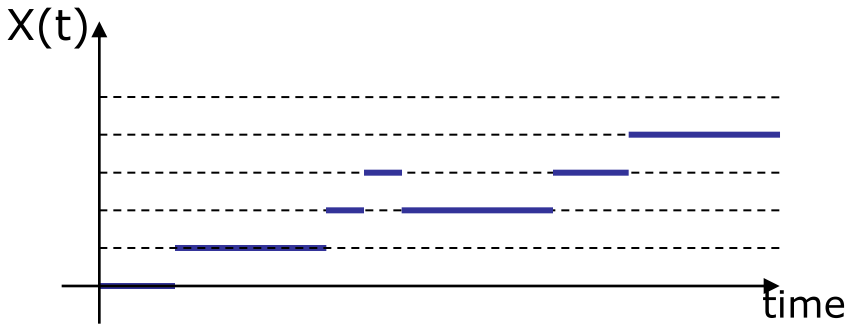
- Transient measures
  - how will the system state change in the near future?
- Stationary measures
  - how does the system behave on the long run?
  - average measures
  - often considered in this course

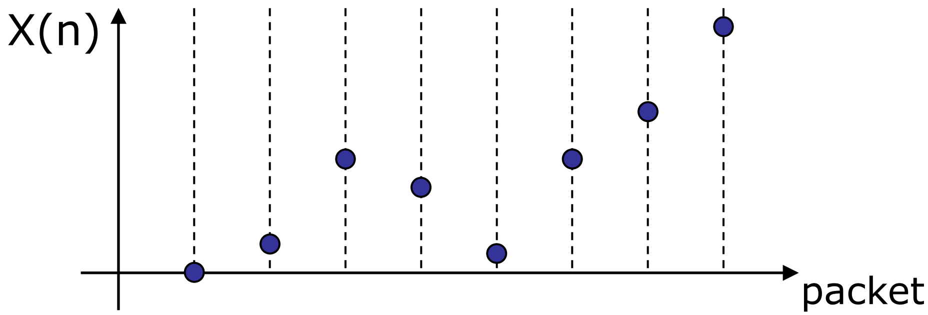
Lecture 1
Queuing systems - introduction

- Teletraffic examples and the performance triangle
- The queuing model
  - System parameters
  - Performance measures
- Stochastic processes recall

Stochastic process

- Stochastic process
  - A system that evolves – changes its state - in time in a random way
  - Family of random variables
  - Variables indexed by a time parameter
    - Continuous time: $X(t)$, a random variable for each value of t
    - Discrete time: $X(n)$, a random variable for each step $n=0,1,\ldots$
  - State space: the set of possible values of r.v. $X(t)$ (or $X(n)$)
    - Continuous or discrete state





- Number of packets waiting:
  - Discrete space
  - Continuous time
- Waiting time of consecutive packets:
  - Discrete time
  - Continuous space

Stochastic process - statistics

- We are interested in quantities, like:
  - time dependent (transient) state probabilities (statistics over many realizations, an *ensemble* of realizations, *ensemble average*):

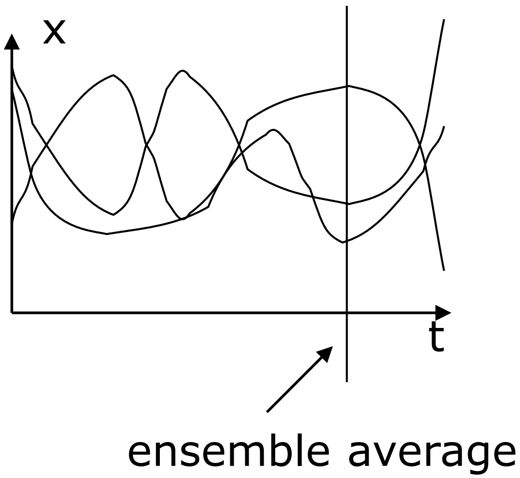
$$f_x(t) = P(X(t) = x),\ \ F_x(t) = P(X(t) \le x)$$

  - n^{th} order statistics – joint distribution over n samples

$$F_{x_1,\ldots,x_n}(t_1,\ldots,t_n) = P(X(t_1) \le x_1,\ldots,X(t_n) \le x_n)$$

  - limiting (or stationary) state probabilities (if exist) :

$$f_x = \lim_{t\to\infty} P(X(t) = x), \quad F_x = \lim_{t\to\infty} P\{X(t) \le x\}$$



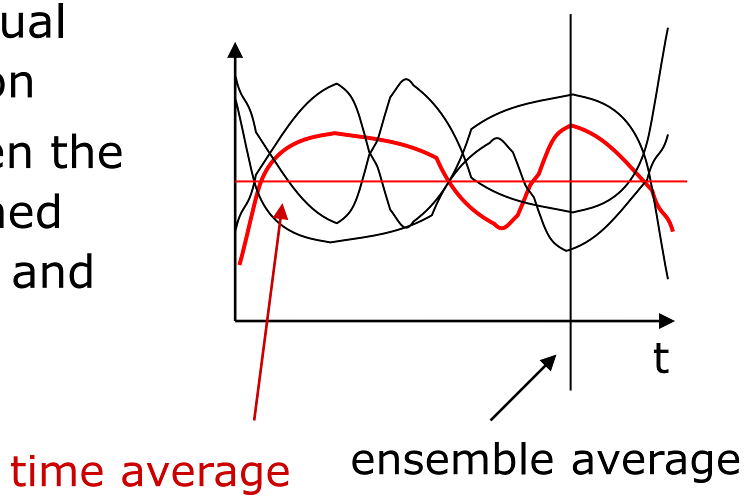
Stochastic process - terminology

- The stochastic process is:
  - stationary, if all n^th order statistics are unchanged by a shift in time:

$$F_x(t+\tau) = F_x(t), \quad \forall t$$

$$F_{x_1,\dots,x_n}(t_1+\tau,\dots,t_n+\tau) = F_{x_1,\dots,x_n}(t_1,\dots,t_n), \quad \forall n, \quad \forall t_1,\dots,t_n$$

  - ergodic, if the ensemble average is equal to the time average of a single realization
  - consequence: if a process ergodic, then the statistics of the process can be determined from a single (infinitely long) realization and vice versa



Stochastic process

- Example on stationary versus ergodic
- Consider a source, that generates the following sequences with the same probability:
  - ABABABAB…
  - BABABABA…
  - EEEEEEEE…
- Is this source stationary?
  - yes: ensemble average is time independent $p(A)=p(B)=p(E)=1/3$
- Is this source ergodic?
  - no: the ensemble average is not the same as the time average of a single realization

Summary

Today:

- Queuing systems - definition and parameters
- Stochastic processes

Next lecture:

- Poisson processes and Markov-chains, the theoretical background to analyze queuing systems

Recitation:

- Probability theory and transforms
- Prepare for the recitation: read Virtamo 1-3 in the course binder or download from the course web
  - Definition of probability of events
  - Conditional probability, law of total probability, Bayes formula, independent events
  - Random variables, distribution functions (discrete and continuous)
  - Z and Laplace transforms