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 dynamically
  - Request may form a queue to wait for service
- Applied probability theory, stochastic processes

Teletraffic systems

- Systems with telecommunication traffic (data networks, telephone networks)
- Are designed and evaluated using queuing theory

Why do we need a whole theory for that?

What is queuing theory?
What are teletraffic systems?

Streaming:

How long should the system pre-fetch to ensure continuous streaming?

Cloud computing:

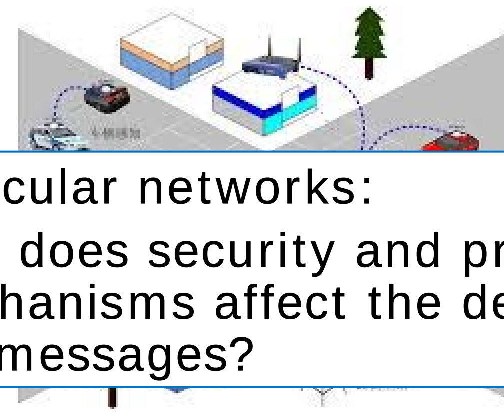
What limits the performance? The server? The network? My device?

Software Defined Networking:

Will increased time of routing decisions degrade network performance?

Smart grid:

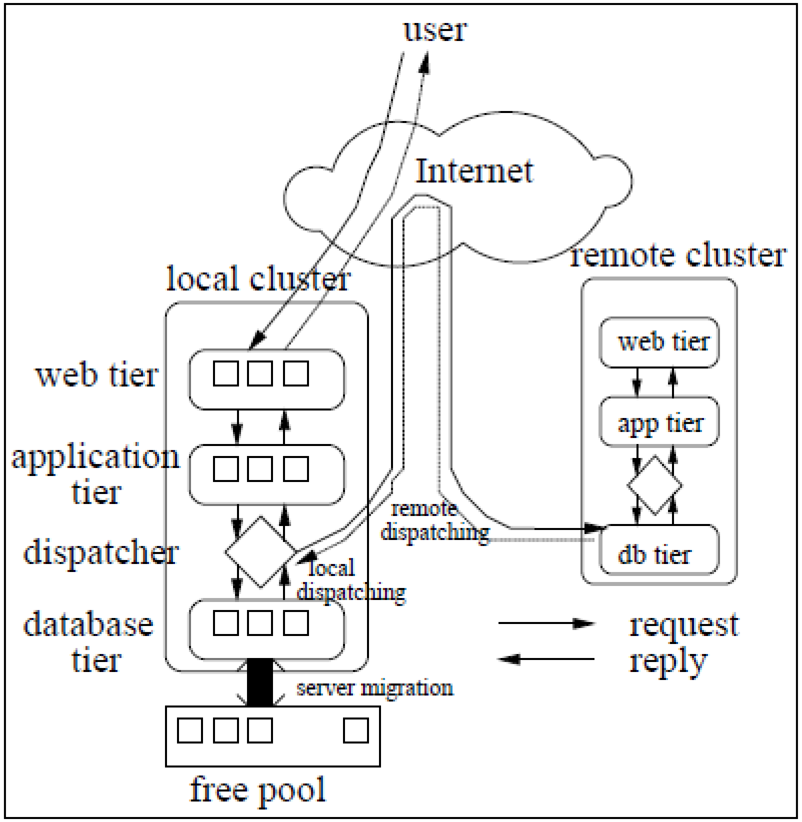
Can the network delay ensure efficient control? Would secure routing paths compromise performance?



Vehicular networks:

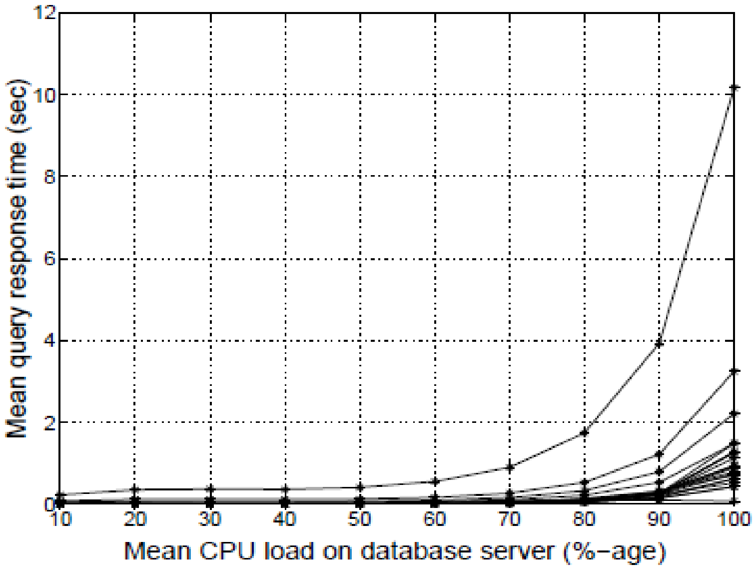
How does security and privacy mechanisms affect the delay of the messages?

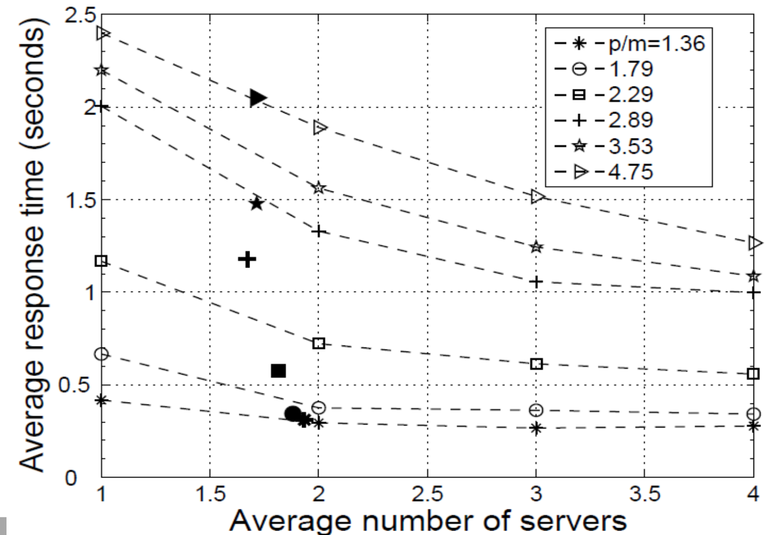
WEB server response time



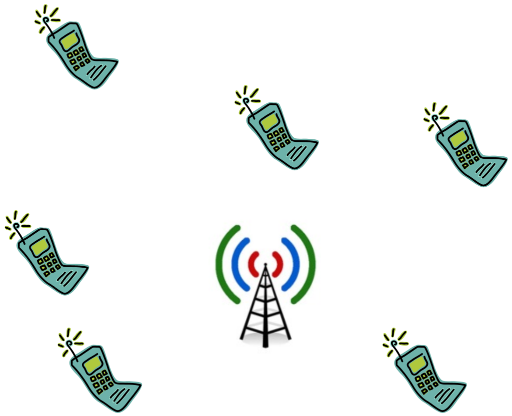
High Performance Resource Allocation and Request Redirection Algorithms for Web Clusters

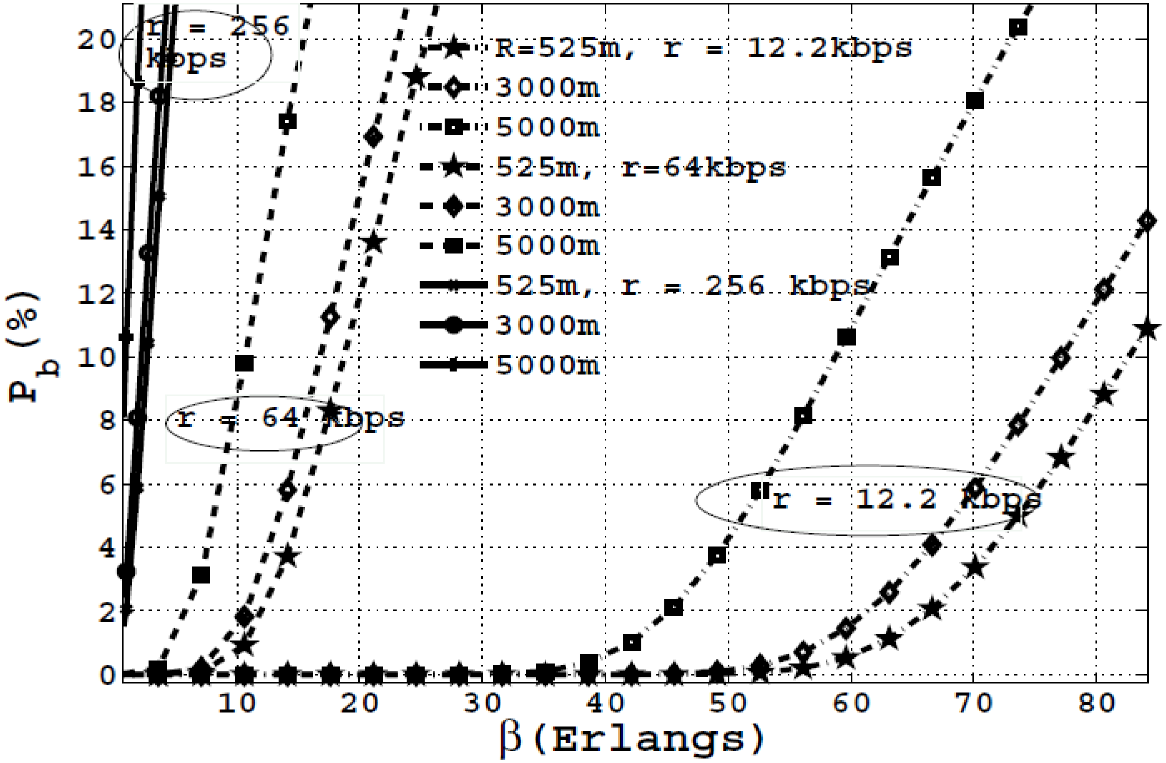
Supranamaya Ranjan, *Member, IEEE,* and Edward Knightly, *Senior Member, IEEE*





Call blocking in mobile networks





Distance dependent Call Blocking Probability, and Area Erlang Efficiency of Cellular Networks

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, lectures and some recitations
  - Viktoria Fodor <vfodor@kth.se>
- Recitations, project work
  - Sladjana Josilo <josilo@kth.se>

- Course web page
  - KTH Social EP2200 (https://www.kth.se/social/course/EP2200/)
  - Course reading material, home assignments, project, messages, updated schedule and course information
    - Your responsibility to stay up to date!
  - Useful resources: applets, calculators
  - Useful links: on-line books
  - Links to probability theory basics

Course material

- All course material on line
  - 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 full solutions
  - Erlang tables
  - Formula sheet, Laplace transforms

- Printed course material: STEX

- 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 theory
- Two home assignments and a project (1.5 ECTS, compulsory, pass/fail)
- Deadlines, information on the web
- Home assignments (beginning and middle of the course)
  - numerical exercises and proofs
  - individual submission, only handwritten version
  - you need 75% satisfactory solution to pass this moment
  - submit in class or at the STEX office
- Small project (end of the course)
  - computer exercise (matlab, C, java, simulation platform…)
  - +5 points for outstanding projects (upper 10%)

Exam

- There is a written exam to pass the course, 5 hours
  - 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 (same as in the course binder and on the web)
- 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 STEX (stex@ee.kth.se) if you are not sure what to do
- Questions: e-mail or via KTH Social, including asking for meeting

PhD students in brief

- Lectures and recitations as for all students
- Submission of home assignments as all students
- Exam as all students (P/F, at least grade B level to pass)

- Additional weekly seminars on advanced material
- Different project

- See specific menu point on the EP2200 KTH Social course web

Step 0

- Queuing theory is applied probability theory
- You need to be able to work with basic probability theory tools

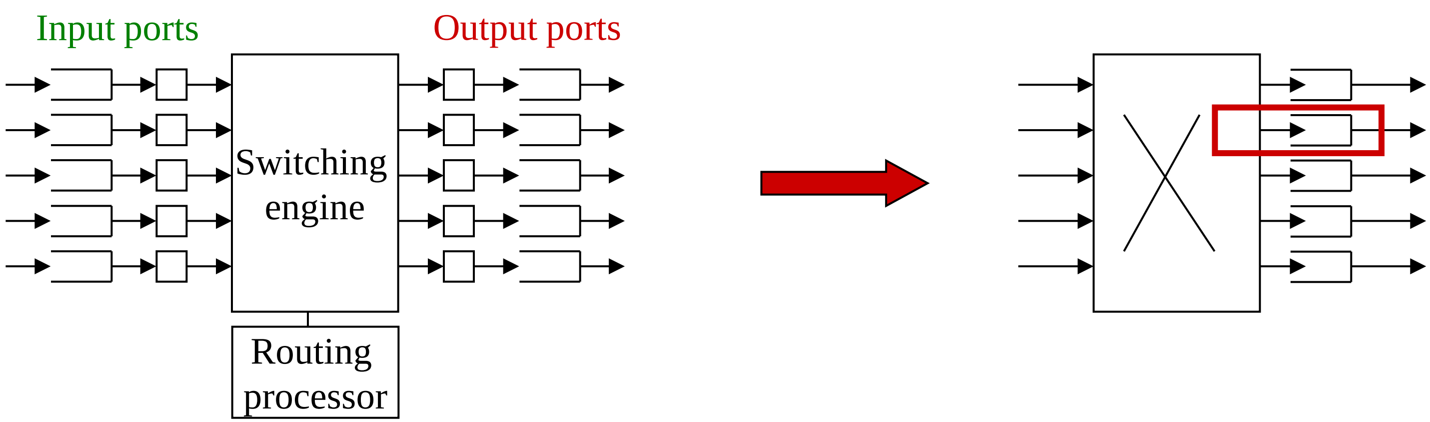
- Short summary in Virtamo notes, chapters 1-4
- Suggested video lectures
- First recitation is dedicated to probability theory overview
- Early home assignment with only probability theory problems

Lecture 1
Queuing systems - introduction

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

Example

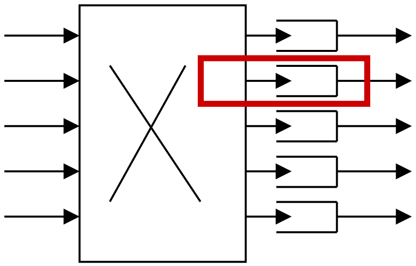
- Packet transmission at a large IP router



- 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 arrive randomly and wait for free output link

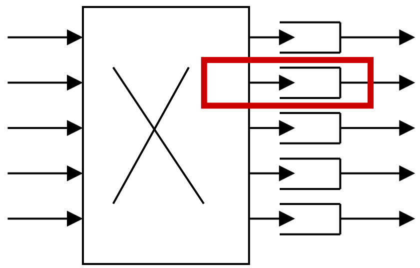


- Performance:
  - Waiting time in the buffer
  - Number of packets waiting
  - Probability of buffer overflow and packet loss

- 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
  - Probability of buffer owerflow

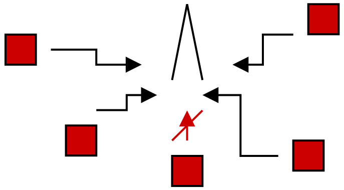
- Depends on:
  - How many packets arrive
  - Packet size

  } → Service demand

  - Link capacity → Server capacity

Example

- Voice calls in a GSM cell – calls arrive randomly and occupy a "channel". Call blocked if all channels busy.



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

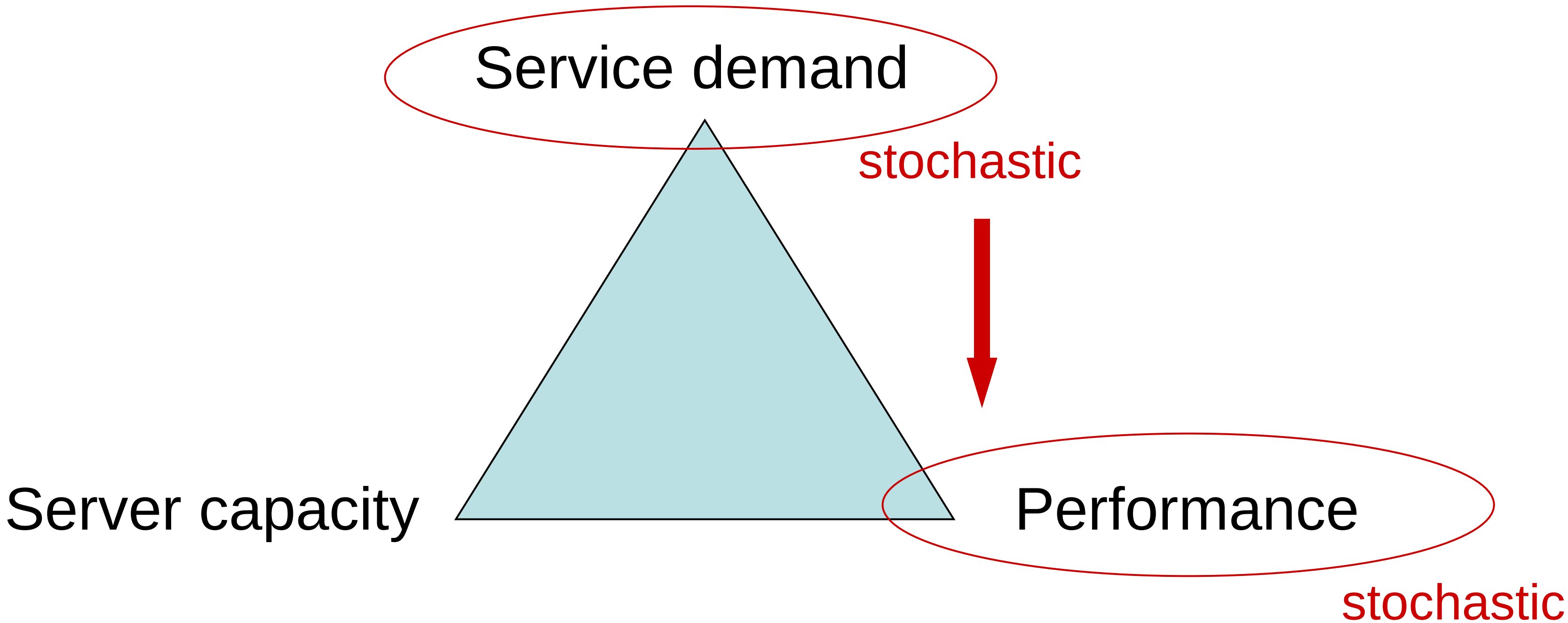
- Depends on:
  - How many calls arrive
  - Length of a conversation

  } → Service demand

  - Cell capacity (number of voice channels) → Server capacity

Performance of queuing systems

- The triangular relationship in queuing



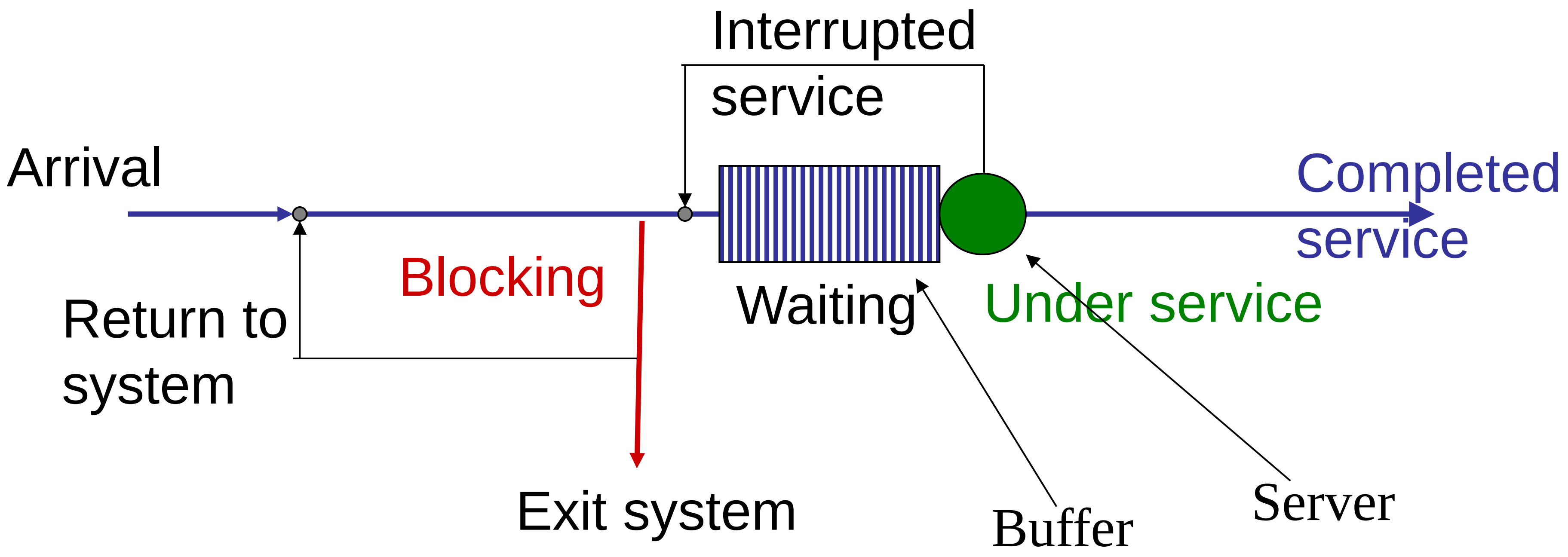
- Works in 3 directions
  - Given service demand and server capacity → achievable performance
  - Given server capacity and required performance → acceptable demand
  - Given demand and required performance → required server capacity

Lecture 1
Queuing systems - introduction

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

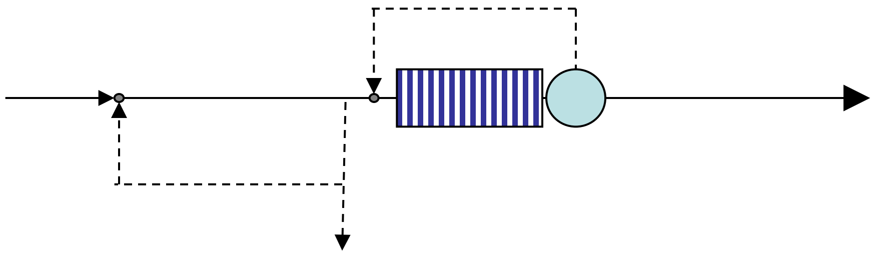
Block diagram of a queuing system

- Queuing system: abstract model of a resource sharing 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

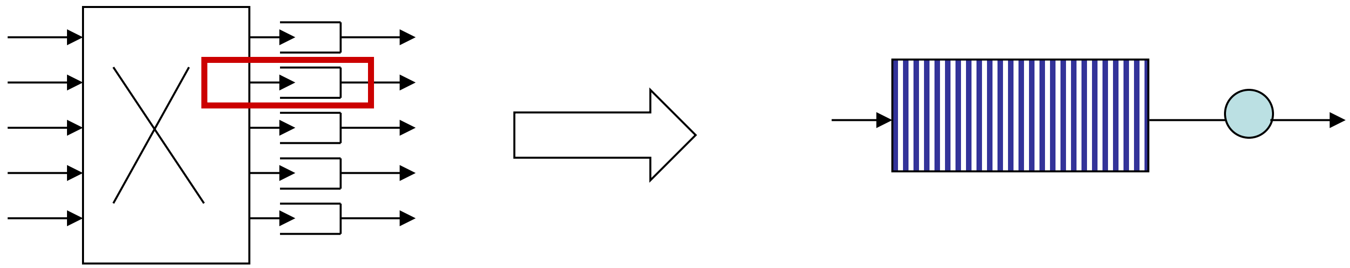
- Server capacity – system parameters
  - Number of servers (customers served in parallel)
  - Buffer capacity
    - Infinite: enough waiting room for all customers
    - Finite: customers might be blocked
  - Order of service (FIFO, random, priority)
- Service demand (stochastic)
  - Arrival process: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



Customer:
- IP packet
- Phone call

Examples in details

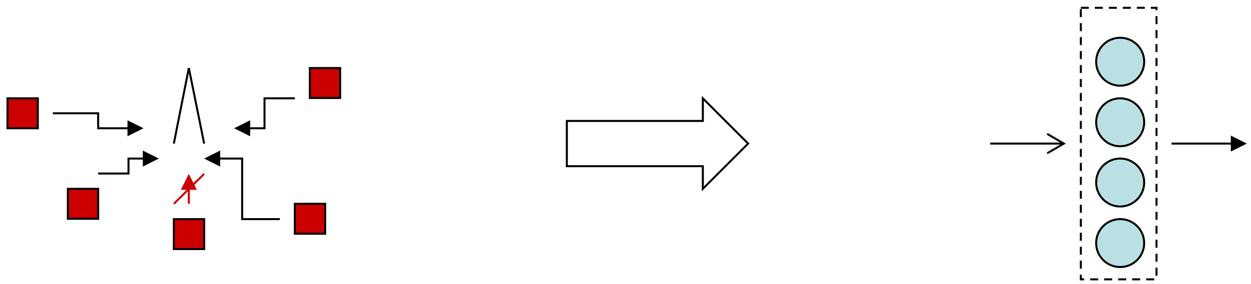
- Packet transmission at the output link of a large IP router



- Number of servers: 1
- Buffer capacity: max. number of IP packets
- Order of service: FIFO
- Arrivals: IP packet multiplexed at the output buffer, packet arrival process or inter-arrival time distribution
- Services: transmission of one IP packet, transmission time distribution (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



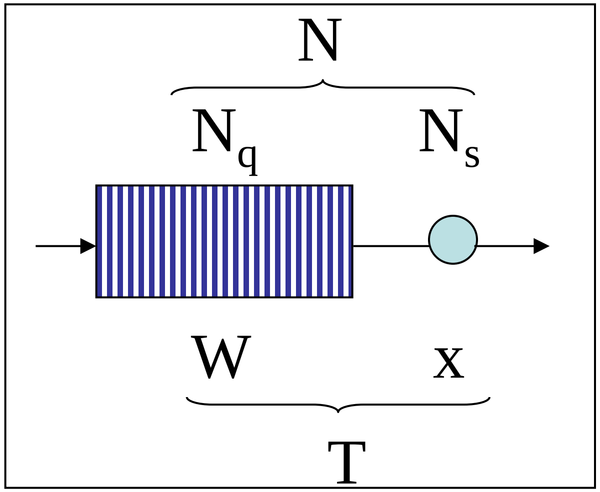
- Number of servers: number of parallel channels
- Buffer capacity: no buffer
- Order of service: does not apply
- Arrivals: call attempts in the GSM cell, call arrival process or call inter-arrival time
- Service: the phone call, call holding time distribution (service time = call holding time)

Group work

1. Service at a bank, with “queue numbers” and several clerks
2. Cashiers at the supermarket, customers select a queue randomly and wait there in a queue
3. Several terminals transmitting in a WLAN

- Draw the block diagram of the queuing systems
- Describe the model: number of servers, buffer capacity, order of service, arrivals, service

Performance measures



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

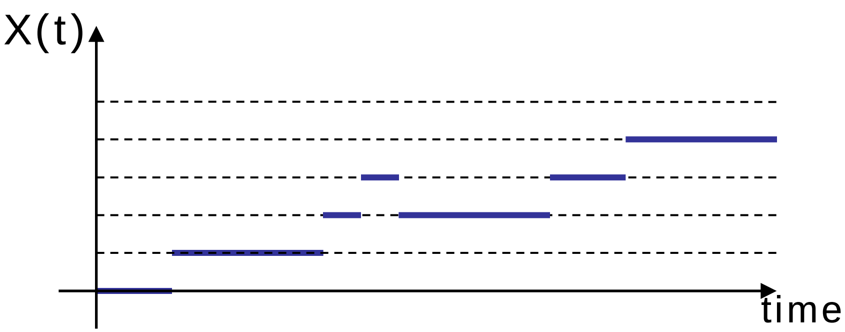
- Transient measures
  - how will the system state change in the near future?
- Stationary measures
  - how does the system behave on the long run?
  - average measures
  - often considered in this course

Lecture 1
Queuing systems - introduction

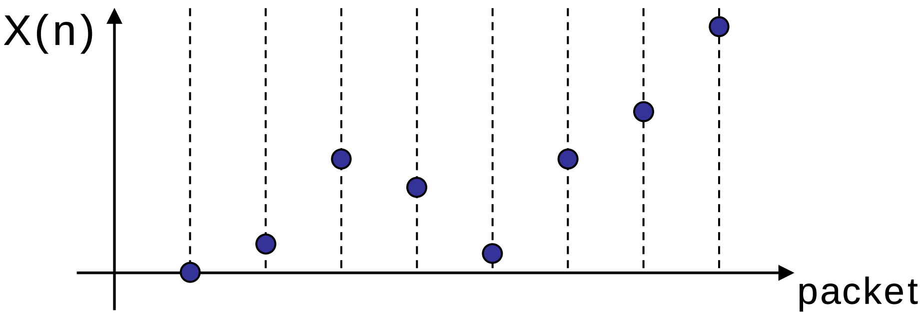
- Teletraffic examples and the performance triangle
- The queuing model
  - Block diagram
  - System parameters
  - Performance measures
- Stochastic processes recall

Stochastic process

- Stochastic process
  - –A system that evolves – changes its state - in time in a random way
  - –Family of random variables (r.v.)
  - –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,…*
  - –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*):

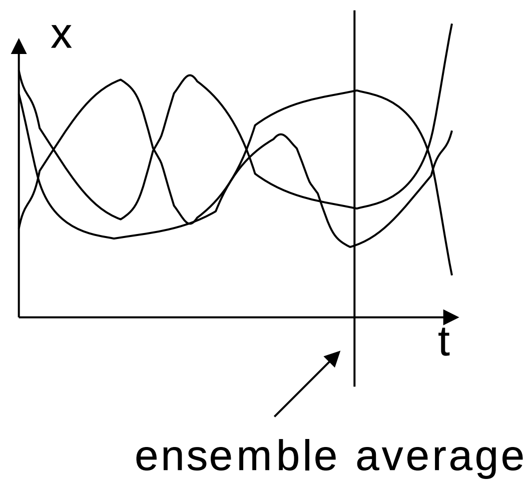
$$f_x(t) = P(X(t) = x),\ \ F_x(t) = P(X(t) \leq x)$$

  – n[th] order statistics – joint distribution over n samples

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

  – limiting (or steady state) probabilities (if exist) :

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



Stochastic process - terminology

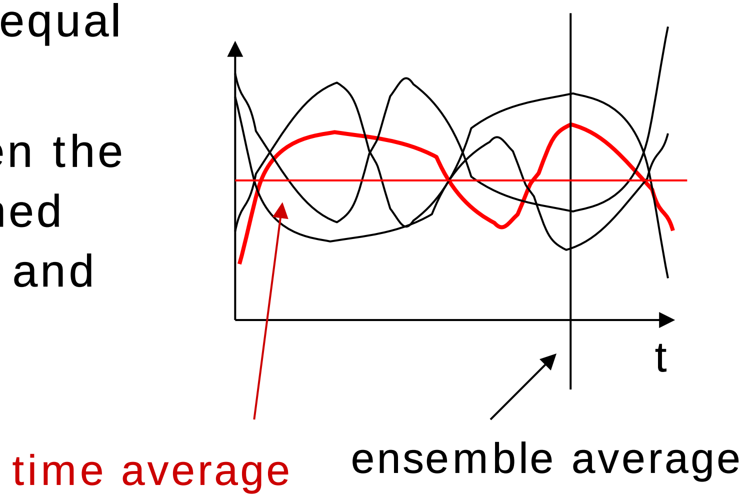
- The stochastic process is:
  – stationary, if all n[th] order statistics are unchanged by a shift in time:

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

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

  – ergodic, if the ensemble statistics are equal to the statistics 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 (state space A,B,E):
  - ABABABAB...
  - BABABABA...
  - EEEEEEEE...
- Is this source stationary?
- Is this source ergodic?

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 (Have a quick look at Virtamo 1-4 before class)
  - Definition of probability of events
  - Conditional probability, law of total probability, Bayes formula, independent events
  - Random variables, distribution functions
    - Bernoulli, Binomial, Geometric, **Poisson**
    - Uniform, **Exponential, Erlang-k**