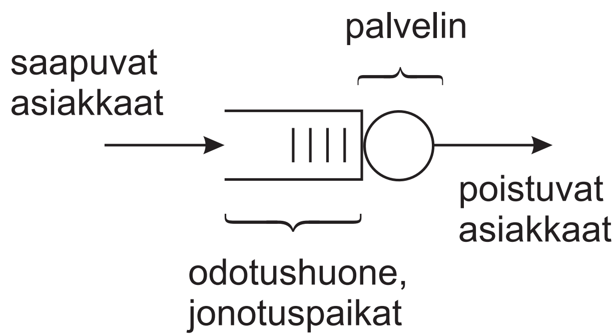
QUEUEING SYSTEMS

General

- Queueing systems constitute a central tool in modelling and performance analysis of e.g. telecommunication systems and computer systems.
- Describes contention on the resources
  - in queueing systems the resources are called servers
  - in applications, the resources may be trunks, capacity ...
- The "customers" arriving at a queue may be calls, messages, packets, tasks ...
- Often the systems are complex (for instance communication network, operating system) and contains many queues, which form a network of queues, i.e. a queueing network.
  - in the beginning we focus on systems consisting of a single queue
  - there are many types of queues, giving rise to a rich theory

Example.

Single server queue



Differentiating factors in queueing systems

- Arrival process
  - interarrival times
  - group arrivals
- Service process
  - service times (requested service work)
- Number of servers
- Number of queues
- Number of waiting places
  - division of the waiting room between the queues
- Service discipline
  - FIFO, LIFO
  - shortest jobs first
  - most profitable jobs first
- Scheduling
  - round robin
  - processor sharing
  - priorities
- Information available
  - upon choice of a queue, does one know the lengths of queues, the service times of individual customers ...
- Discrete time (slotted) / continuous time queues
- Other factors (in real life)
  - screening of the customers
  - bribing
  - ...

The notation of queueing systems (Kendall)

For a unique definition of queueing systems, the following notation is usually used: $A/S/m/c/p$, where

$$\underbrace{A}_{\text{arrival process}} \;/\; \underbrace{S}_{\text{service process}} \;/\; \underbrace{m}_{\text{number of servers}} \;/\; \underbrace{c}_{\text{number of system places}} \;/\; \underbrace{p}_{\text{size of customer population}}$$

- A and S are substituted by one of the commonly used symbols as the case may be.
- Usually the term queue length refers to the total number of customers in the system (including both waiting customers and those in service).
- The parameter c includes both waiting places and service places
  - may be omitted from the notation, whence by default its value is infinite
- The size of the customer population s also on optional parameter
  - may be omitted from the notation, whence by default its value is infinite

A (arrival process)

- Defines the type of arrival process
- Often it is thought that the interarrival times are independent (renewal process), whence the process is determined by the type of interarrival distribution.

Commonly used symbols are

M	exponential interarrival distribution (M = Markovian, memoryless); Poisson process
D	deterministic, constant interarrival times
G	general (unspecified)
E_k	Erlang-k distribution
PH	phase distribution
Cox	Cox distribution

- More abbreviations are introduced as needed.

S (service process)

- Defines the distribution of the customer's service time
- The service time is affected by two factors
  - the required work requested by the customer (e.g. the size of a data packet to be sent, kB)
  - the service rate of the server (e.g. kB/s)
  - the service time is the ratio of these
- In Kendall's notation, the type of the service time distribution is indicated by substituting an appropriate symbol for S; commonly the same symbols (M, D, G, etc.) are being used as for defining the type of the interarrival time distribution

Example 1. The queue $M/M/1$

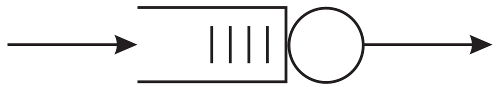
- Poisson arrival process
- exponential service time distribution
- single server
- unlimited number of waiting places

Example 2. The queue $M/M/m/m$

- Poisson arrival process
- exponential service time distribution
- m servers and m system places $\Rightarrow$ no waiting room, so called loss system

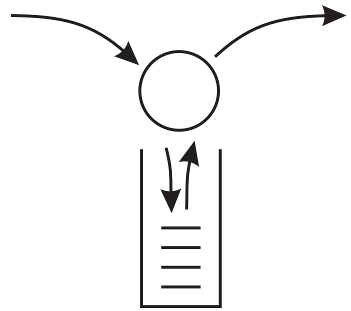
Queueing discipline / scheduling

- Ordinary queue, service in the order of arrivals



$\left\{ \begin{array}{ll} \underline{\text{FCFS}} & \underline{\text{First Come First Served}} \\ \underline{\text{FIFO}} & \underline{\text{First In First Out}} \end{array} \right.$

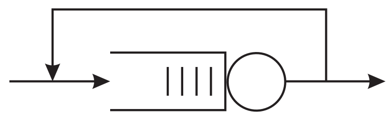
- Stack, the latest arrival is being served first



$\left\{ \begin{array}{ll} \underline{\text{LIFS}} & \underline{\text{Last Come First Served}} \\ \underline{\text{LIFO}} & \underline{\text{Last In First Out}} \end{array} \right.$

- There are three sub-cases of a stack
  - pre-emptive resume
    the arriving customer pre-empts the ongoing service, which is then resumed when the interrupted customer is again taken into the server, continuing from the same point on as at the time of interruption
  - pre-emptive restart
    the arriving customer pre-empts the ongoing service; the service is started from the beginning when the interrupted customer is again taken into the server
  - non-pre-emptive
    the arriving customer waits until the ongoing service is finished before being taken into the server

Queueing discipline / scheduling (continued)

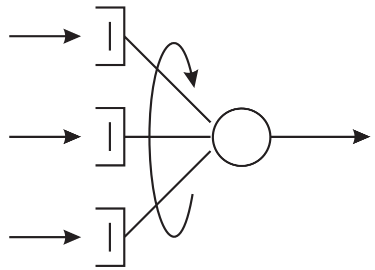


- Service in rotating order

  RR Round robin

  – each customer receives, in turn, a small "time slice" of service

  – polling



- Sharing the capacity of the server

  PS Processor sharing

  – all customers in the queue are receive service simultaneously

  – the capacity is shared evenly between the customers (the service rate received by each customer is inversely proportional to the number of customers in the queue)

  – an idealized form of RR (the time slices tend to zero)

Other service disciplines are e.g.

- SIRO (Service In Random Order)
- SSF (Shortest Jobs First): the service time has to be known in advance; this minimizes the mean waiting time

Queueing discipline / scheduling (continued)

- A queueing discipline is called word conserving, if the capacity of the server / servers is not wasted, i.e. no server is idle if there is at least waiting customer in the system.
- Not all disciplines are work conserving, e.g.
  - LCFS / pre-emptive restart
  - systems, where the server can take a "vacation"