Queueing Systems

Dr Ahmad Khonsari

ECE Dept.

The University of Tehran

Excerpt from virtamo slides

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 . . .

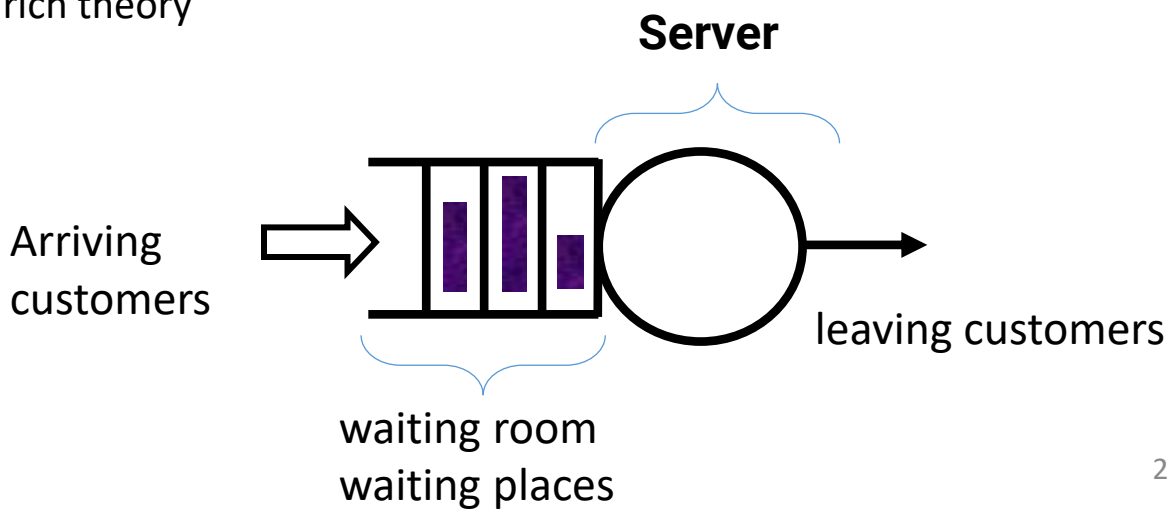
• Often the systems are complex (for instance communication network, operating system) and contains many queues, which form a network of queues, i.e. a queueing network.

– 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

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

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

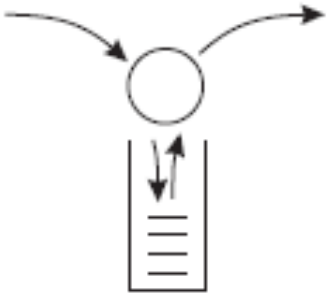
- Poisson arrival process
- exponential service time distribution
- m servers and m system places ⇒ no waiting room, so called loss system

Queueing discipline / scheduling

- Ordinary queue, service in the order of arrivals

FCFS First Come First Served

FIFO First In First Out



- Stack, the latest arrival is being served first

LIFS Last Come First Served

LIFO Last In First Out

- 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

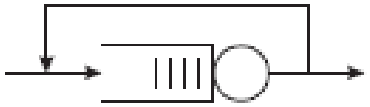
  - pre-emptive restart

    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)

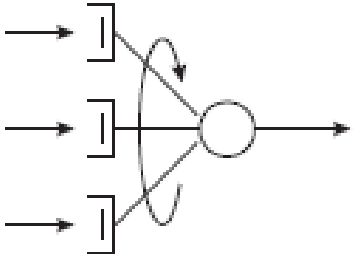


- Service in rotating order

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)
- SJF (Shortest Jobs First): the service time has to be known in advance; this minimizes the mean waiting time

Queueing discipline / scheduling (continued)

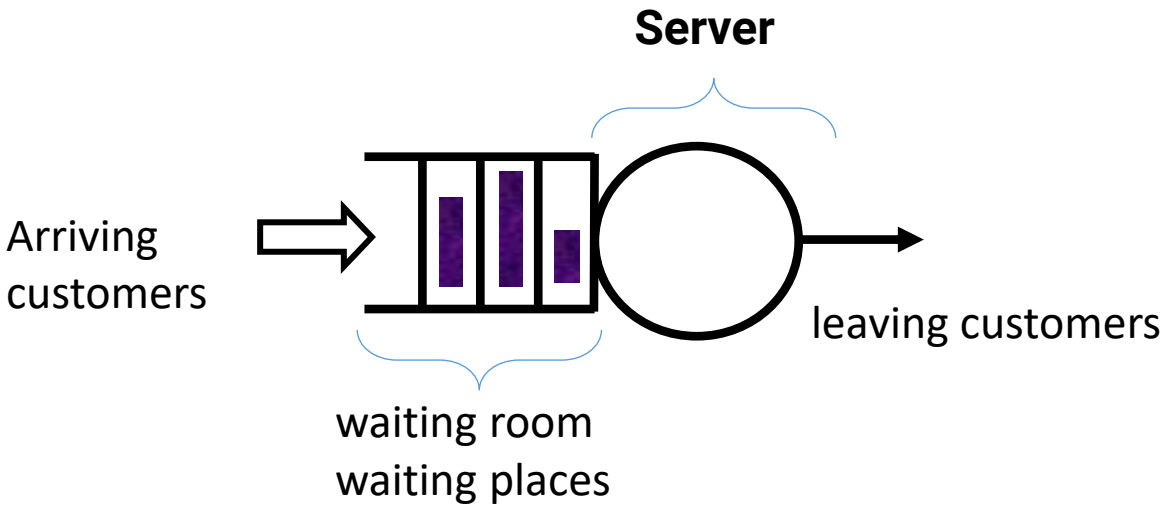
- A queueing discipline is called work conserving, if the capacity of the server / servers is not wasted, i.e. no server is idle if there is at least waiting customer in the system.

- Not all disciplines are work conserving, e.g.
  - LCFS / pre-emptive restart
  - systems, where the server can take a “vacation”

Waiting systems

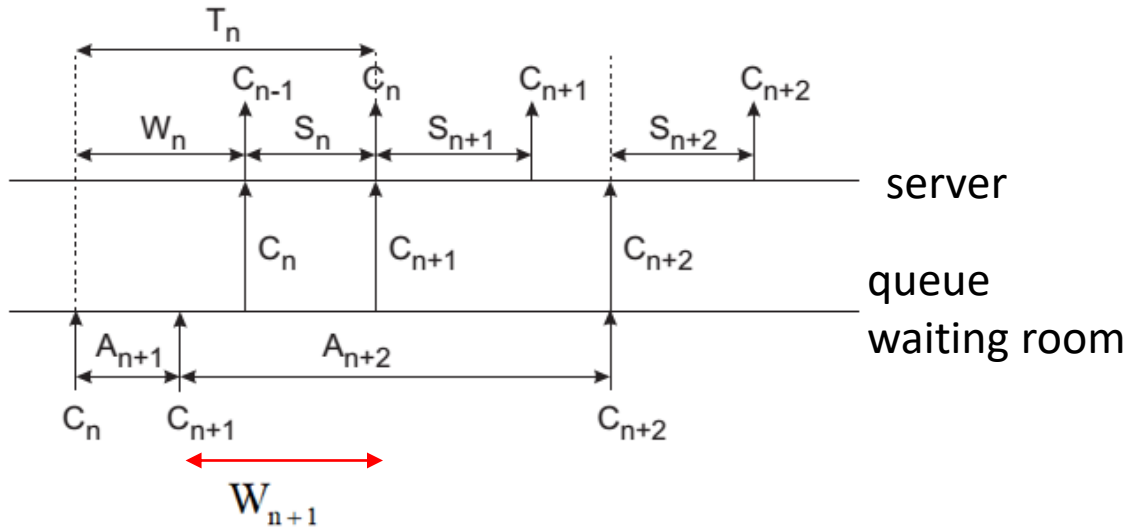
Now we turn our focus on waiting systems. These are the genuine queues where there is a waiting room and the customers may have to wait for the service.

The basic elements of a (single server) queue are as shown in the figure.



Double time axis (in a single server system)

C_n	customer *n*
S_n	service time of customer *n (time it takes to discharge the work)*
X_n	service requirement of customer *n (the work required)*
W_n	waiting time of customer *n*
T_n	*W_n +S_n the total time spent in the system by customer n*
	time in system, sojourn time
A_n	(or t_n) *the interarrival time between customers n − 1 and n*
C	the service rate or capacity of the server (also denoted by *c or μ)*



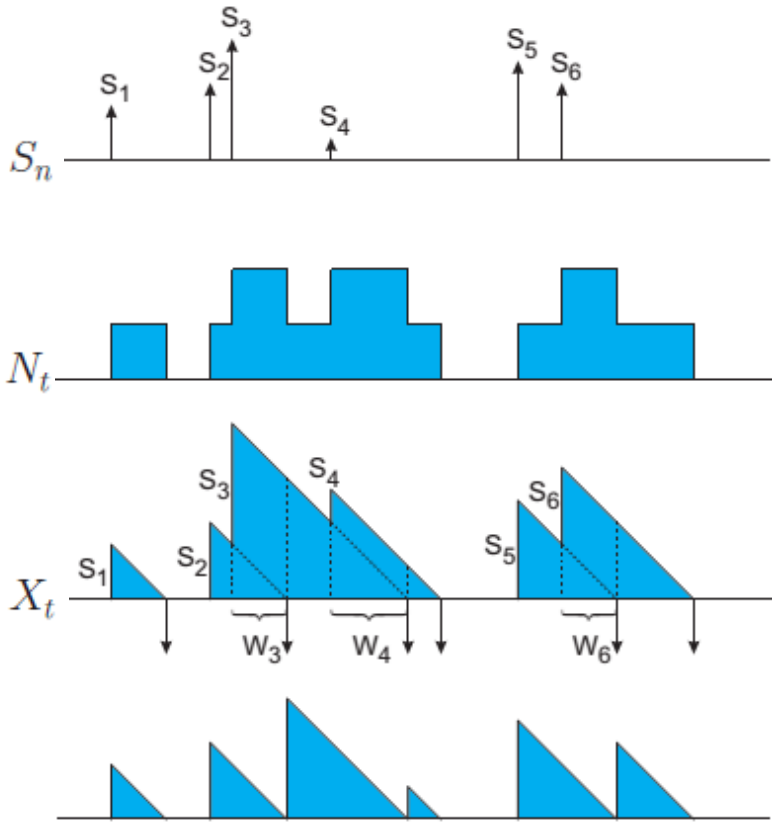
The service time depends on the service requirement (work) and the service rate: $S_n = X_n / C.$

In telecommunication applications the service may mean transmission of a packet on the line. Then the work may be measured e.g. in units of kbit and the service rate is measured in kbit/s.

By inspection, one sees that for FIFO $W_{n+1} = (W_n + S_n - A_{n+1})^+$ where $(x)^+ = \max(x, 0)$

Queue length, unfinished work and virtual waiting time

- N_t (or Q_t *or* L_t *number of customers in* system ("number in system", "queue length")
- S_n service time of customer *n* (time to discharge the work)
- X_t unfinished work (**volume** of the work) in the queue at time *t*
- V_t virtual waiting **time** at time *t*
- W_n the real waiting time of customer *n*
- C the service rate or capacity of the server (also denoted by *c or μ)*



service times, service claims

queue length, number in the system

Virtual waiting time, unfinished work in the system at arbitrary t

remaining period of service (unfinished work) at the server at the instance of a departure

• Virtual waiting time V_t *means the time which* a customer would have to wait for service if the customer happened to arrive at time *t (in* a FIFO queue).

V_t *is the time it takes to discharge the unfinished* work in the queue, $\mathrm{X_t}$, $\mathrm{i.e.,V_t = X_t/C.}$

• In the case of Poisson arrivals the distribution of W_n *is by the PASTA property the same* as the stationary distribution of $\mathrm{V_t}$.