



POLITECNICO
MILANO 1863

QUEUE MANAGEMENT

part 1

Prof. Alberto Portioli Staudacher

Bassel Kassem

Lean Excellence Center www.lean.polimi.it

Politecnico di Milano

Dep. Management, Economics and Industrial Engineering

Bassel.kassem@polimi.it

- ★ What is a queuing system?
- ★ Why does it form?
- ★ Why is it important to study and to manage queues as best as possible?
- ★ What are the queuing system applications?
- ★ How to model a single queue system

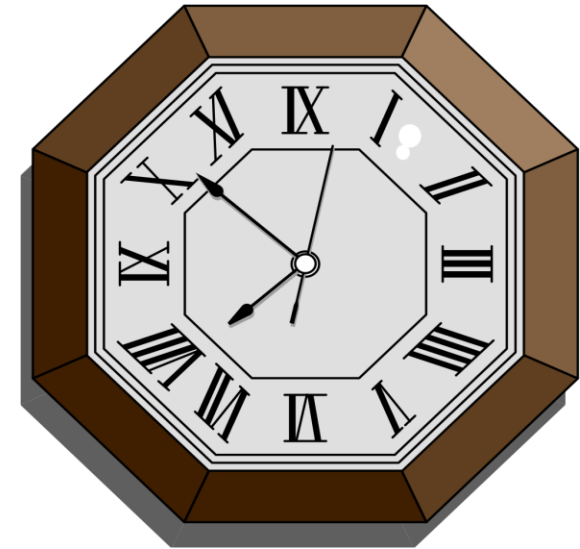
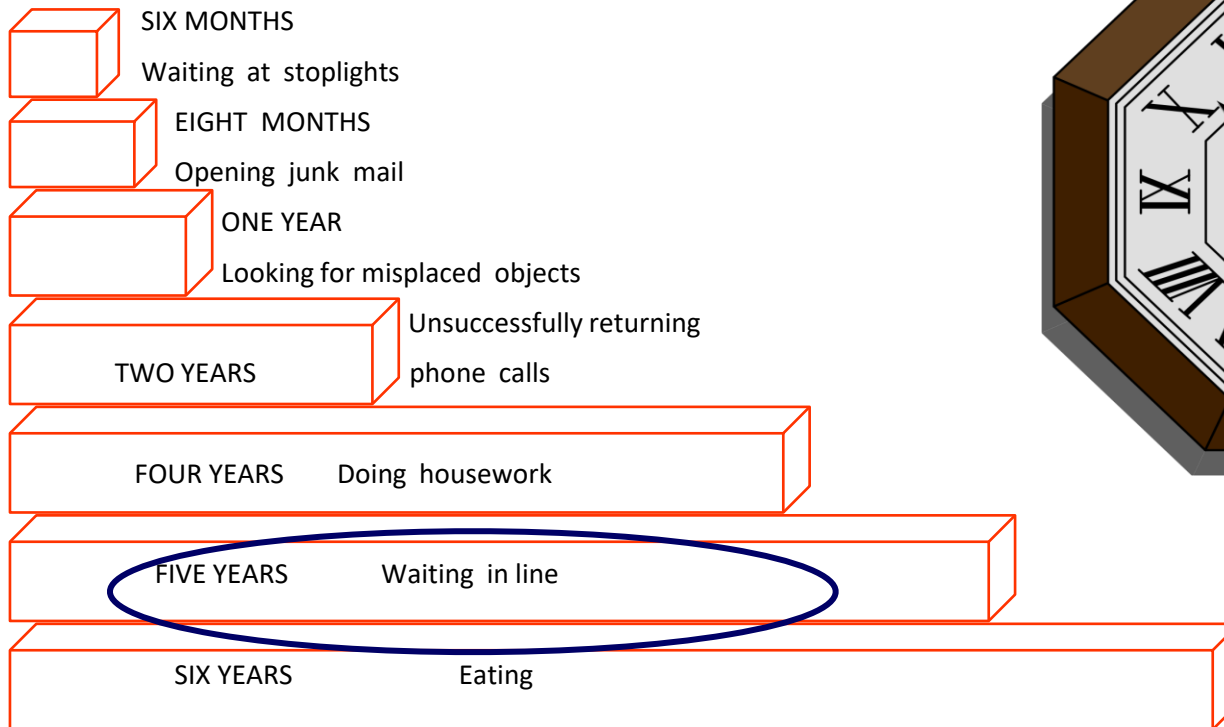
- ✦ How to model a system (in service/manufacturing company) to understand which are the main criticalities of a company and where they are concentrated.
- ✦ How the modelling through queuing theory supports companies, especially service ones, to take better operations decisions (number of resources, bottle-necks, waiting time).
- ✦ How companies (especially service ones) use queueing theory in order to improve the service level for their customers, managing the waiting/throughput time.

- A queueing system is formed by one or more “customers” waiting to be served by one or more “servers”
- Customers’ examples
 - People waiting at the cashier’s desk for a bank operation
 - Pieces waiting to be processed by a lathe in a job-shop
 - Container waiting to be loaded on chassis
- Corresponding servers
 - Cashiers
 - Job-shop’s lathes
 - Overhead traveling cranes’ array for the loading

Queues are unavoidable

5

In a life time, the average
American will spend--



- ✦ Bank historical data show that on Monday there are on average 12 customers/hour at the only counter
- ✦ The counter operator is able to serve 10 customers/hour on average

IS A QUEUE GOING TO FORM ?

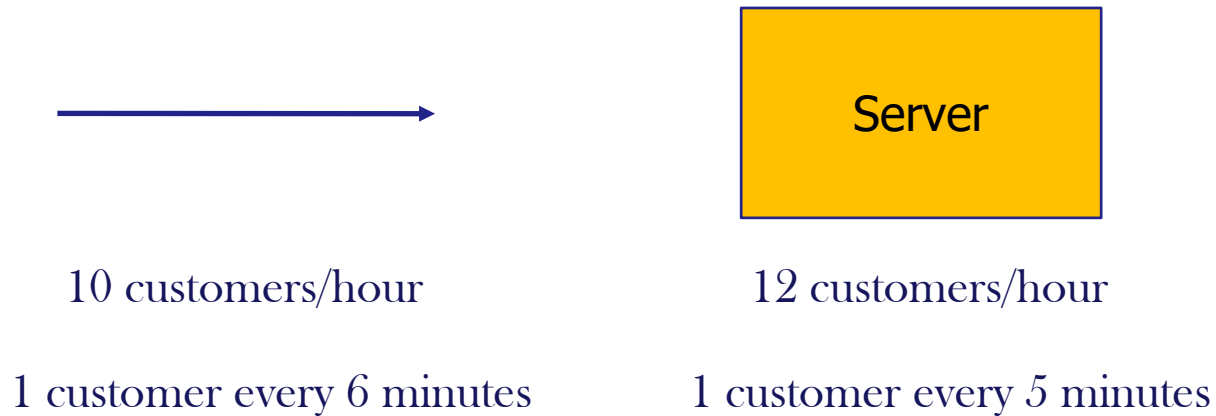
yes !

- ✦ Bank historical data show that on Monday between 10 and 11 am there are on average 10 customers/hour (1 customer every 6 minutes on average) at the only counter
- ✦ The counter operator is able to serve 12 customers/hour on average (1 customer every 5 minutes on average)

IS A QUEUE GOING TO FORM ?

Is it possible to have queue in this system?

8



Queues form due to an imperfect balance between demand rate and service rate

- Structural imbalances
- Incidental imbalances:
 - Variability
 - Uncertainty

Why do we stand in queue?

- Structural imbalance (a case of no variability)

A: customer arrives every 30 min

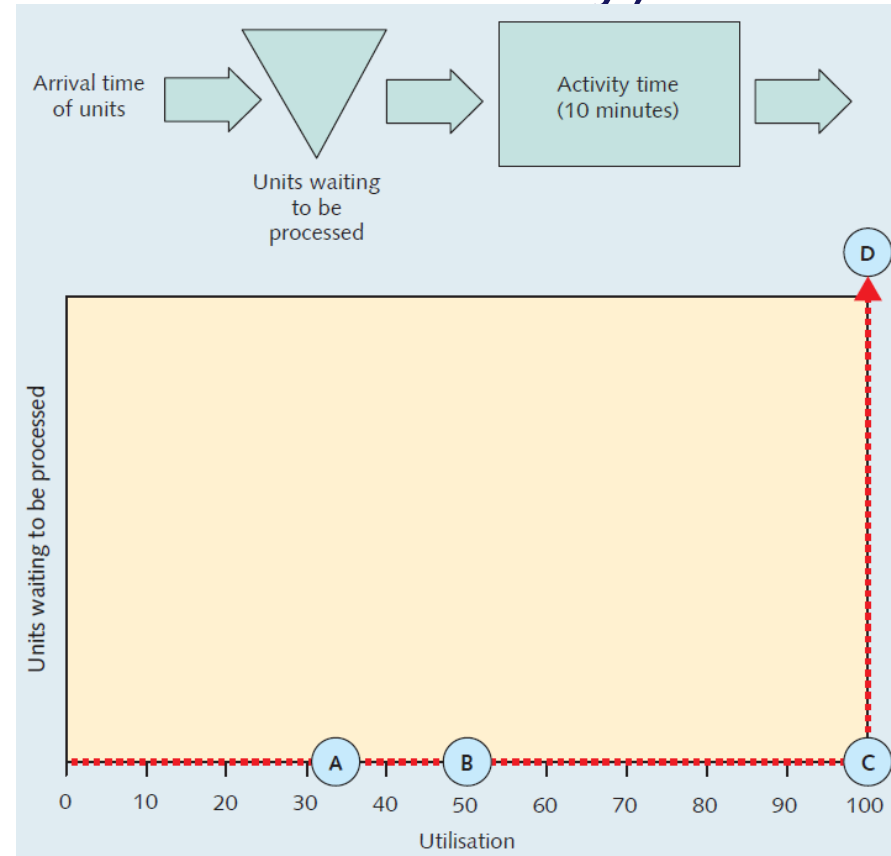
B: customer arrives every 20 min

C: customer arrives every 10 min

D: customer arrives every <10 min

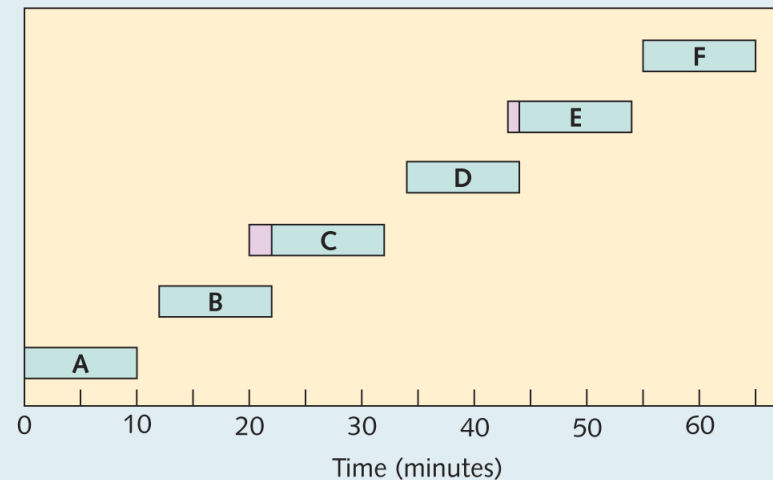
infinite queue

“In a perfectly constant and predictable world, the relationship between process waiting time and utilisation is a rectangular function as shown by the dotted line”



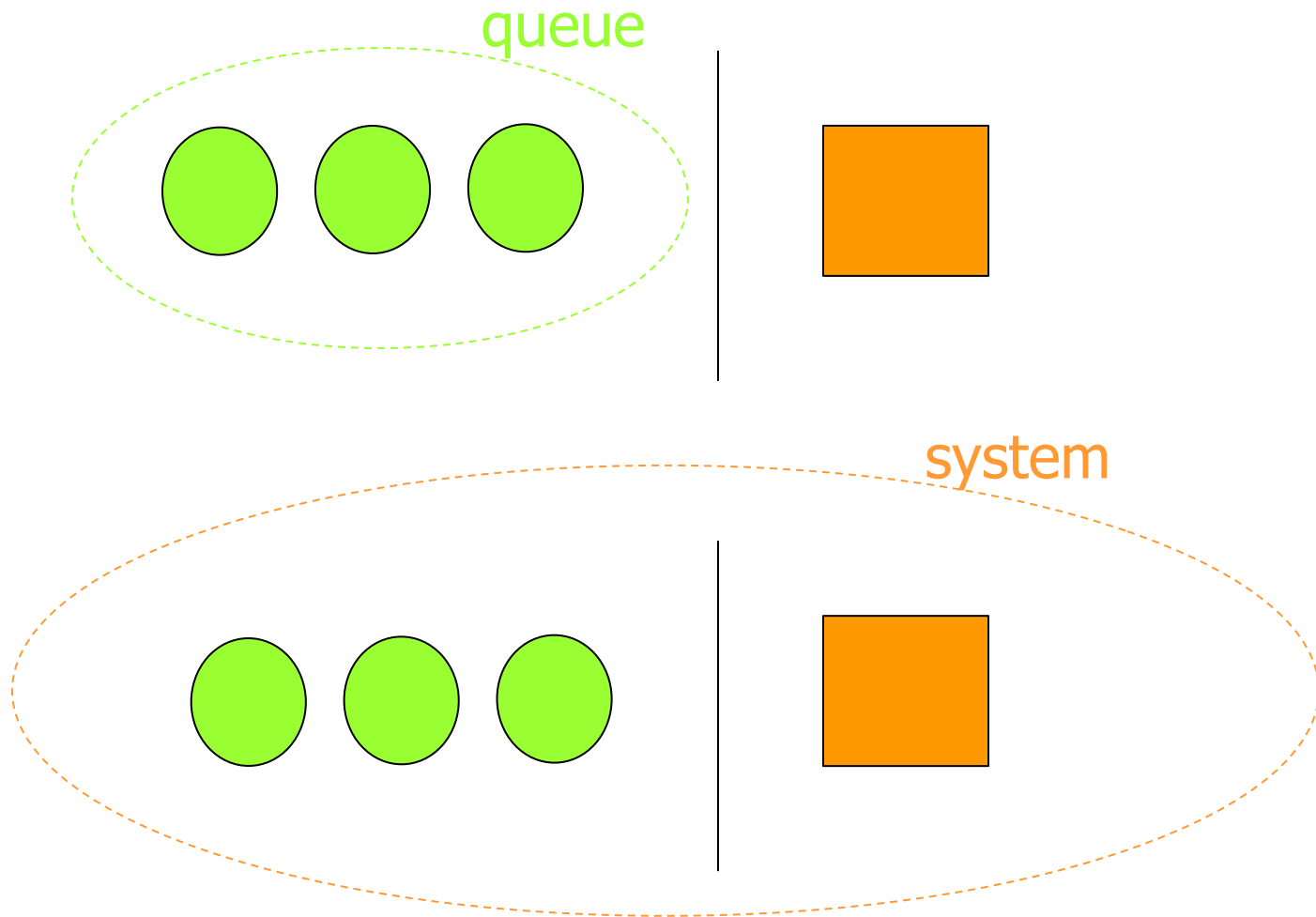
- Incidental imbalance variability

Unit	Arrival time	Start of activity	End of activity	Wait time
A	0	0	10	0
B	12	12	22	0
C	20	22	32	2
D	34	34	44	0
E	43	44	54	1
F	55	55	65	0



→ units waiting to be processed
under-utilisation of the processes' resources
.... over the same period

- ✦ Usually, the tendency is to consider average values only; however, there is always a variability that needs to be considered and managed.
- ✦ Forecast and historical data help in mitigating possible queue impacts on system performances, but they are not the only solution.
- ✦ The structural condition of a system is that the average pace of customers' arrival into the system is lower than the average of the server's serving rate.
- ✦ Queues are a result, a symptom.



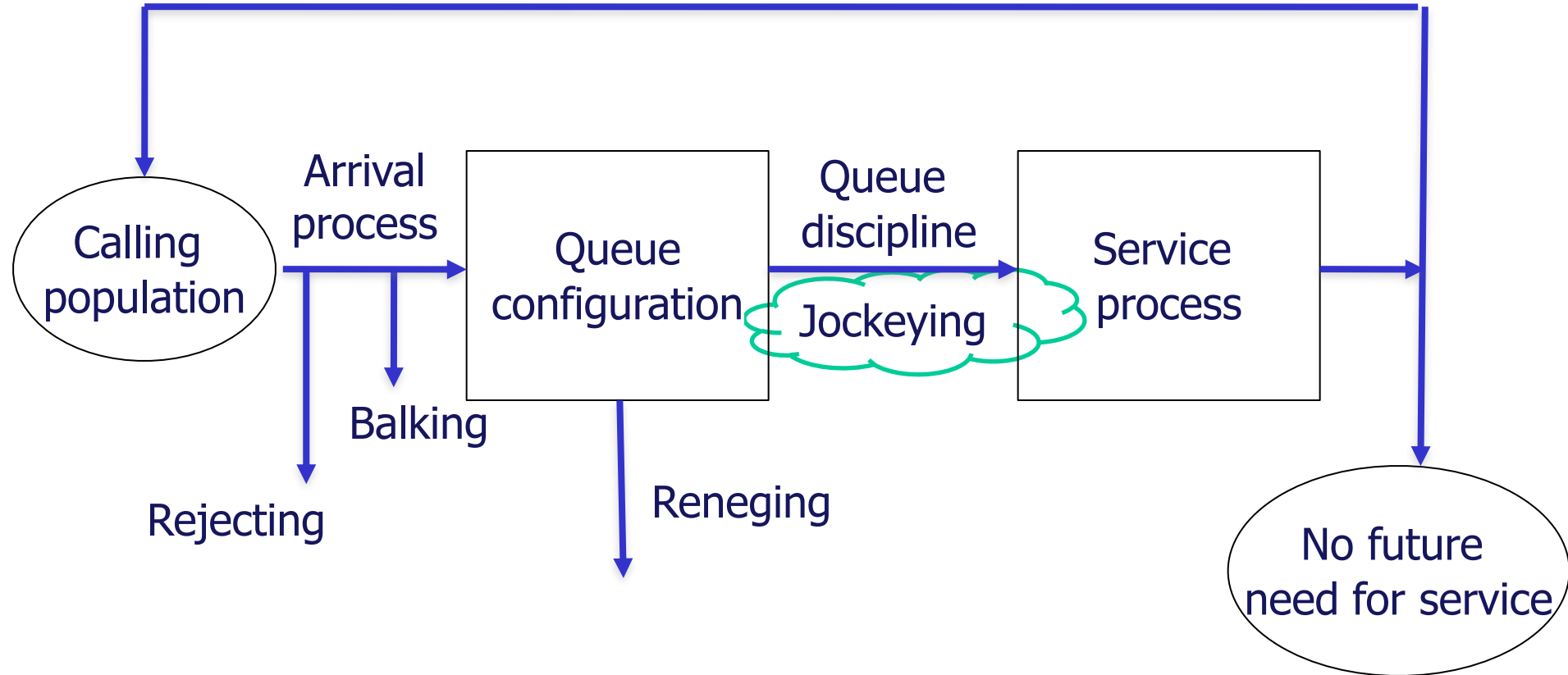
Characteristics	Stocks in a productive system	Waiting lines in services
Costs	Capital opportunity cost	Time opportunity cost
Space	Warehouse	Waiting area
Quality	Inefficiencies index	Negative Impression
Decouplement	Allow to decouple the productive process phases	Allow job division and specialization
Utilization level	WIPs allow to keep machines always busy	Waiting customers keep the servers busy
Coordination	There's no need of a detailed scheduling	Allow to not precisely balance demand and offer

Why are queues particularly critical in the service companies?

15

customers can talk





★ Rejecting

- When a customer is rejected by the system because she doesn't respect acceptance requirements

★ Balking

- When a customer decides not to enter a queue because it's already too long

★ Reneging

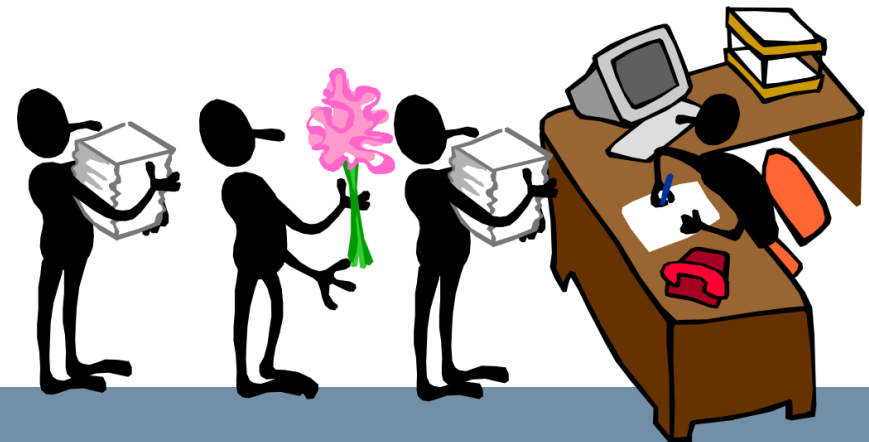
- When a customer already in queue gives up the service and goes away without being served

★ Jockeying

- When a customer tries to trick the rules to get advantage, e.g. shifting from one queue to another to shorten the waiting

Queuing system parts

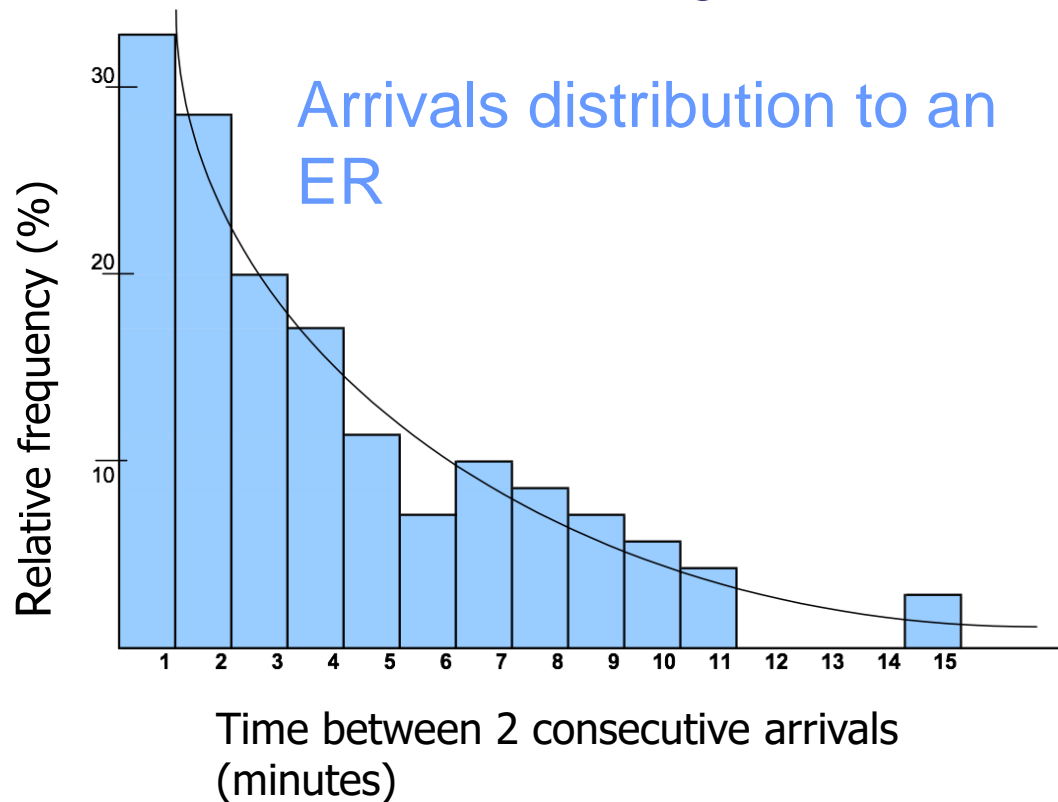
- ★ Calling population
- ★ Arrival process
- ★ Service process
- ★ Queue configuration
- ★ Queue discipline



- ★ Customers' population: it's the input source
 - Finite: if the potential number of new customers for the system is significantly affected by the number of customers already in the system
 - Ex: Arriving to a copier shared between three secretaries
 - Ex: Hotel breakfast
 - Infinite: if the number of clients in the system does not affect the demand pace for the service made by new customers.
 - Ex: people arriving at the ER

Arrival process

The arrival process describes how customers show up. It's described by the distribution of interarrival time, which represents the time intervals occurring between two consecutive arrivals.



Generally, data are given by collecting the actual arrivals time

Different empirical studies show that very often the distribution of inter-arrival times is well described by a negative exponential distribution

The negative exponential distribution has a continuous probability density function as:

$$f(t) = \lambda * e^{-\lambda t} \quad t \geq 0$$

λ = arrival rate: average number of arrival per time unit

t = interarrival time: time interval between 2 consecutive arrivals

$$\text{Mean} = \frac{1}{\lambda} \quad \text{Variance} = \frac{1}{\lambda^2}$$

The cumulative probability function is:

$$F(t) = 1 - e^{-\lambda t} \quad t \geq 0$$

The cumulative probability function describes the probability that the time between 2 consecutive arrivals is t or less than t.

Arrival, negative exponential distribution

The average inter-arrival time between 2 patients to the doctor is 2.4 minutes.

Considering that a patient is just arrived, what is the probability that another will arrive in the next 5 minutes?

$$\lambda = 1/2.4 = 0.4167 \text{ arrivals/minute}$$

$$F(5) = 1 - e^{-\lambda t} = 1 - e^{-0.4167 \cdot 5} = 1 - 0.124 = 0.876$$

Arrival, Poisson distribution

The Poisson distribution is in a one-to-one relationship with the exponential distribution. When the inter-arrivals time are exponential, the number of events $N(t)$ that takes place in a given time t is a Poisson process.

The Poisson distribution has a discrete probability function as:

$$f(n) = P(N(t) = n) = \frac{(\lambda t)^n e^{-\lambda t}}{n!} \quad n=0,1,2,3,\dots$$

The Poisson distribution gives the probability of n arrivals during an interval time t

λ = arrival rate: average number of arrival per time unit

t = number of interest time periods (usually $t=1$)

n = number of arrivals (0,1,2,3,...)

Mean = $t\lambda$ Variance = $t\lambda$

At a telephone exchange phone, calls arrive at a pace of 12 calls per hour. Knowing that the number of calls can be shaped with a Poisson distribution, answer the following questions.

1) What is the probability that 10 calls will arrive in the next hour?

$$\lambda = 12 \text{ calls/h} \quad n = 10 \text{ calls} \quad t = 1$$

$$P(10 \text{ in } 1 \text{ h}) = \frac{(\lambda t)^n e^{-\lambda t}}{n!} = \frac{(12 * 1)^{10} * e^{-12 * 1}}{10!} = 0.148$$

2) What is the probability that there won't be any call in the next 5 minutes?

$\lambda=12$ calls/h $n=0$ calls $t=5$ minutes

$$5 \text{ min} * \frac{1}{60 \frac{\text{min}}{\text{h}}} = 0.083 \text{ h}$$

$$P(0 \text{ in } 0.083 \text{ h}) = \frac{(\lambda t)^n e^{-\lambda t}}{n!} = \frac{(12 * 0.083)^0 * e^{-12 * 0.083}}{0!} = 0.3679$$

3) What is the probability to have more than 2 calls in the next 10 minutes?

$$P(N > 2) = 1 - P(N \leq 2) = 1 - [P(N=0) + P(N=1) + P(N=2)]$$

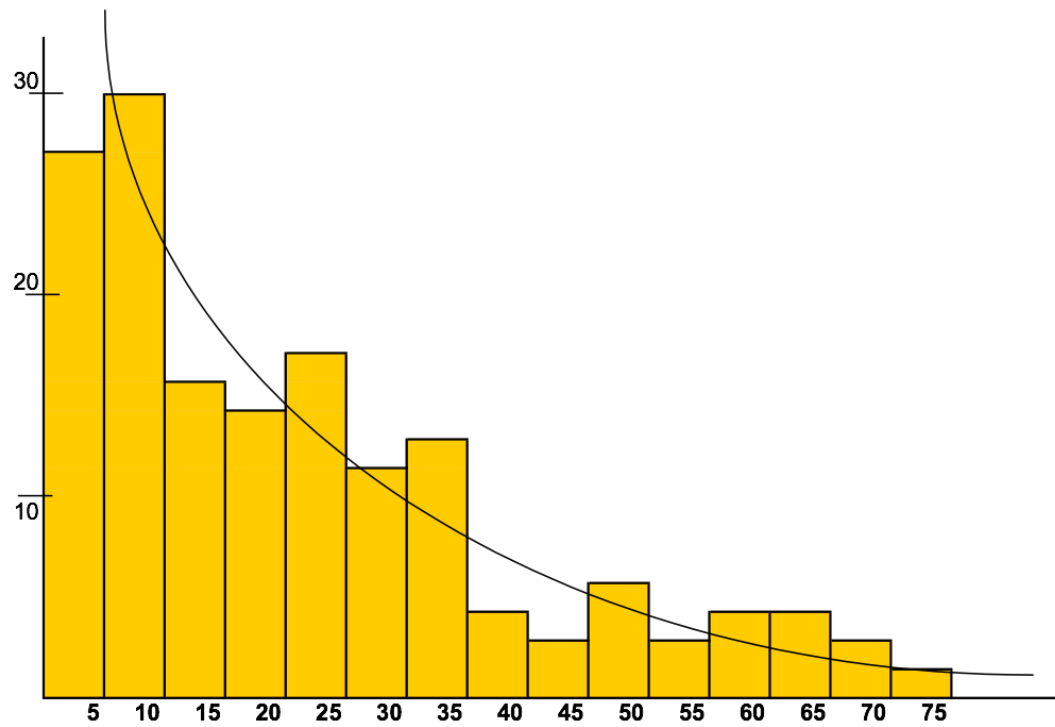
$$10 \text{ min} * \frac{1}{60 \frac{\text{min}}{h}} = 0.167h$$

$$P(0 \text{ in } 0.167h) = \frac{(12 * 0.167)^0 * e^{-12 * 0.167}}{0!} = 0.135$$

$$P(1 \text{ in } 0.167h) = 0.27013 \quad P(2 \text{ in } 0.167h) = 0.27067$$

$$P(N > 2) = 0.3242$$

The service process describes how every server delivers the service, in particular it defines its duration. It's defined in terms of service times distributions.

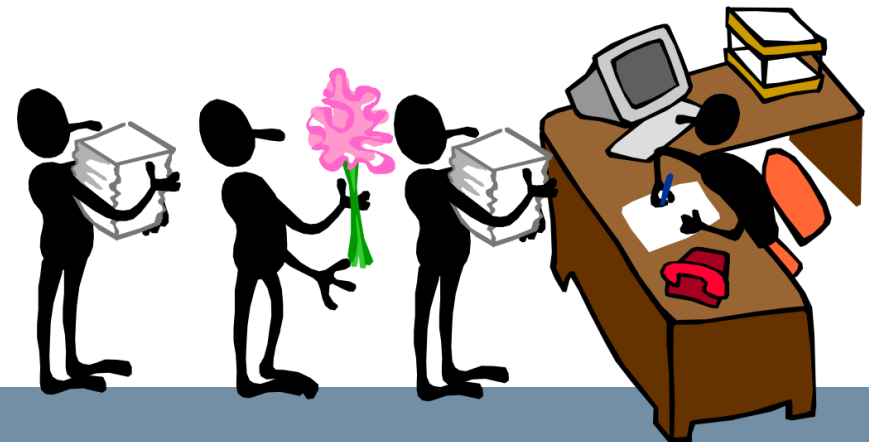


Usually data are given by collecting the actual service times

The service times distribution is often well represented by a negative exponential distribution

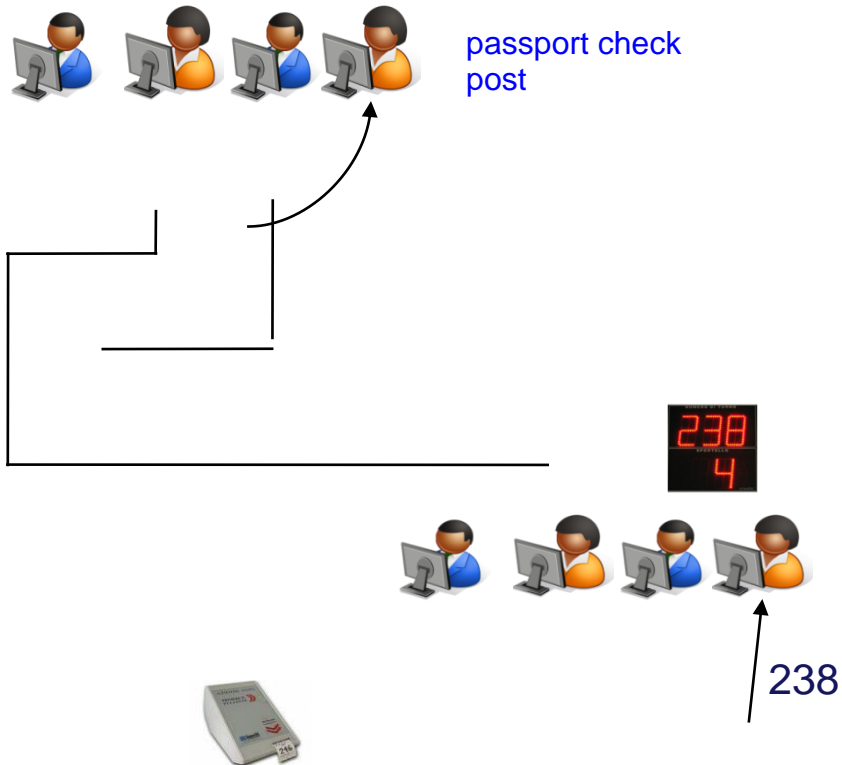
Queuing system characteristics

- ★ Calling population
- ★ Arrival process
- ★ Service process
- ★ Queue configuration
- ★ Queue discipline



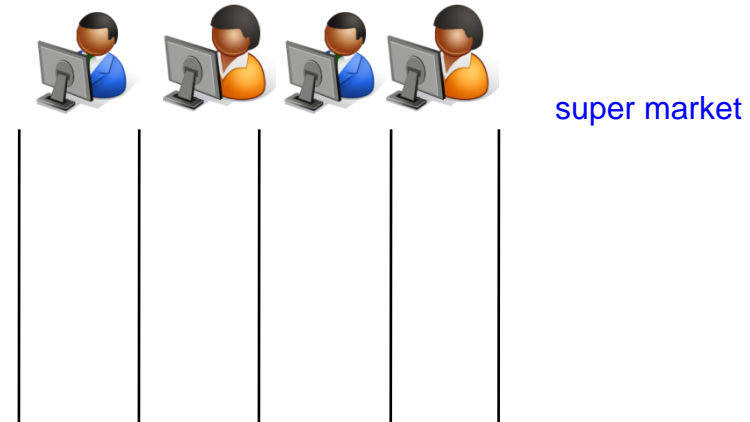
- ★ Queue is formed by customers waiting to be served
- ★ It's possible to distinguish between
 - Queue configuration
 - Single queue
 - Multiple queue
 - Take a number
 - Endless servers
 - Queue capacity
 - Limited queue
 - Unlimited queue

1. Single queue



3. Take a number

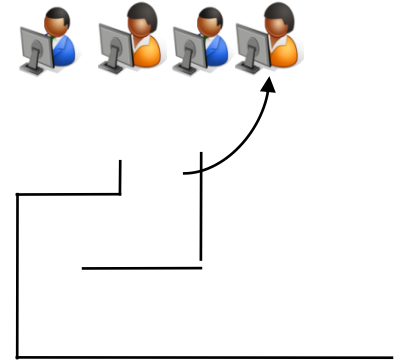
2. Multiple queue



4. Infinite Servers

Call center
online shops

SINGLE QUEUE

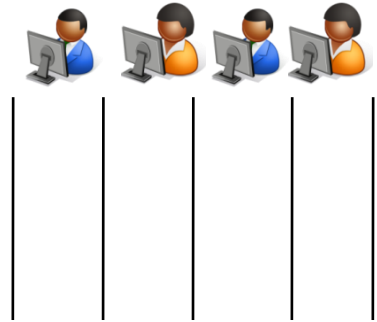


Pros

- Assure a FCFS type of service
- Avoid concern that another queue could be faster
- Minimize the average waiting time
- Reneging actions are less frequent

Cons

- It could "scare the customer"



MULTIPLE QUEUE

Pros

- Allow to diversify the service
- Allow to diversify the work
- Customer can chose a server of his/her liking
- Balking actions are less frequent

Cons

- With the same number of servers, the average waiting time is greater with respect to single queue

TAKE A NUMBER: it's a variation of the single queue



Pros

- Assure a FCFS type of service
- Avoid anxiety related to that another queue could be faster
- Minimize the average waiting time
- Customers have the possibility to relax and dedicate themselves to other things during the wait

Cons

- An absent-minded customer could risk losing its turn
- It could "scare the customer" (if he/she doesn't have anything else to do)

Some example of queue configuration

34

Type of service

Type of queue

Parking lot

Self-service

Cafe

Multiple queue

Toll booth

Multiple queue

Supermarket

Self-service for the acquisition from shelf,

Take a number at the fish counter

Multiple queue at the cash registers

Hospital

Complex system with different type of queues

- ★ Service customization degree
 - ↑ customization → best multiple queue
- ★ Service times variability
 - ↑ variability and unpredictability → best single queue

Limited
queue:

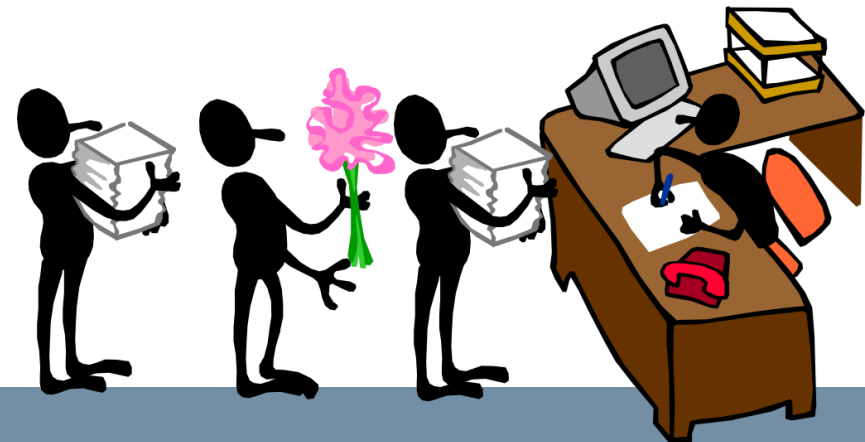
The number of customers in line plus the ones that are being served is limited. Customers who arrive, after the capacity is overfilled, are rejected.

Unlimited
queue:

The number of customers in line plus the ones who are being served is potentially unlimited.

Queuing system characteristics

- ★ Population
- ★ Arrival process
- ★ Service process
- ★ Queue configuration
- ★ Queue discipline

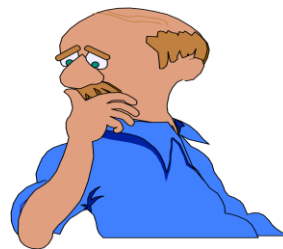


- ★ Queue discipline (or “ranking rule”) is the rule or set of rules related to the order in which customers are served, it's often strictly related to the queue configuration
 - Static: the pertaining order among the customers in queue doesn't change in time and/or at the changing of the system conditions (ex. FCFS,...)
 - Dynamic: the pertaining order between the customers can change in time and/or at the changing of the system conditions

✦ QUEUE SYSTEMS ANALYSIS TOOLS

- ✦ Deterministic Analysis
- ✦ Queueing Theory
- ✦ Simulation (hints)

increasing degree of
precision and detail



✦Pros

- Simple and intuitive to apply

✦Cons

- Doesn't take transitory into account
- Doesn't take arrivals and service times variability into account
- Queue as an ON-OFF aspect

✦ Pros

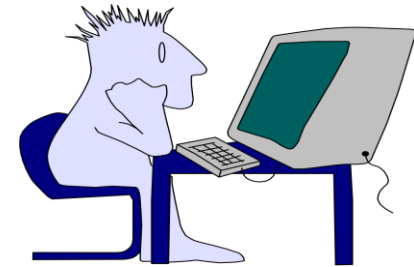
- Takes inter-arrivals and service times variability into account
- Allows to calculate a set of significant variables (W_q , L_q , P_n etc..)

✦ Cons

- Doesn't take transitory into account
 - Mathematic analysis can become very complex or even intractable in case of complex waiting lines
 - Customers complex behaviors (balking, reneging, jockeying)
-
- In these cases simulation can become a useful analysis tool

★Pros

- Takes transitory into account
- Very flexible
 - For the type of systems that can be shaped
 - For the type of data that can be obtained (ex. times (%) that a specific customer waits more than 1 minute)
 - Economic and user friendly softwares are always more widespread



★Cons

- Time consuming
- Skills needed to design, production and analysis

✦ Stationary process

- This assumption could be very restrictive: usually the arrival process isn't a static process, it changes during the day. For example, in a bank, there are very different values of arrival rates in a whole day
→ a possible solution is to make more analysis considering different time slot.

✦ Peculiar customers behaviors are not expected

✦ No balking, reneging, or jockeying

✦ True only for certain inter-arrival and service times distributions

Kendall's codification

Kendall's codification turns out to be $A/B/c/K/m/Z$

A identifies customers' arrival distribution

B identifies service times distribution

c identifies number of servers

K identifies queue capacity (buffer by default: endless)

m identifies population dimension (by default: endless)

Z identifies service discipline (by default: FIFO)

Usually stop at the first 3 components (ex. $M/G/1$), in detail

M suggests a negative exponential distribution (Markovian)

G suggests a Gaussian distribution

Queue and system..

A queue system is formed by 1 queue and one or more servers to offer the service. Therefore, hypothesizing that arrivals and service rates are deployed as negative exponentials:



M/M/4 system



one system
M/M/1/3 type



4 system M/M/1
type

★ Standard M/M/1

Most common

★ Standard M/M/c

★ Finite-queue M/M/1 (M/M/1/K) Particularly useful in estimating lost sales due to an inadequate waiting area or to an excessive long queue

★ Finite queue M/M/c (M/M/c/K)

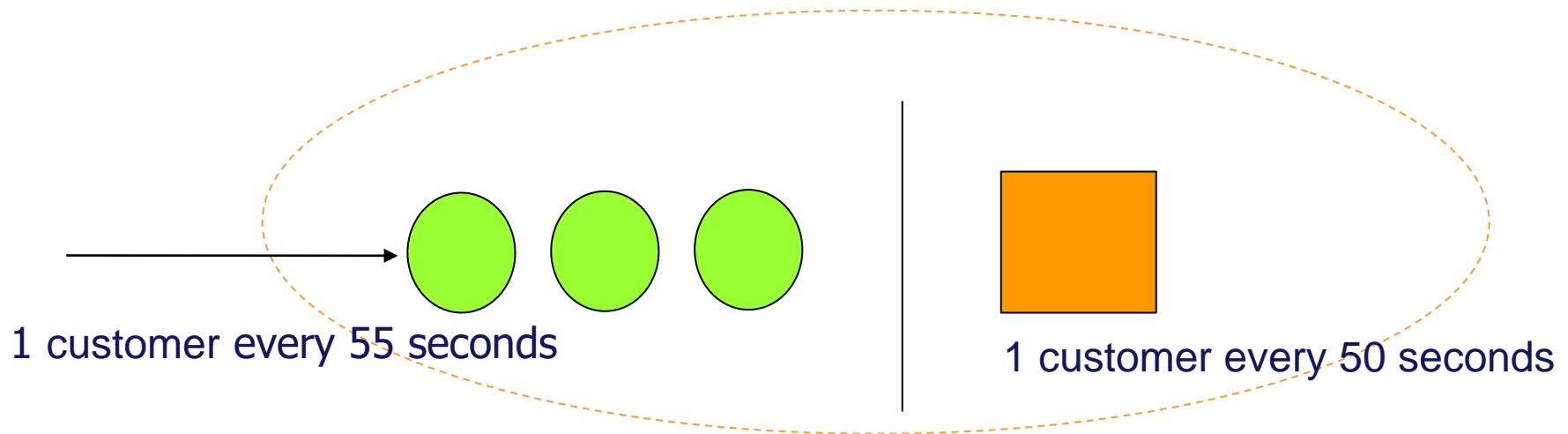
★ General Self Service M/G/ ∞

The total number of customers in the process varies due to arrival and service times variability. This model is also useful to shape, rounding situations where you rarely have to wait (ex. Emergency services).

...some numerical examples

Pierre Cahon, barman of one of the most popular discotheque of St. Catherine Street, Montreal, serves cocktails at a pace of 1 every 50 seconds. Recently, during an evening when the club was particularly full, every 55 seconds someone was at the bar asking for a drink.

- How to shape this system?



Which are the relevant parameters?

- λ Arrival rate (customer/time unit)
- μ Service rate (customer/time unit)
- n Number of customers in the system (**waiting + serving**)
- L_q Average number of customers in line
- W_q Average waiting time in line by the customer
- L_s Average number of customers in the system
- W_s Average waiting time in the system by the customer

Which are the relevant parameters?

50

$$\frac{\lambda}{\mu * s}$$

System utilization rate

It's the coefficient that describes the utilization degree of the system, where S represents the number of servers

λ = arrivals rate

μ = service rate

P_n Probability that there are n customers in the system

P_0 It indicates the probability not to have any customer in the system

...some numerical examples

In the Pierre Cahon's example case:

- a) Assuming that Pierre is able to serve customers following a FCFS method, how much would you expect to wait before you'll have your drink in hand (W_s)?
- b) How many people would you expect to find in line on average?
- c) What is the probability that three or more people will be in the system?
- d) What is the probability that you can be served immediately without none in line?
- e) What is Pierre's saturation level? (how busy is he?)

they are given in the exam

P_n =probability that there are n customers in the system

$$\left\{ \begin{array}{l} P_0 = 1 - \rho \\ P(n \geq k) = \rho^k \\ P_n = P_0 \rho^n \end{array} \right.$$

L_s =average number of customers in the system:

$$L_s = \frac{\lambda}{\mu - \lambda}$$

L_q =average number of customers in line:

$$L_q = \frac{\rho \lambda}{\mu - \lambda}$$

W_s =average time waiting in the system by the customer:

$$W_s = \frac{1}{\mu - \lambda}$$

W_q =average time waiting in line by the customer:

$$W_q = \frac{\rho}{\mu - \lambda}$$

$$\lambda = 60 / 55 = 1.09 \text{ customers/minute} \quad \mu = 60 / 50 = 1.2 \text{ customers/minute}$$

- a) Assuming that Pierre is able to serve customers following a FCFS method, how much would you expect to wait before you'll have your drink in hand (W_s)?

$$W_s = [1 / (\mu - \lambda)] = 9.09 \text{ minutes}$$

- b) How many people would you expect to find in line on average?

$$L_q = [(\rho * \lambda) * (1 / (\mu - \lambda))] = 9.0 \text{ customers}$$

- c) What is the probability that three or more people will be in the system?

$$P(n \geq 3) = (\lambda / \mu)^3 = 0.75$$

$$\lambda = 60 / 55 = 1.09 \text{ customers/minute} \quad \mu = 60 / 50 = 1.2 \text{ customers/minute}$$

d) What is the probability that you can be served immediately without none in line?

$$P_0 = 1 - (\lambda / \mu) = 0.092$$

e) What is Pierre's saturation level? (how busy is he?)

$$\lambda / \mu = 0.908$$

...some numerical examples

- ✦ A consultant is searching for a way to understand the exact number of counters to leave open in a new branch bank to offer a good service to the customer. How many counters should be left open in the rush hour if you want to keep the waiting time in line below 15 minutes? Each counter will have its own customers' queue in front of it.
- ✦ The consultant, from previous surveys, knows that during the rush hour there are on average 100 customers per hour and that each server at the counter takes on average 5 minutes to be able to serve each customer (arrival and service distribution are negative exponential).

...some numerical example

λ	100
μ	12
λ/μ	8,33

customers /hour

customers / hour

Which preliminary indications can we give to the consultant based on these data?

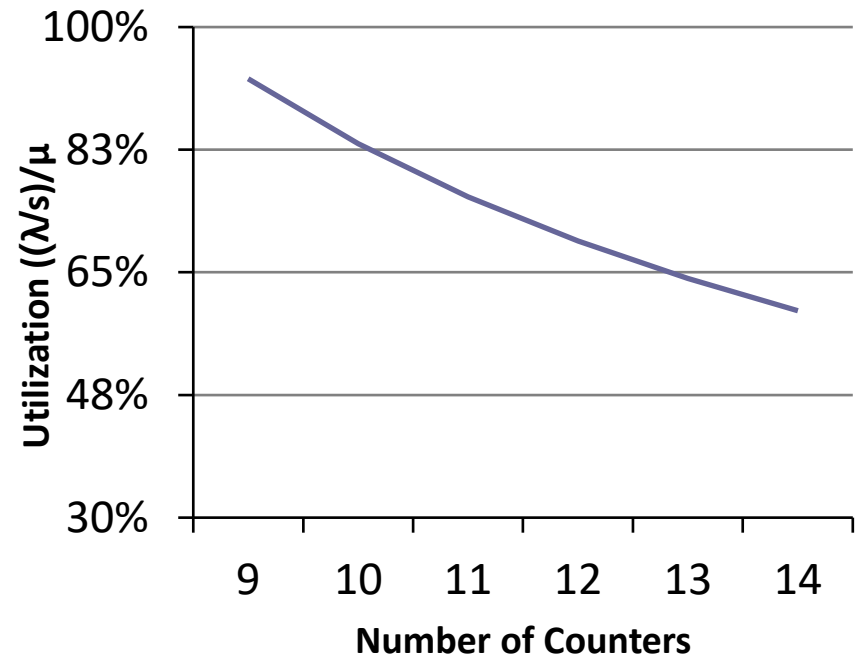
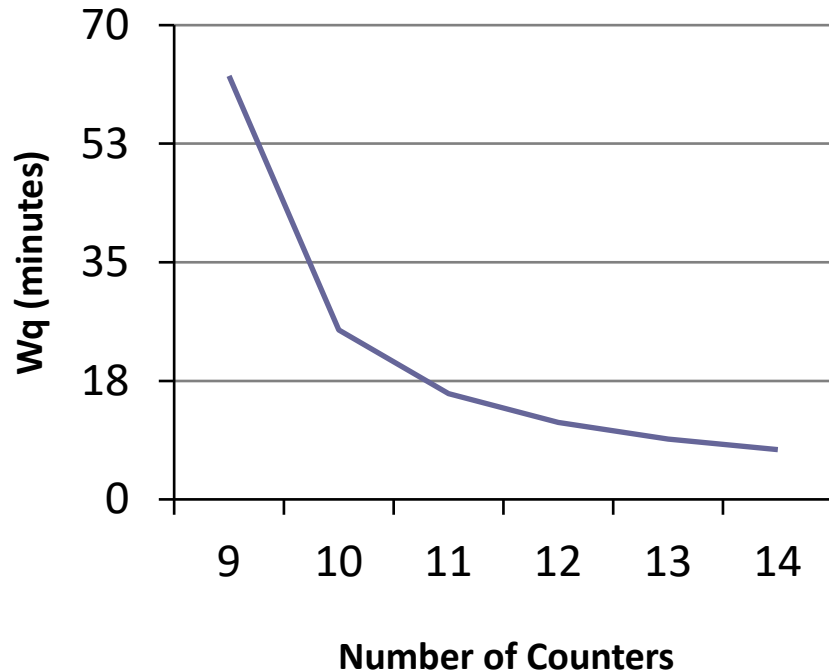
→ Minimum number of servers has to be greater than 8,33 to avoid a structural imbalance

→ At least 9 servers

...some numerical example

λ	100
μ	12
λ/μ	8,33

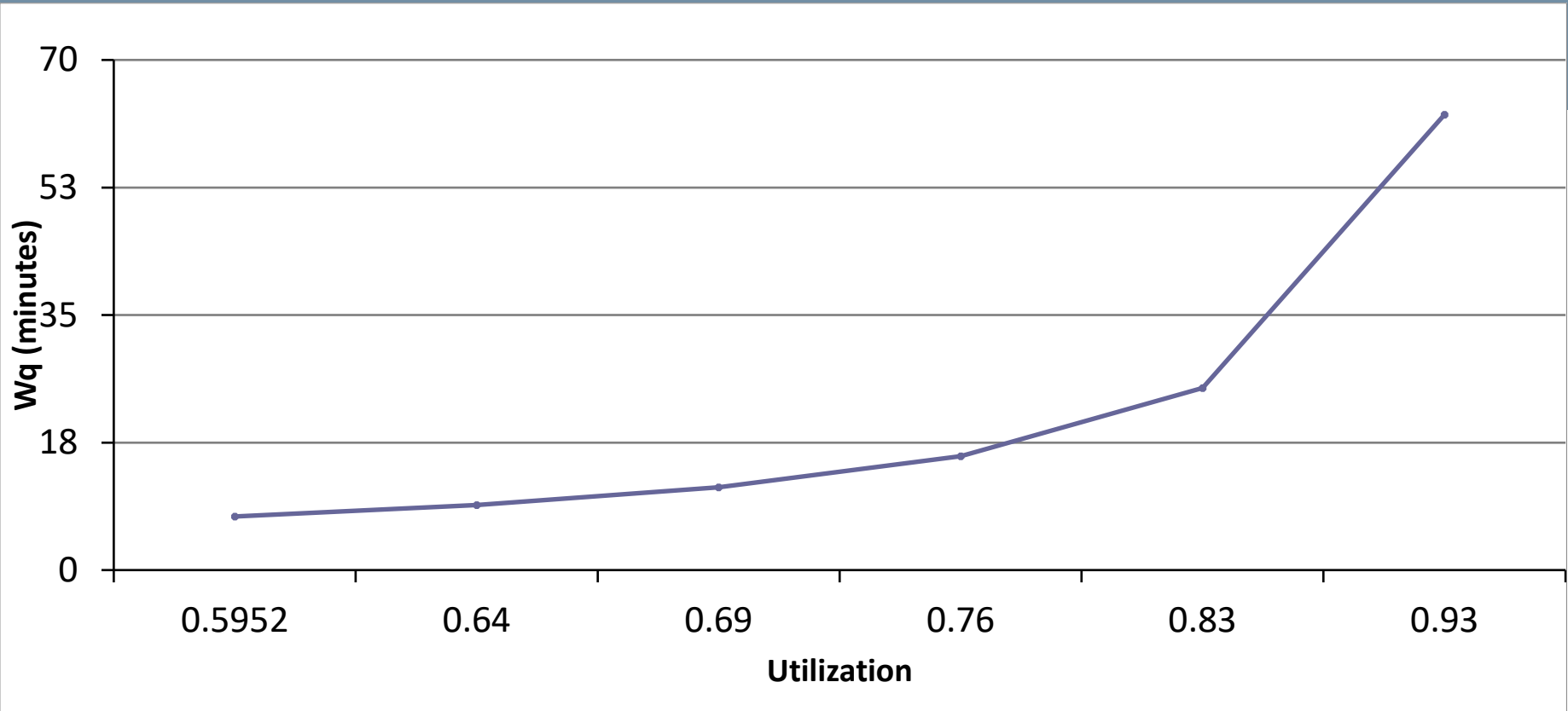
	Number of counters					
	9	10	11	12	13	14
$(\lambda/s)/\mu$	92,59%	83,33%	75,76%	69,44%	64,10%	59,5%
Wq (min)	62,5	25	15,625	11,36	8,93	7,35
Ws (min)	67,5	30	20,625	16,36	13,93	12,35



Increasing by 1 the number of servers from 9 to 10 (+11%) decreases the throughput time of 60%!!

- THE TREND IS NOT LINEAR -

Waiting time vs saturation



- The waiting time doesn't grow in a linear way with the increase of the system saturation time.
- Even with a system utilization rate rather low (around 60%) there will be waiting times!!

- ✦ **System sizing and designing** (ex. consultant who needs to understand how many counters at least have to be left open)
- ✦ **Evaluation of costs and gains to improve the service** (ex. barman who tries to understand how to make his customers more satisfied)
 - Queue length (Long lines suggest a poor customer service and cause congestion and dissatisfaction)
 - Average waiting time in line (Long waits are linked to a poor service)
 - Average waiting time in the system (it can suggest problems with customers, servers' efficiency or capacity)
 - System utilization rate (Check costs without unacceptable reductions in services)

★ “Service quality” principles

– Average waiting time of the customer

- Banks, restaurants

$$E(W^*) < \beta$$

Ex. sizing with the goal that the expected value of the time that the customer has to wait before sitting at a table is less than 10 minutes (check ex of the bank)

– Maximum waiting time

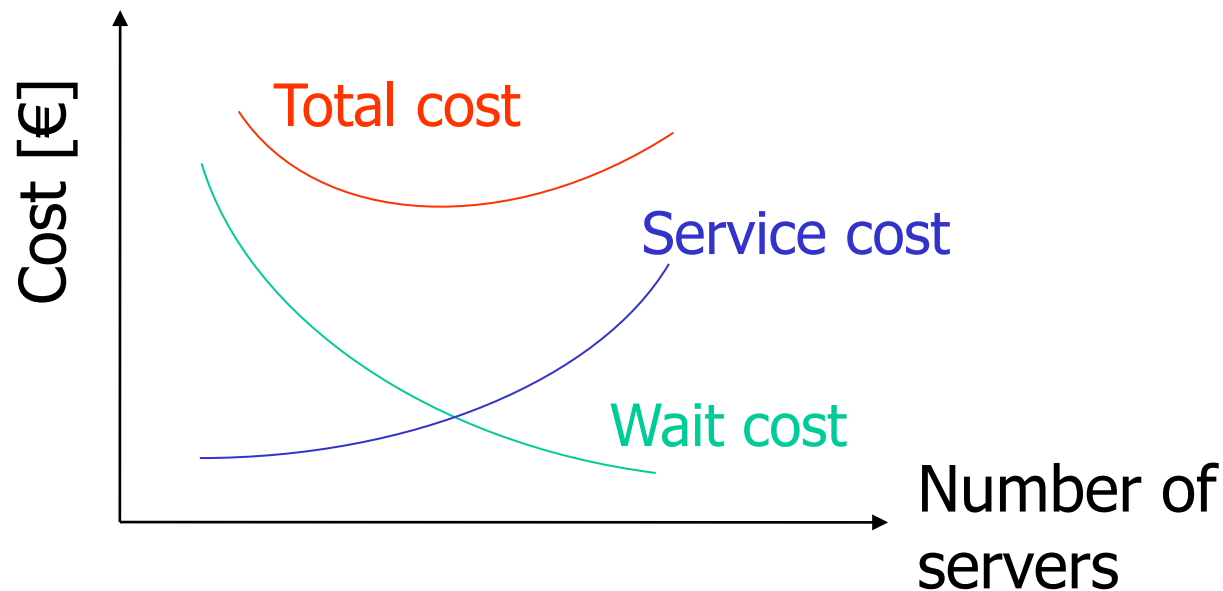
- Public services (ambulance, fire brigade)

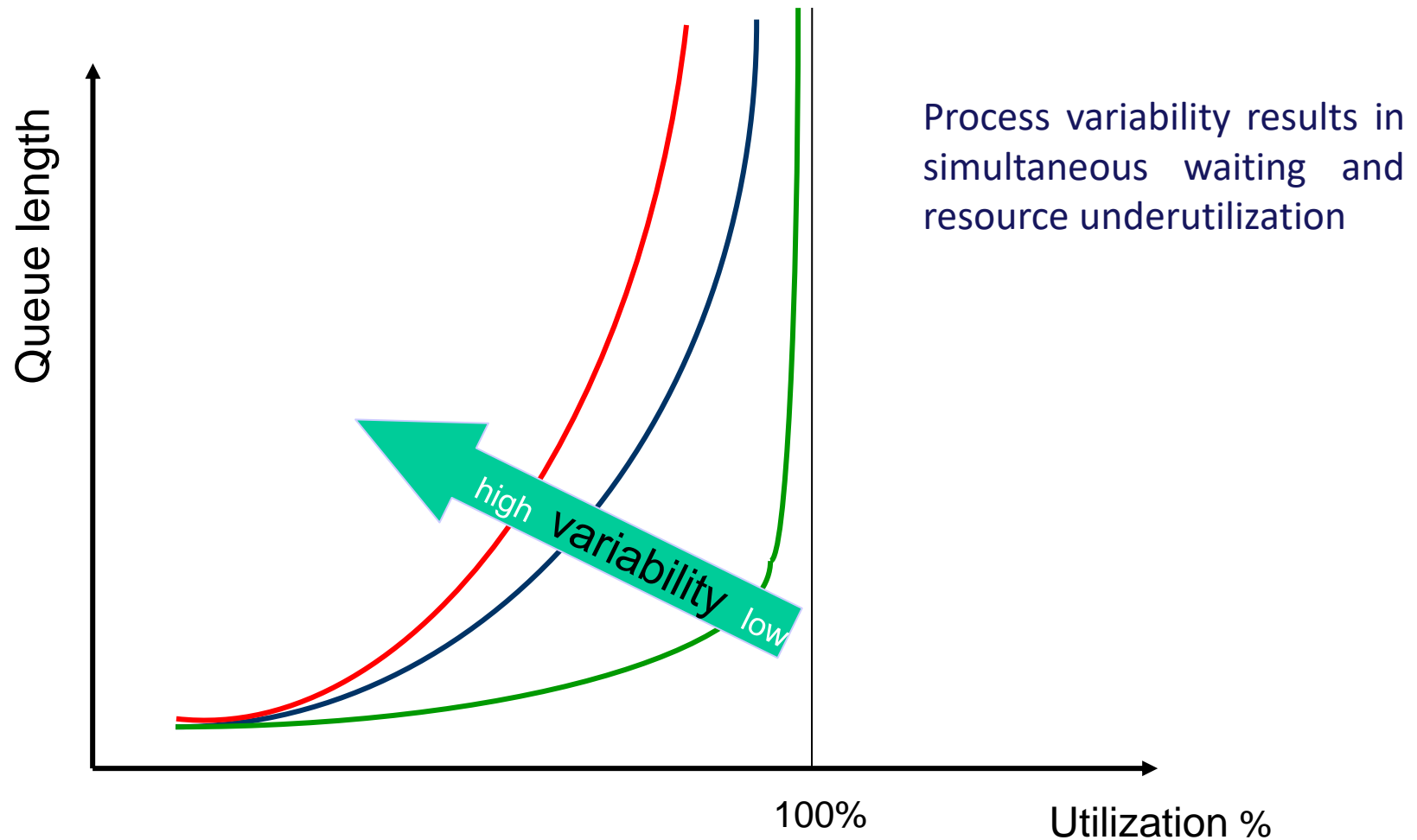
$$P(W^* > w) \leq \alpha$$

Ex. sizing with the goal of wanting that the probability that the client has to wait more than 5 minutes before receiving the service is less than 5%-heart attack ex

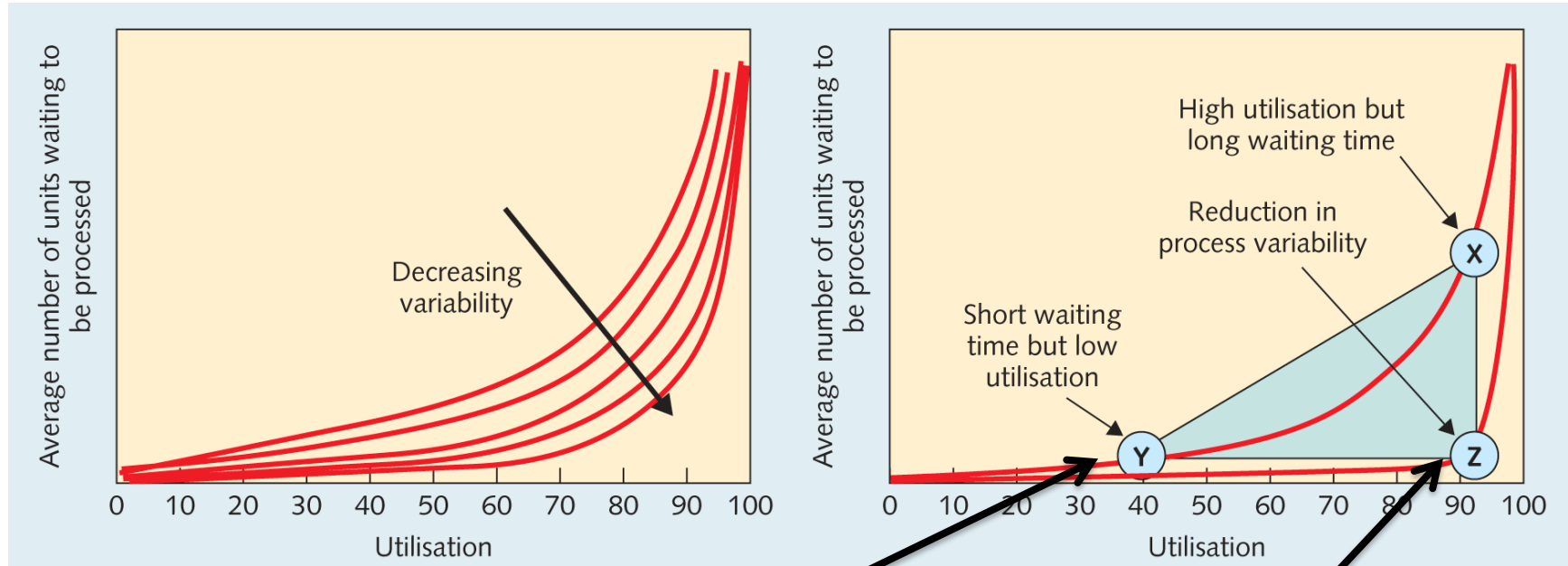
★ Cost principles

- Minimization of the sum of the wait and service costs (trade off)





“(X) is the AS-IS system performance. The queue is too long (then also the waiting time is too long). We set a new queue length target. How to get there?”



Two ways to reach the same target:

- (Y) utilization reduction: e.g. increasing the number of resources, same system conditions.
- (Z) system variability reduction: e.g. booking system introduction, same number of resources.

variability = different between actual and average
Uncertainty = different between actual and expected

✦ MANAGERS MODIFY QUEUEING
SYSTEM IN ORDER TO GET
BETTER PERFORMANCES

What set of levers does the manager have to take to manage queues?⁶⁶

- ✦ Levers and countermeasures-offer side
- ✦ Levers and countermeasures-demand side
- ✦ Levers and countermeasures for soft aspects/related to the psychology of waiting

★Service

- Configuration changes
- Add/Remove/Move resources
- Decrease service time
 - Technology utilization
 - Training
- Increase resources flexibility
- Decrease service times variability

✦Arrivals

- Decrease uncertainty level
 - Booking
 - Improve forecasts
- Decrease variance
 - Incentives for non rush hours moments
 - Booking

Service laws

I. Satisfaction=Perception-Expectation

Perception is more important than reality

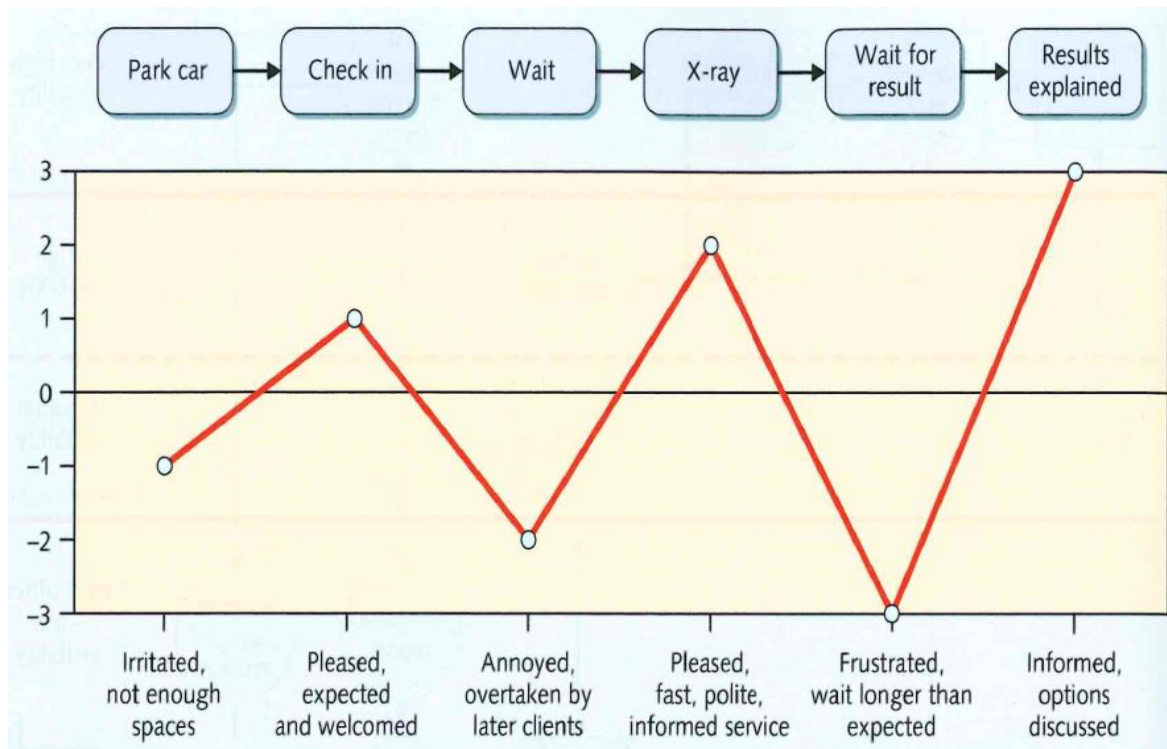
II. It's hard to play catch-up ball

First impression is the most important



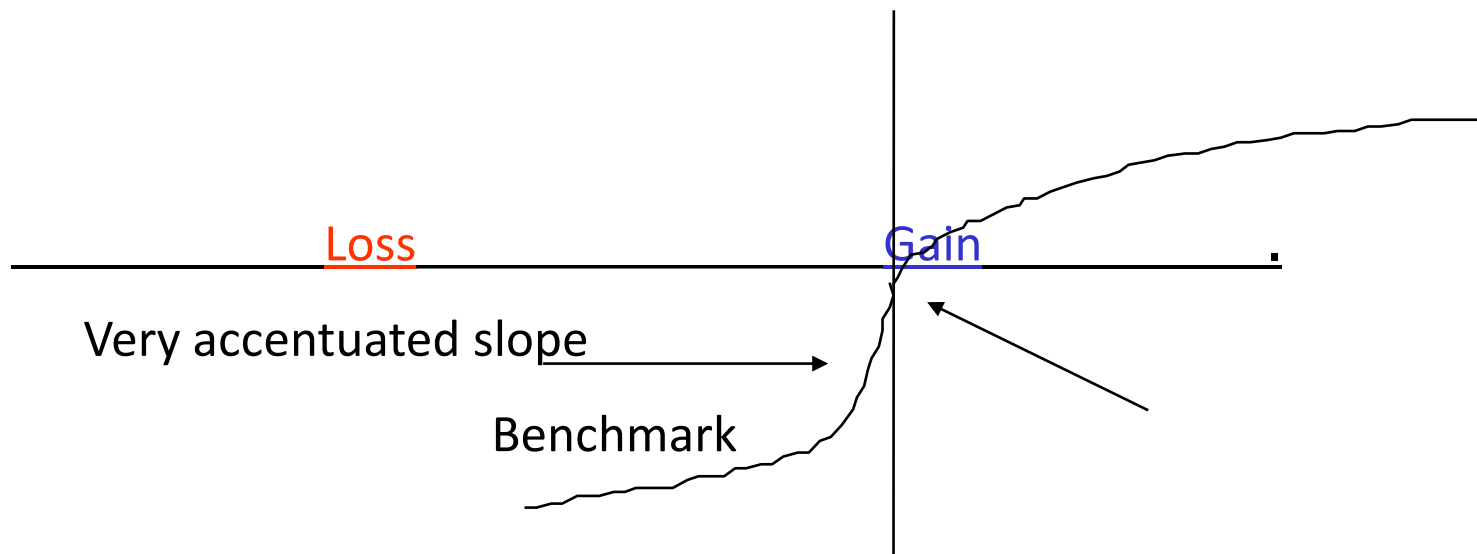
Come on! Get
in line!

Measure the customer satisfaction is important to understand where the critical point of the system is. This is a customer experience map of a visit for an X-ray investigation



Prospect theory (Tversky and Kahneman, contemporary psychology)

Individual Utility






- ✦ Gain or loss perception depends on the benchmark
- ✦ Loss sensitivity is greater than gain sensitivity

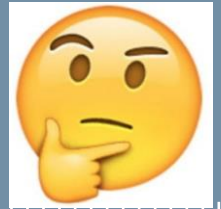
- ✦ The unoccupied time seems longer than occupied time
 - Distract and entertain with small related or unrelated activities
 - Small show at Disneyland, mirrors near elevators, recorded music, magazines at the doctor
- ✦ Pre-process wait seems longer than in-process wait
 - Taking care of the customer as soon as possible to make him/her “in-process” (tracking package with Amazon)
- ✦ Anxiety makes the wait to look longer
 - Communicate and update frequently the waiting time to the customer

- ✦ An unfair wait seems longer than a fair one
- ✦ The more the service is of value or the more the customer senses its utility, the more she/he is willing to wait
- ✦ Waiting alone seems longer than waiting in a group
- ✦ A customer exposed to an exaggerated wait will be a more difficult one to serve or an ex-customer

Therefore, be careful to all these factors and make sure to organize some actions to avoid that the customer senses the supplied service in a negative way!

- ✦ How to model a system (in service/manufacturing company) to understand which are the main criticalities of a company and where they concentrated 
- ✦ How the modelling through queuing theory supports companies, especially service ones, to take better operations decisions (number of resources, bottle-necks, waiting time) 
- ✦ How companies (especially service ones) use queueing theory in order to improve the service level to their customers managing the waiting/throughput time of customers 

Let's Try!– How to shape the following systems?



1. At a post office 5 postal clerks serve customers following this logic: the first free clerk serves the next customer in line (There's space 40 customers at top at the branch)
2. At the BuonaSpesa supermarket there are three cash counters. The firsts two serve customers indistinctly, customers choose autonomously the line. The third cash counter is dedicated to customers with 5 products in the bash at top. The supermarket capacity is not a restriction to customers' possibility of getting in line.
3. From Milan - Linate airport, airplanes take off from the airport only strip
4. Decathlon managers are aware about the importance of the service quality, considering the huge increase of the customers number during Christmas holidays, they would like to find a solution to increase the service rate by changing the queue type.



POLITECNICO
MILANO 1863