QUEUE MANAGEMENT

part 2

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

Agenda of Today

- M/M/c Queueing System
- NETWORKS of QUEUEING SYSTEMS and Nodes Balancing
- Priority rules

Takeaways from the Lecture

- How to model a complex system with queuing theory
- How to model different customer types and a multiple-servers queueing configuration (M/M/c)
- How to model and manage customer’s priority rules in the system

…some numerical example

In the first morning hours customers arrive at the post office where there's a single take a number queue, at an average pace of 54 customers/hour (poisson) while each servers can manage to complete a service in an average time of 4 minutes (negative exponential)

a) If there are 6 servers at the counter: what is the average number of customers in the system, the average waiting time in line and in the system? → **(Ls,Wq,Ws)**

b) Determine the smallest number of counters that need to be opened to keep the average time in the system lower than 10 minutes → **(Ws<10min)**

c) If an employee costs 30€/h and customer's waiting time in line stands for a cost of 20€/h for customer, what would be the optimal number of servers according to purely economic considerations?

Table

…solution

In the first morning hours customers arrive at the post office where there's a single take a number queue, at an average pace of 54 customers/hour (Poisson) while each servers can manage to complete a service in an average time of 4 minutes (negative exponential)

a) If there are 6 servers at the counter: what is the average number of customers in the system, the average waiting time in line and in the system?

Situation can be shaped as a M/M/6 system, where:

λ= 54 customers/hour μ= 1 customer every 4 minutes =15 customers/hour

a) $L_s = L_q + \rho$

$\rho = \lambda/\mu = 54/15=3.6$ c=6 → from table (see next slide for the computation)

L_q=0,295 (average number of customers in line)

$L_s = L_q + \rho$ = 0,295 + 3,6 = 3,895 customers (average number of customers in the system)

$W_q = L_q / \lambda$ = 0,00546 hours → around 20 seconds (19,67 seconds)

$W_s = L_q / \lambda + 1 / \mu$ = 0,072 hours → around 260 seconds (259,67 seconds)

Lq results of model M/M/c

λ/μ	c=1	c=2	c=3	c=4	c=5	c=6	c=7	c=8
0,15	0,026	0,001						
0,20	0,050	0,002						
0,25	0,083	0,004						
0,30	0,129	0,007						
0,35	0,188	0,011						
0,40	0,267	0,017						
0,45	0,368	0,024	0,002					
0,50	0,500	0,033	0,003					
0,55	0,672	0,045	0,004					
0,60	0,900	0,059	0,006					
0,65	1,207	0,077	0,008					
0,70	1,633	0,098	0,011					
0,75	2,250	0,123	0,015					
0,80	3,200	0,152	0,019					
0,85	4,817	0,187	0,024	0,003				
0,90	8,100	0,229	0,030	0,004				
0,95	18,050	0,277	0,037	0,005				
1,0		0,333	0,045	0,007				
1,1		0,477	0,066	0,011				
1,2		0,675	0,094	0,016	0,003			
1,3		0,951	0,130	0,023	0,004			
1,4		1,345	0,177	0,032	0,006			
1,5		1,929	0,237	0,045	0,009			
1,6		2,844	0,313	0,060	0,012			
1,7		4,426	0,409	0,080	0,017			
1,8		7,674	0,532	0,105	0,023			
1,9		17,587	0,688	0,136	0,030	0,007		
2,0			0,889	0,174	0,040	0,009		
2,1			1,149	0,220	0,052	0,012		
2,2			1,491	0,277	0,066	0,016		
2,3			1,951	0,346	0,084	0,021		
2,4			2,589	0,431	0,105	0,027	0,007	
2,5			3,511	0,533	0,130	0,034	0,009	
2,6			4,933	0,658	0,161	0,043	0,011	
2,7			7,354	0,811	0,198	0,053	0,014	
2,8			12,273	1,000	0,241	0,066	0,018	
2,9			27,193	1,234	0,293	0,081	0,023	
3,0				1,528	0,354	0,099	0,028	0,008
3,1				1,902	0,427	0,120	0,035	0,010
3,2				2,386	0,513	0,145	0,043	0,012
3,3				3,027	0,615	0,174	0,052	0,015
3,4				3,906	0,737	0,209	0,063	0,019
3,5				5,165	0,882	0,248	0,076	0,023
3,6				7,090	1,055	0,295	0,091	0,028

To define Lq we need to calculate ρ and define the number of servers (c)

ρ = λ/μ = 54/15=3.6
(Arrival rate/Service rate))

c=6
(Number of Servers)

Lq=0,295
(Average number of customers in line)

…solution

b) Determine the smallest number of counters that need to be opened to keep the average time in the system lower than 10 minutes

Situation can be shaped as a M/M/6 system, where:

λ= 54 customers/hour μ= 1 customer every 4 minutes =15 customers/hour

b) Goal: Ws < 10 min. having the smallest number of open counters

Ws= Lq / λ + 1 / μ → (when c varies. so does Lq!)

Lq / 54 + 1 / 15 <10 min / (60min/hour) →

Lq < 5,4 (with ρ=3,6 it results that c≥5 servers)

…solution

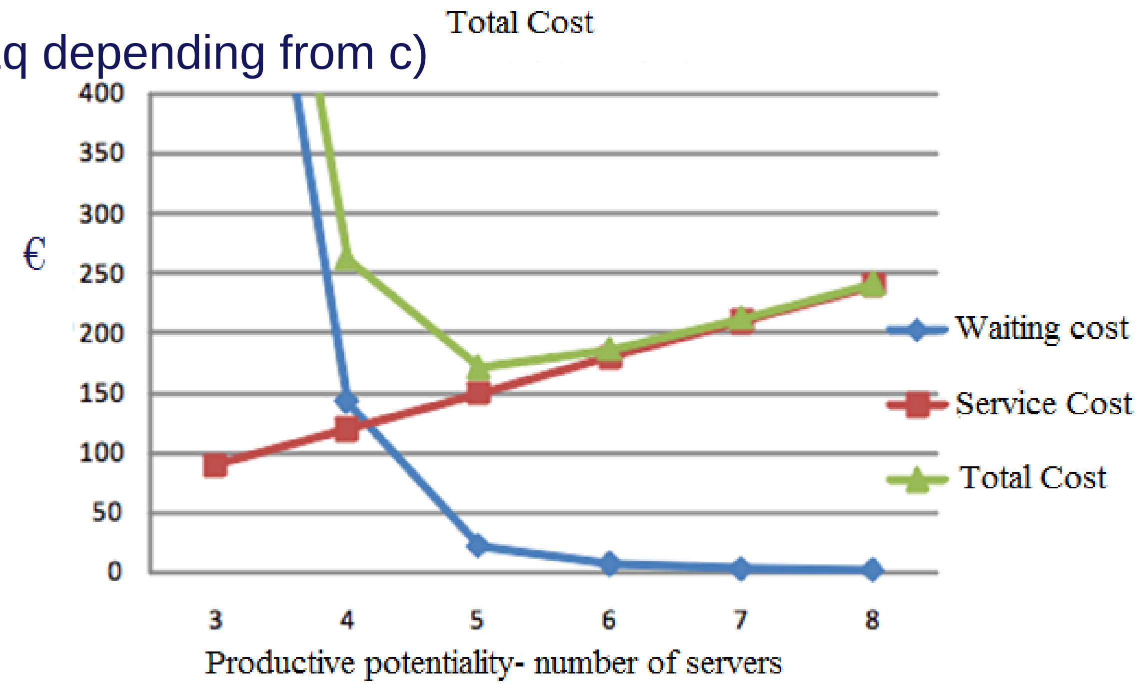
c) If an employee costs 30€/h and customer's waiting time in line stands for a cost of 20€/h for customer, what would be the optimal number of servers according to purely economic considerations?

c) Goal: calculate the optimal number of servers that minimizes service cost and wait cost.

Server cost: 30 €/h * c

Wait cost: 20 €/h * Lq (with Lq depending from c)

goal: min 30 * c + 20 * Lq



…some numerical example

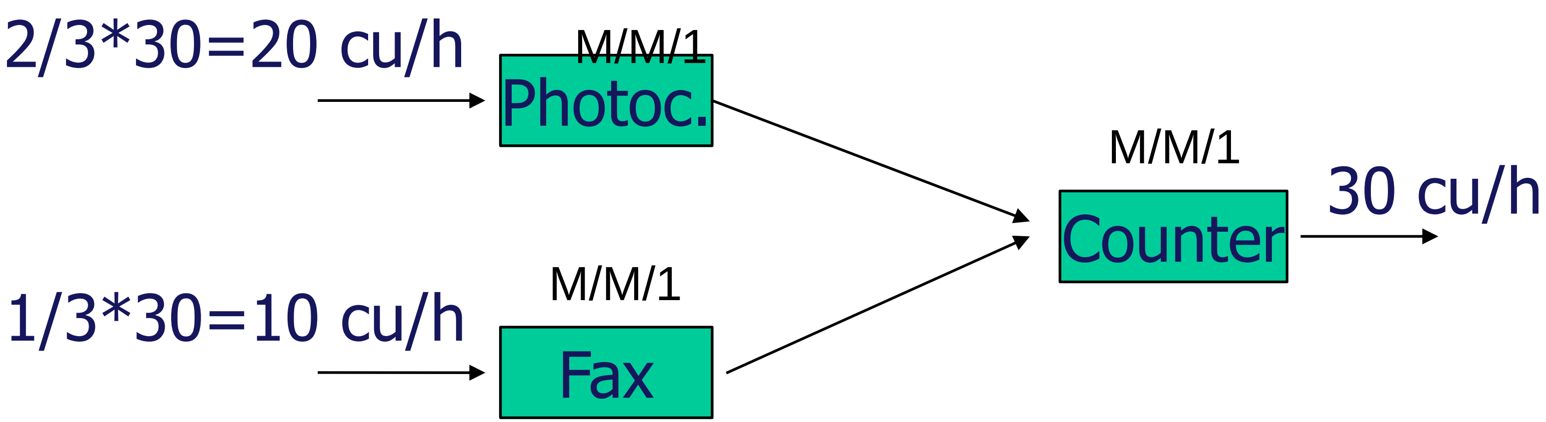
At a photocopier shop, customers usually enter to make copies or send a fax. Once finished, they go to the cash counter to pay. The customers' arrival rate is 30 cu/h. Of those, 2/3 enter to make copies while 1/3 to send fax. The arrival rates are distributed like a Poissonian. On average a customer takes 2 minutes at the photocopier; 4 minutes at the fax; 1 minute to pay.

Supposing that service rates are distributed like a negative exponential:

1. What is the probability that the shop is empty?
2. What is the system average throughput time?
3. How many people on average are in line at the cash counter?
4. How many customers on average are at the shop?
5. What is the probability to have less than 6 people in line at the photocopier?
6. What is the probability to have 1 customer at the shop?

…solution

System shaping:



	Photocopier	fax	Cash counter
λ	30 cu/h * 2/3 = 20 **cu/h**	30 cu/h * 1/3= 10 **cu/h**	20 cu/h + 10 cu/h = 30 **cu/h**
μ	1/2 cu/min * 60 min/h = 30 **cu/h**	1/4 cu/min* 60 min/h = 15 **cu/h**	1 cu/min * 60 min/h = 60 **cu/h**

…solution

1. Probability that the shop is empty

   Probability: $P = P_{0photocopier} * P_{0fax} * P_{0cashcounter}$

$$P_0 = 1 – \rho = 1 – \lambda / \mu$$

Photocopier	1 - 20/30=0.333333
Fax	1 - 10/15=0.333333
Cash counter	1 - 30/60=0.5

Probability: 5,55%

…solution

2. Average system throughput time (Ws)

- Photocopier throughput time Ws = $1/(\mu - \lambda) = 1/(30-20) = 1/10$ hours = 6 minutes
- Fax throughput time Ws = $1/(\mu - \lambda) = 1/(15-10) = 1/5$ hours = 12 minutes
- Cash counter throughput time Ws = $1/(\mu - \lambda) = 1/(60-30) = 1/30$ hours = 2 minutes

System expected throughput time: weighted sum of all types of customers' throughput times (There are two types of customers)

photocopier customers → cash counter 20 / 30 customers

fax customers → cash counter 10 / 30 customers

20/30 * (6 min + 2 min) + 10/30 * (12 min + 2 min) = 10 minutes

…solution

3. Average people in line at the cash counter:

$Lq = \rho * \lambda / (\mu - \lambda) = (30/60 * 30) / (60-30) = 0{,}5$ customers

4. Average number of customers at the shop

Ls shop = Ls photo + Ls fax + Ls cash counter

$Ls = \lambda / (\mu - \lambda)$

	Photocopier	fax	cashcounter	total
Ls	20/(30-20)= **2**	10/(15-10)= **2**	30/(60-30)= **1**	2+2+1= **5**

…solution

5. Probability to have less than 6 customers in line at the photocopier

Probability to have less than k customers <u>in the system</u>

(IN THE SYSTEM = LINE + IN SERVICE)

$P(n<k) = 1 – P(n≥k) = 1- \rho^{k}$

Probability to have less than 6 customers in line

$P(n<7) = 1 – P(n ≥7) = 1- \rho^{7}$

[6 customers in line + 1 in service]

$P(n<7) = 1- \rho^{7} = 1 – (20/30)^{7}= 94.15\%$

…solution

6. What is the probability to have 1 customer at the shop?

$P = P_{1\ photo} * P_{0\ fax} * P_{0\ cashcounter} + P_{0\ photo} * P_{1\ fax} * P_{0\ cashcounter} + P_{0\ photo} * P_{0\ fax} * P_{1\ cashcounter}$

$P_1 = P_0 * \rho^n$

$P_0 = 1 – \rho = 1 – \lambda / \mu$

	Photocopier	fax	cashcounter
P_1	0.2222222	0.2222222	0,25
P_0	0.333333	0.333333	0.5

Probability: 10,18%

… some messages from queue system shaping

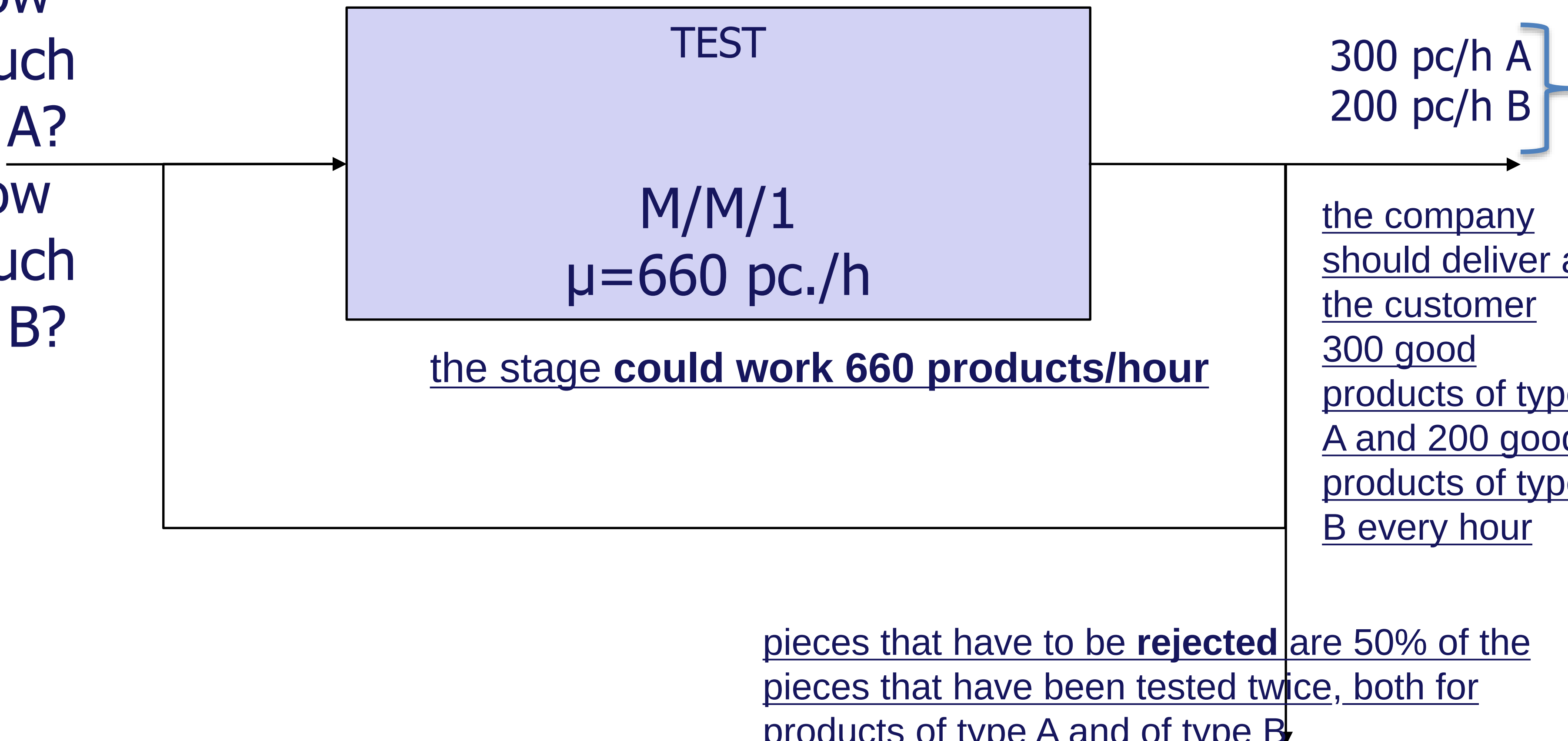
- A complex system is the composition of more elemental sub-systems that interact among themselves (mapping the system and identifying every step of the process).
- Very often in a complex system different types of customers exist (we must know in advance every type of customers in the system).
- The system expected throughput time is a representation of an **"average"** customer and not of the specific one!
- The throughput time varies according to the variation of the workload in the systems and at the variation of the input.

Nodes balancing to flows calculation

- (From the Operations Management Exam of September 25, 2008)

The test phase of a stage is a stage where enter, without different priority logics, product A and B that form the productive selection of the company. In particular, the testing machine isn't of the latest generation. Indeed, every once in a while it could signal as nonstandard even products that actually are good. For this reason, in case <u>nonstandard products</u> are signaled, they <u>are tested once more</u>. <u>On average 15% of products A and 15% of products B that the machine tests have to be tested again</u>. The <u>pieces that have to be rejected are 50% of the pieces that have been tested twice, both for products of type A and of type B</u>. The rest of the products instead are good and can be delivered to the customer. In case <u>the stage could work 660 products/hour</u> and in case <u>the company should deliver at the customer 300 good products of type A and 200 good products of type B every hour</u>, <u>how much material is it necessary</u> to be put in the test phase every hour? Shape the system reporting all the significant parameters.

How much of A? How much of B?



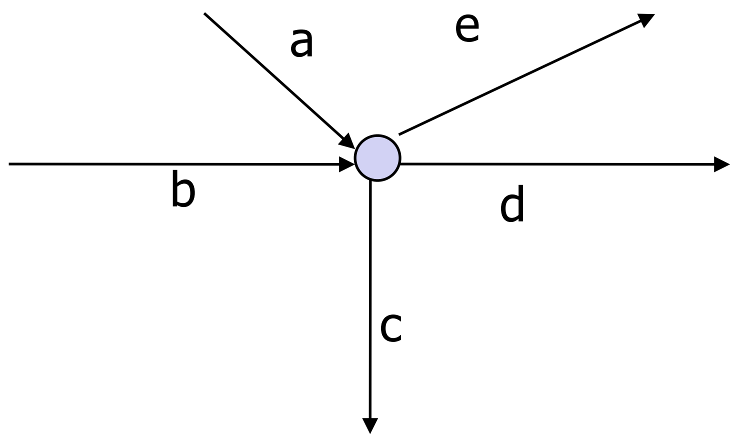
the stage **could work 660 products/hour**

the company should deliver at the customer 300 good products of type A and 200 good products of type B every hour

pieces that have to be **rejected** are 50% of the pieces that have been tested twice, both for products of type A and of type B

Everything that enters in the node is equal to everything that comes out from it.

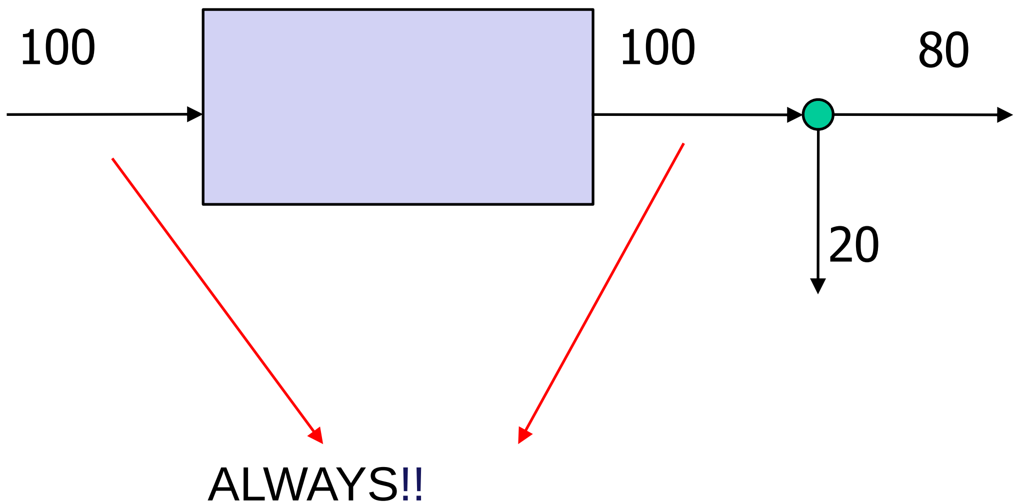
Just like Energy, Nothing is created or destroyed!!



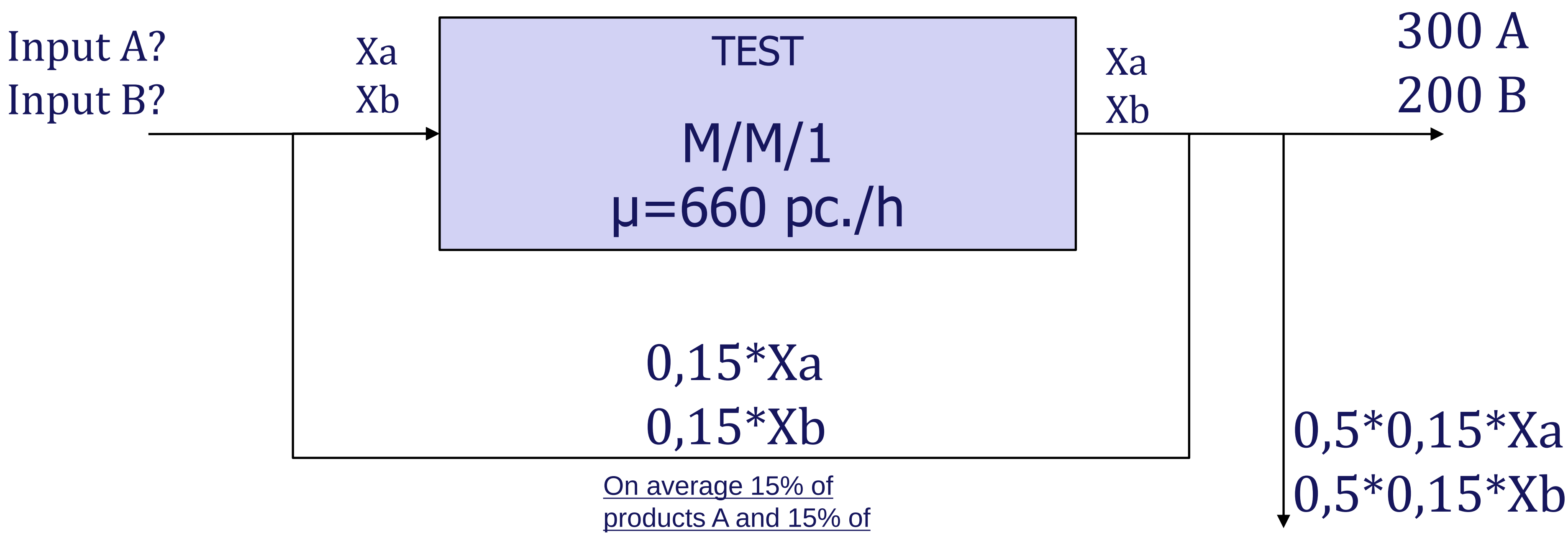
$a+b = c+d+e$ ALWAYS!!!!

The flow in the system (line + server)

Everything that enters in the system is equal to everything that comes out. In the system there isn't any flow loss. (the service process works on customers. With neither creating nor destroying any!)



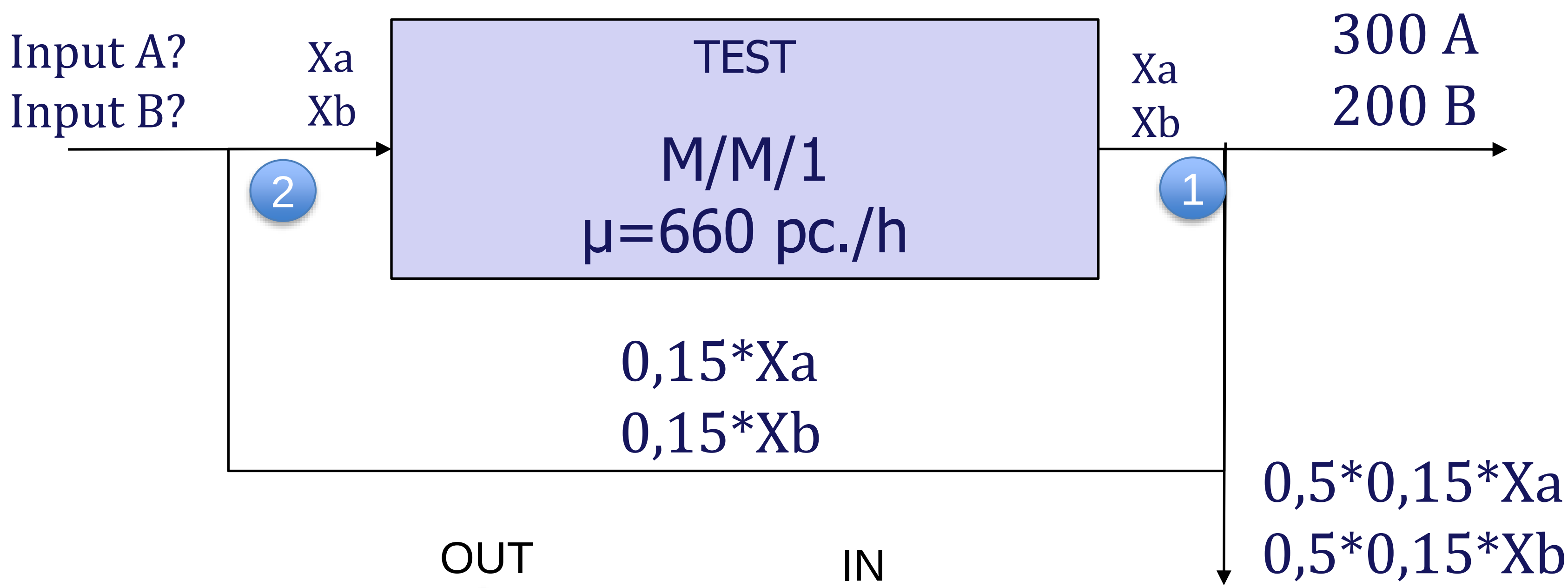
The flow in this system



On average 15% of products A and 15% of products B that the machine tests have to be tested again

pieces that have to be **rejected** are 50% of the pieces that have been tested twice, both for products of type A and of type B

Nodes Balancing

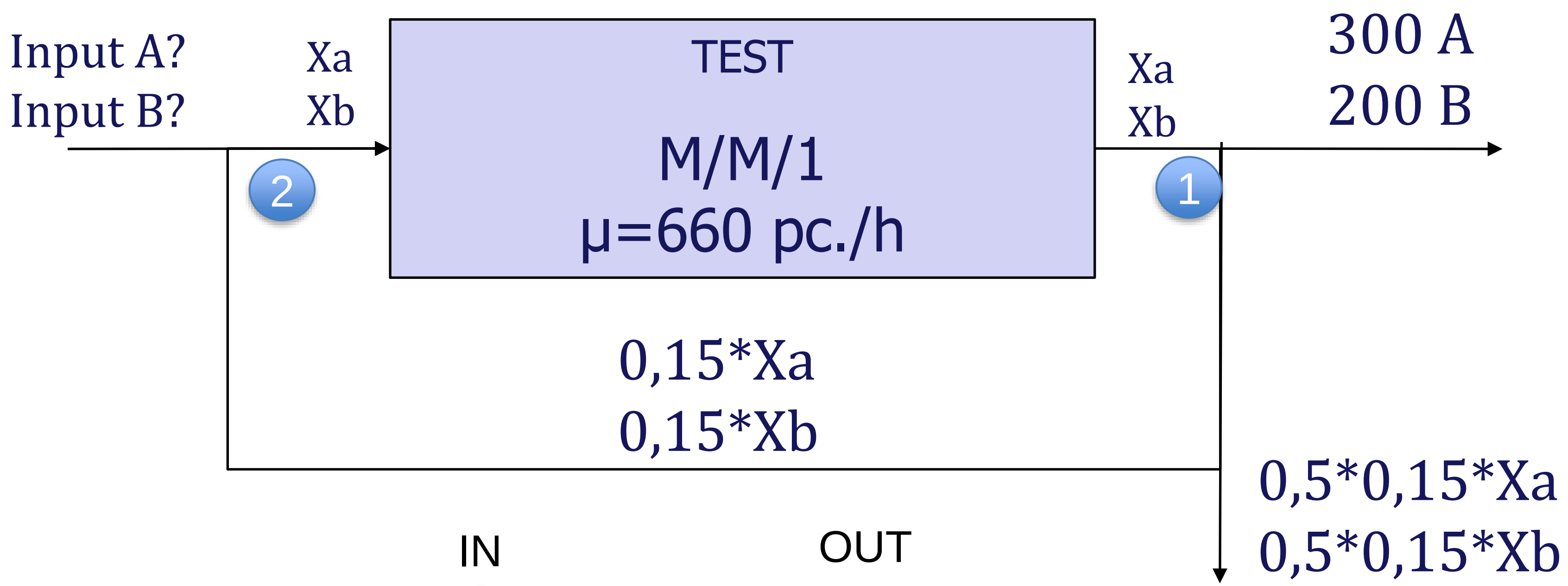


1

A) $\overbrace{300+0{,}15Xa+0{,}5*0{,}15Xa}^{\text{OUT}}=\overbrace{Xa}^{\text{IN}} \rightarrow Xa=387{,}09\ pc/h$

B) $200+0{,}15Xb+0{,}5*0{,}15Xb=Xb \rightarrow Xb=258{,}06\ pc/h$

Nodes Balancing

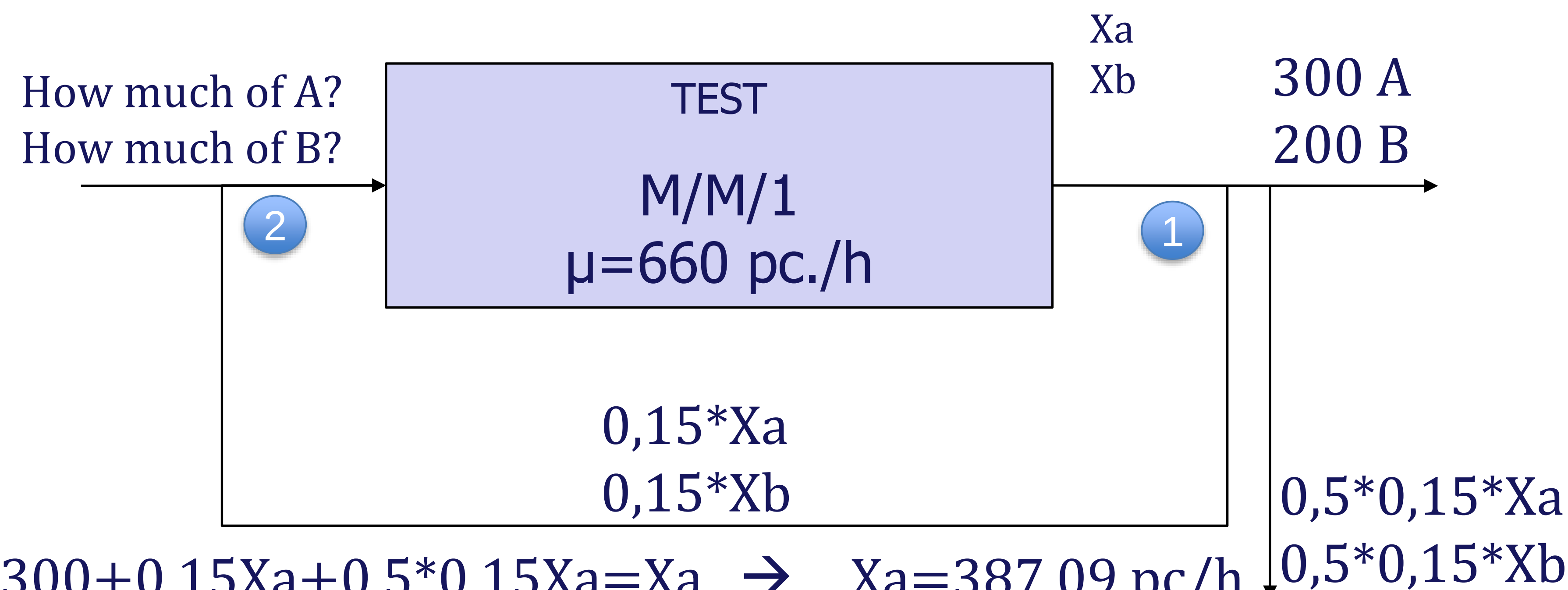


2

$$\underbrace{\text{Input A: input A} + 0{,}15Xa}_{\text{IN}} = \underbrace{Xa}_{\text{OUT}} \rightarrow \text{input A} = 329{,}02 \text{ pc/h}$$

$$\text{Input B: input B} + 0{,}15Xb = Xb \rightarrow \text{input B} = 219{,}35 \text{ pc/h}$$

…to sum up



1. 300+0,15Xa+0,5*0,15Xa=Xa → Xa=387,09 pc/h
   200+0,15Xb+0,5*0,15Xb=Xb → Xb=258,06 pc/h

2. Input A: input A + 0,15Xa=Xa → input A = 329,02 pc/h
   Input B: input B + 0,15Xb=Xb → input B = 219,35 pc/h

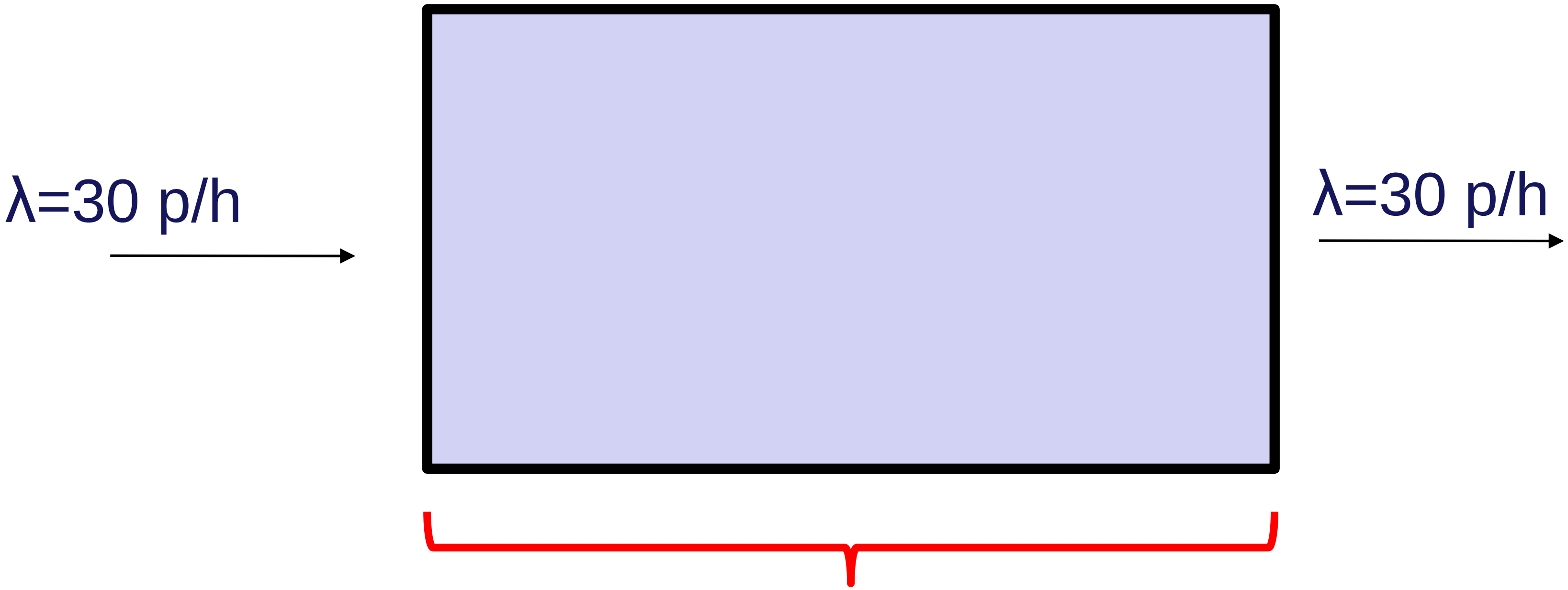
EX 1
RINASCIMENTO CLINIC

"Rinascimento" clinic is a small health care facility, located in Austin, Texas.
The medical examinations take place without appointment and the flow is described as follow: a patient enters in the clinic and goes to reception, where the receptionist, according to the patient problem, sends her to Dr Carter ,a general practitioner, or sends her to Dr Romano, a specialised physician. In some cases the patient could be seen by both Dr Carter and Dr Romano.
The collected data are depicted in the following table. Arrival rate and service rate are distributed as a negative exponential.

Variable	Average
Patients arrival rate	30 p/h
Patients that are sent to Dr Carter (fraction of the total amount entering in the system)	2/3
Patients that are seen by Dr Carter and then by Dr Romano (percentage of the number of patient that are sent to Dr Carter)	0.15
Receptionist service rate	40 p/h
Dr Carter service rate	30 p/h
Dr Romano service rate	15 p/h

1. The priority rule is FCFS. You are required to calculate:
   a. Average waiting time in each queue.
   b. Average waiting time in the system for each patient typology.
   c. The expected LT (of a generic patient) .
   d. Average time of inactivity (expressed in minutes each hour) for the receptionist, Dr Carter and Dr Romano.
2. Rinasciemento clinic is considering implementing a priority system to manage in a different way the visits held by Dr Romano. The idea is that the patients going to Dr Romano after being seen by Dr Carter, have a non-pre-emptive priority on the patients that arrive directly to Dr Romano. Which are, in your point of view, the main effects introducing this priority rule has on the LT (for the 3 typologies and for the generic patient) ?

EX 1
RINASCIMENTO CLINIC



Average lead time: around 25 minutes

The system average throughput time is around 25 minutes. But the patients complain of a very variable time and that the reason behind it it's not clear. Time varies from less than 5 minutes to more than 45!

Customers' complains risk damaging the center image and losing customers!

↓

A specific analysis of the process is necessary to comprehend the causes of variability

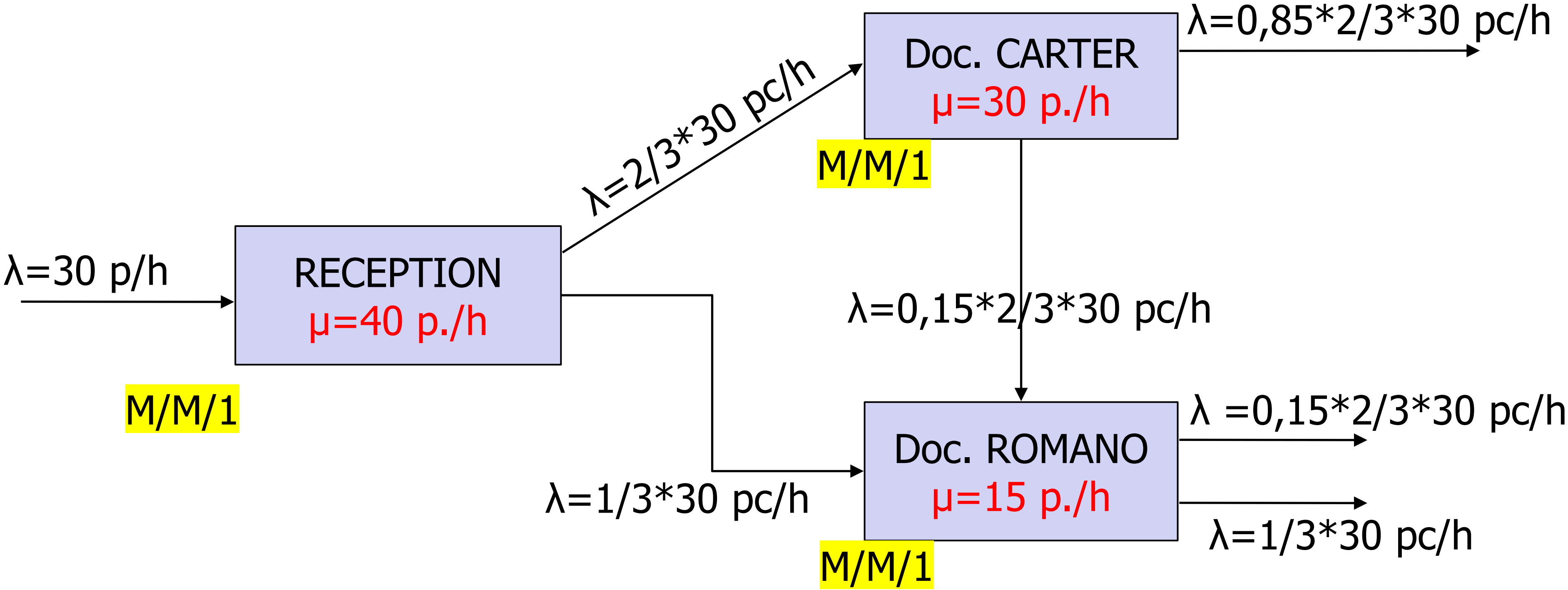
- Each process is characterized by a certain patients' arrival rate and a certain service capacity. Both parameters influence queue length and throughput times!

  → Know the system

- The system is formed by different entities (different type of patients) who have different behaviors and paths

  → Identify different typology/families

RINASCIMENTO CLINIC



	Reception	Doctor Carter	Doctor Romano
λ	30 p/h	20 p/h	10+3 =13 p/h
μ	40 p/h	30 p/h	15 p/h

Wq, Ws, Inactivity time

Wq= $(\lambda/\ \mu)/(\mu-\ \lambda)$	Ws=$1/(\mu-\ \lambda)$	inact.t.=$(1-\ \rho)$*60 min/h

	Reception	Doc. Carter	Doc. Romano
Wq	(30/40)/(40-30)*60= **4,5 min**	(20/30)/(30-20)*60= **4 min**	(13/15)/(15-13)*60= **26 min**
Ws	1/(40-30)*60= **6 min**	1/(30-20)*60= **6 min**	1/(15-13)*60 = **30 min**
Inactivity time	(1-(30/40))*60= **15 min**	(1-(20/30))*60= **20 min**	(1-(13/15))*60= **8 min**

- Ws is given by the sum of Wq and service time (do the math!)
- Why does it simultaneously exist an average waiting time and a server's inactivity time?

Average waiting time in line for each one of the three typologies

We have 3 different types of customers / types of path in the system:

1) Reception → Doctor Carter
2) Reception → Doctor Carter → Doctor Romano
3) Reception → Doctor Romano

	Reception (Wq)	Doc. Carter (Wq)	Doc. Romano (Wq)	Wq
Type 1	4,5 min	4 min	-	4,5+ 4 = 8,5 min.
Type 2	4,5 min	4 min	26 min	4,5+4+26 = 34,5 min.
Type 3	4,5 min	-	26 min	4,5 + 26 = 30,5 min.

Occurences estimation

3 different types of customers / route in the system:

1) Reception → Doctor Carter
2) Reception → Doctor Carter → Doctor Romano
3) Reception → Doctor Romano

	Occurrences
Type 1	**((2/3)*30) - (0,15*(2/3)*30))/ 30 = 0.5666**
Type 2	**(0,15*(2/3)*30) / 30 = 0.1**
Type 3	**((1/3)*30) / 30 = 0.3334**

Expected throughput time

3 different types of customers / route in the system:

1) Reception → Doctor Carter
2) Reception → Doctor Carter → Doctor Romano
3) Reception → Doctor Romano

	Reception (Ws)	Carter (Ws)	Romano (Ws)	Ws	Occurence	Expected time
Type 1	6 min.	6 min.	-	**12 min.**	**0.5666**	12*0,5666 +42*0,1+36*0,3334=
Type 2	6 min.	6 min.	30 min.	**42 min.**	**0.1**	
Type 3	6 min.	-	30 min.	**36 min.**	**0.3334**	
EXPECTED THROUGHPUT TIME					**1**	**23 min.**

expected throughput time→ weighted sum of the average throughput times for each type of product

Expected throughput time, messages

- The system expected throughput time refers to an "average product". It doesn't have to be taken as an absolute benchmark
- It's crucial to know the system and to map it: which phases? which products? which problems?
- Each type of product has its variability
- If it's possible to assign the type of product at the beginning of the process, the uncertainty decreases and it's possible to better align performances to expectations, therefore it's possible to start effective countermeasures

M/M/1 system with assigned priority to a customers' class

It's a system formed by just one server who is able to serve two classes of customers (1 and 2) that have a different service admission

<u>Hp:</u> customers of class 1 have a service admission priority higher than customers of class 2

The following cases will be taken into account:

- Preemptive priority
- Non preemptive priority

M/M/1 system with assigned priority to a customers' class

- Preemptive priority

If a customer of class 1 enters the system while a customer of class 2 is being served, the service to the customer of class 2 is interrupted in order to give immediate precedence to the customer of class 1

Following formulas allow to calculate the system throughput time of the two classes of customers

$$E(S_1) = \frac{1/\mu}{1-\rho_1},$$

$$E(S_2) = \frac{1/\mu}{(1-\rho_1)(1-\rho_1-\rho_2)}..$$

M/M/1 system with assigned priority to a customers' class

- Preemptive priority

Example: λ_1= arrival rate of type 1 customers= 0,2

λ_2= arrival rate of type 2 customers = 0,6

μ = service rate = 1

Without priority

$$E(S) = \frac{1}{1-0.8} = 5$$

In case that customer 1 has absolute priority

$$E(S_1) = \frac{1}{1-0.2} = 1.25, \qquad E(S_2) = \frac{1}{(1-0.2)(1-0.8)} = 6.25.$$

M/M/1 system with assigned priority to a customers' class

- Non Preemptive priority

If a customer of class 1 enters the system while a customer of class 2 is being served, the service to the customer of class 2 is completed and then the customer of class 1 is taken care of

Following formulas allow to calculate the system throughput time of the two classes of customers

$$E(S_1) = \frac{(1+\rho_2)/\mu}{1-\rho_1},$$

$$E(S_2) = \frac{(1-\rho_1(1-\rho_1-\rho_2))/\mu}{(1-\rho_1)(1-\rho_1-\rho_2)}.$$

M/M/1 system with assigned priority to a customers' class

- Non Preemptive priority

Example: λ_1= arrival rate of type 1 customers= 0,2

λ_2= arrival rate of type 2 customers = 0,6

μ = service rate = 1

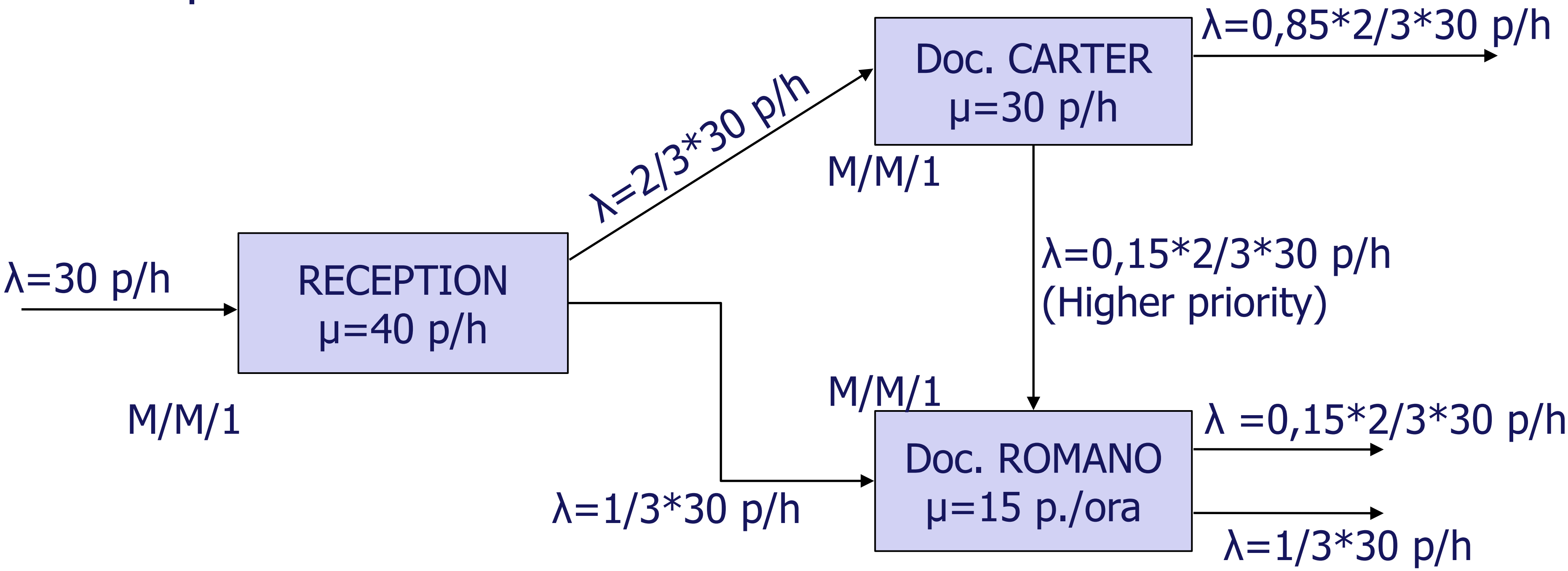
In case that customer 1 doesn't have absolute priority on customer 2

$$E(S_1) = \frac{1+0.6}{1-0.2} = 2, \qquad E(S_2) = \frac{1-0.2(1-0.8)}{(1-0.2)(1-0.8)} = 6.$$

System with priority – Rinascimento clinic

Class with higher priority: customers arriving from doc. Carter

Class with lower priority: customers arriving from the reception



Class with high priority: customers arriving from doc. Carter (S1)

Class with low priority: customers arriving from the reception (S2)

$\rho 1 = \lambda 1 / \mu = 3/15 = 0{,}2$

$\rho 2 = \lambda 2 / \mu = 10/15 = 0{,}667$

$E(S1) = ((1 + \rho 2)/ \mu)/(1- \rho 1) = ((1{,}667)/15) / (1 – 0{,}2) = 0{,}1389$ h $= 8{,}333$ minutes

$E(S2) = (((1 - \rho 1*(1- \rho 1- \rho 2))/ \mu)/((1- \rho 1)*(1- \rho 1- \rho 2)) = 0{,}6083$ h $= 36{,}5$ minutes

	Reception (Ws)	Carter (Ws)	Romano (Ws)	Ws	OCCURRENCE	EXPECTED TIME
Type 1	6 min.	6 min.	-	**12 min.**	0.5666	
Type 2	6 min.	6 min.	8,33 min. (priority 1)	**20,33 min.**	0.1	
Type 3	6 min.	-	36,5 min. (priority 2)	**42,5 min.**	0.3334	
EXPECTED THROUGHPUT TIME					**1**	**23 min.**

Priorities…

The alteration of the priority logic doesn't change the system expected throughput time…

…However, in some cases it's essential to increase the customer's satisfaction degree and to improve customer service

Some examples – How to shape the following systems?

1. At a post office, the chief observed that people renege usually after on average 15 minutes waiting in the system. How the chief can manage this task?

2. At the Linate Airport it has been seen the necessity to classify customers according to their needs. For this reason they have introduced tickets with priority. Evaluate how the single customer throughput time could change by hypothesising the data, considering as base case, the case in which none has a specific priority.

Take Aways from the Lecture

- How to model a complex system with queuing theory ✓
- How to model different customer types and a multiple servers one queue configuration (M/M/c) ✓
- How to model and manage customer's priority rules in the system ✓

Take aways

1. M/M/C
- specific way to calculate performance
- More efficient than c M/M/1

1. Network
- System performance is the combination of the single system performance
- Node balancing: inputs must be always equal to the sum of the outputs

1. Priority rules
- can customize specific service for specific product
- Average performance remains the same

POLITECNICO
MILANO 1863