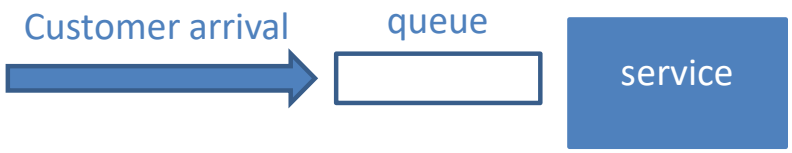
Queuing theory

Queuing theory highlights

- In the world of innovation, the word “flexibility” evoked everybody the meaning of chaoticity that only a computer could have faced and the queuing theory seemed the best tool in the hand of researched to handle the issue.
- The Queuing Theory studies the phenomena that can be summarized in the following model:



- There are three main fields of study:
  - Open stationary systems made of infinite customers with an arrival rate minor than service rate (otherwise the queue becomes infinite)
  - Closed stationary systems consisting of a finite numbers of customers
  - Dynamic systems

- Manufacturing systems are generally of the second type.

The Little's law

In mathematical Queuing theory Little's law is a theorem by John Little which states that the average number W of customers in a stationary system is equal to the average effective arrival rate P multiplied by the average time T that a customer spends in the system. Expressed algebraically the law is:

$$\mathbf{P} = \frac{\mathbf{W}}{\boldsymbol{T}}$$

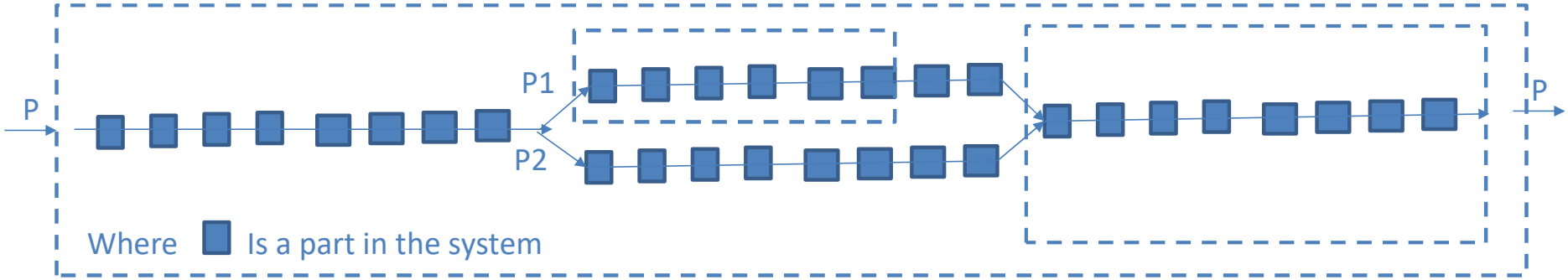
The relationship is not influenced by the arrival process distribution, the service distribution, the service order, or practically anything else. In most queuing systems, service time is the bottleneck that creates the queue.

The result applies to any system, and particularly, it applies to systems within systems.

In Manufacturing P is the throughput, production per time unit. W the Work in progress, the number of parts in the systems, and T the time to go through the system.

Perimeter of application of the Little's law

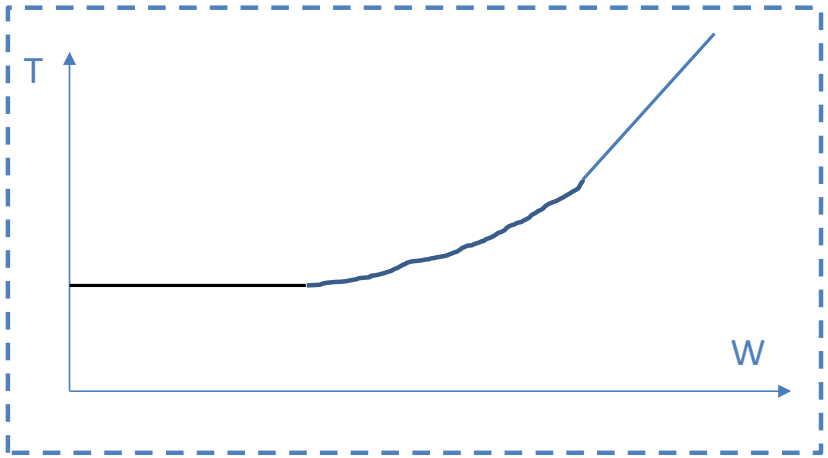
- The law as said can be applied to every manufacturing system I a stationary state where the customers are represented by the part in production and the service by the machines.
- Arrival rate and output rate in a stationary state is the same of course.
- It is valid for any perimeter of the production system we want to analyze:



- In the larger rectangle P= 32/TT
- In the smallest P1=6/T1
- In the medium P=6/TM

The function T(W)

- It has been said that the Little's law is valid for average values and stationary systems but it is really interesting to analyze how the system works with different value of W.
- This is really the main value added of this study to the Manufacturing practice since it is used to dimension the length of the transport, the buffer, the number of pallet or fixture and in some cases to assess the max throughput,
- Let start to study T(W)
  - When the population of customers is small, the Time inside the system is constant and is the sum of all the technical time to transfer and the cycle times.
  - When the population is high, everything is packed and the time in the system is linearly growing with the increase of the customers
  - In the middle there is an evolutionary behavior that can be studied with the Queuing theory.

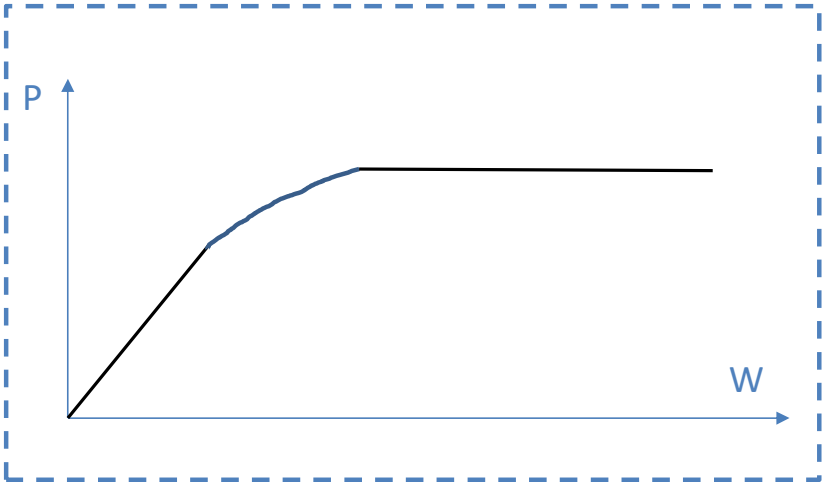


The function P(W)

- Having defined the function T(W) it is immediate define algebraicly the function P(W) that is:

$$\mathbf{P}(\boldsymbol{W}) = \frac{\mathbf{W}}{\boldsymbol{T}(\boldsymbol{W})}$$

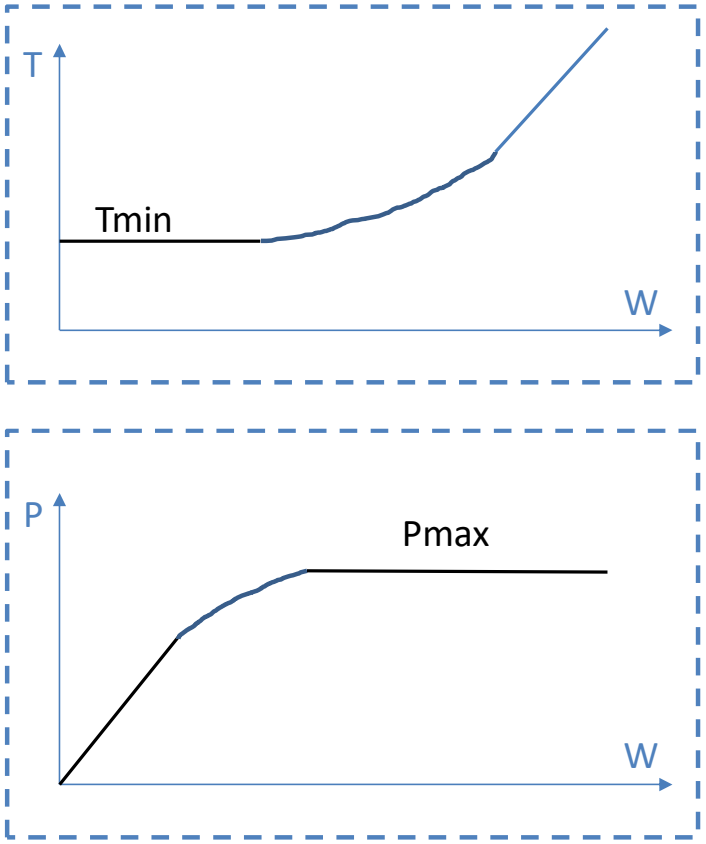
- When the population of customers is small, the throughput grows linearly
- When the population is high, everything is packed and the throughput is equal to the bottleneck rate
- In the middle there is an evolutionary behavior that can be studied with the Queuing theory.
- The statistical theory says something different from the practice and the P(W) curve is asymptotic to the bottleneck rate



Formulas for average numbers

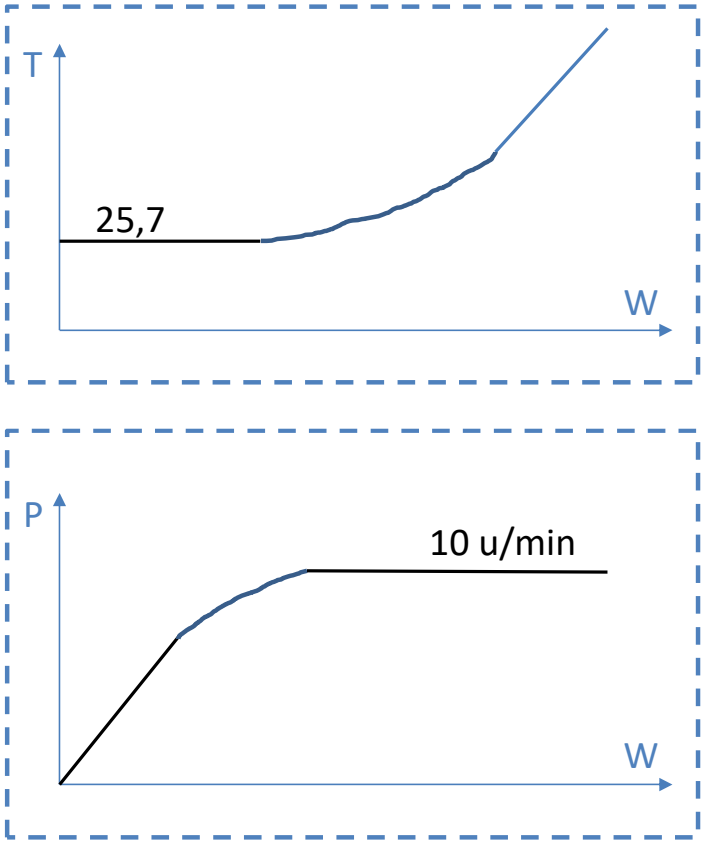
- With a small number of customers, no queue:
  - T= Tmin= time for a turn without queue
  - P=W/Tmin<Pmax

- With a high number of customer and queue:
  - P=Pmax
  - T=Tmin+TQ=W/Pmax
  - TQ=(W/Pmax)-Tmin



Practical example: the isolated skilift

- Let consider the case of a isolated skilift with following features:
  - Flow Pmax 1u/6s= 10 u/min=600 u/h
  - Distance between 2 hooks=6 m
  - Speed 1 m/s=60 m/min=3,6 km/h
  - Length 1 km
- The slope has a length of 3 km and the average speed of the customers is 20km/h= 333 m/min=5,6 m/s
- Min time in the system= 1000/60+3000/333=16,7+9= 25,7 min
- Min number of customers to start the que= Wcr=Tmin*Pmax= 257 u
- Time in the que for 500 customers= (W/Pmax)-Tmin= (500/10)-25,7=24,3 min



The statistical theory: open system/single server

- The first system studied by the theory is the open system with single server (aka M/M/1).
- The system is characterized by:
  - Infinite customers with arrival rate **λ** u/min according to Poisson statistical law
  - Single service with service rate **μ** u/min according to Poisson statistical law
- **Immediately we note that it must be λ<μ.**
- **Formulas:**
  - Probability that there is no queue= 1- λ /μ
  - Average time that a customer spend in the system T=1/(μ- λ)
  - Average time that the customer wait in the queue TQ= λ /μ (μ- λ)
  - Average length of the queue (number of customer waiting) WQ= λ *TQ= λ **2/μ (μ- λ)

- **Immediately we note that it is difficult to apply to the skilift model as to the pallettized line.**

The statistical theory: open system/multiple server

- The second system studied by the theory is the open system with multiple server (aka M/M/S).
- The system is characterized by:
  - Infinite customers with arrival rate **λ** u/min according to Poisson statistical law
  - Multiple S servers with service rate **μ** u/min each according to Poisson statistical law
- **Immediately we note that it must be λ<S*μ.**
- **Formulas:**
  - Probability that there is no queue $$P(0 = \frac{1}{\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \cdot\frac{(\lambda/\mu)^s}{s!}\left(\frac{1}{1-\lambda/s\mu}\right)}$$
  - Average length of the queue (number of customer waiting) $$WQ = \frac{P_0\lambda^{s+1}}{(s-1)!\mu^{s-1}(s\mu-\lambda)^2}$$
  - Average time that the customer wait in the queue $TQ = WQ/\lambda$
  - Average time that a customer spend in the system $T = TQ + 1/\mu$

The statistical theory: limited customers

- A little more interesting is the theory of closed system where the number of customer is limited and the arrival rate is reduced by the effect of the queue and the services (aka M/M/S/N).
- The assumption is anyway to have Poisson distribution for arrival and services, that is not so realistic indeed but it could be used for maintenance service and other similar analysis.
- This system is characterized by:
  - N= customer in the system
  - S =number of servers
  - λ =arrival rate of each customer
  - μ service rate of each server

- To evaluate the value of T, TQ, W, WQ is important to define the probability of each state that could be done with an excel template.

Formulas for the limited customers' model

In this model the process require to define:

1. **Arrival rate lambda λ**
2. **Service rate mu μ**
3. **Server number S**
4. **Customer number N**
5. **Adapt raw numbers and formula for N and S**
6. Evaluate the incremental ratio of probability p(n+1)
7. Evaluate the ratio vs p(0)
8. Define p(0)
9. Define all p(n)

Note that the steps from 6 to 9 are automatically done by the file

lambda	mu	S	N	p(n+1)/p(n)=			
		2	2	n=0	n<S	n>=S	
customer n	server busy	free servers	p(n)	p(1)/p(0)= N*lambda/mu	= (N-n)* lambda/ (n+1)*mu	= (N-n)* lambda/ S*mu	p(n)/p(0)
0	0	S					
1	1	1					
2	S	0					
3	S	0					
4	S	0					
N	S	0					
…	…	…					…
N	N-S						
							0,00

Finally, can be calculated automatically:

- Probability that there is no queue p(0)
- Average length of the queue (number of customer waiting) WQ
- Effective rate (due to limited customers' effect) λeff
- Average time that the customer wait in the queue TQ= WQ/λeff
- Average time that a customer spend in the system T= W/λeff

Exercise 1 - Maintenance

- WE have 5 machines that has breakdown with the rate of 1/hr. Two repair men that takes 15 min to repair so 4/hr. The model is:

lambda	**mu**	**S**	**N**	p(n+1)/p(n)=				
1	4	2	5,00	n=0	n<S	n>=S		
customer n	server busy	free servers	p(n)	p(1)/p(0)= N*lambda/mu	= (N-n)* lambda/ (n+1)*mu	= (N-n)* lambda/ S*mu	p(n)/p(0)	
0	0	2	0,315	1,25				-0,630
1	1	1	0,394		0,5		1,25	-0,394
2	2	0	0,197			0,375	0,63	0,000
3	2	0	0,074			0,25	0,23	0,074
4	2	0	0,018			0,125	0,06	0,037
5	2	0	0,002			0	0,01	0,007
…	…	…					…	
N	N-S							
							2,18	0,118
Probability that there is no queue				p(0)=	0,315			
Average length of the queue (n.of customer waiting)				WQ=	0,118			
Average n.of customer in the system				W=	1,094			
Effective rate (due to limited customers' effect)				λeff=	3,906	3,906		
Average time that the customer wait in the queue				TQ= WQ/λeff=	0,030			
Average time that a customer spend in the system				T= W/λeff=	0,280			

Exercise 2= Battery charger

- We have #10 AGV that has down battery with the rate of 4/g (charge duration= 6 h).
- #5 charger to repair each with 8/g (recharge time=3hr).
- The number in queue is very little 0,15
- We could propose a reduction

lambda	mu	S	N	p(n+1)/p(n)=				
4	8	5	10,00	n=0	n<S	n>=S		
customer n	server busy	free servers	p(n)	p(1)/p(0)= N*lambda/mu	= (N-n)* lambda/ (n+1)*mu	= (N-n)* lambda/ S*mu	p(n)/p(0)	
0	0	5	0,0169	5				
1	1	4	0,084		2,25		5,00	-0,337
2	2	3	0,190		1,33		11,25	-0,569
3	3	2	0,253		0,88		15,00	-0,506
4	4	1	0,221		0,60		13,13	-0,221
5	5	0	0,133			0,5	7,88	0,000
6	1	0	0,066			0,4	3,94	0,066
7	2	0	0,027			0,3	1,58	0,053
8	3	0	0,008			0,2	0,47	0,024
9	4	0	0,002			0,1	0,09	0,006
10	5	0	0,000			0,0	0,01	0,001
…	…	…					…	
N	N-S							
							58,34	0,150
Probability that there is no queue				p(0)=	0,017			
Average length of the queue (n.of customer waiting)				WQ=	0,150			
Average n.of customer in the system				W=	3,518			
Effective rate (due to limited customers' effect)				λeff=	26,939	25,928		
Average time that the customer wait in the queue				TQ= WQ/λeff=	0,006			
Average time that a customer spend in the system				T= W/λeff=	0,131			

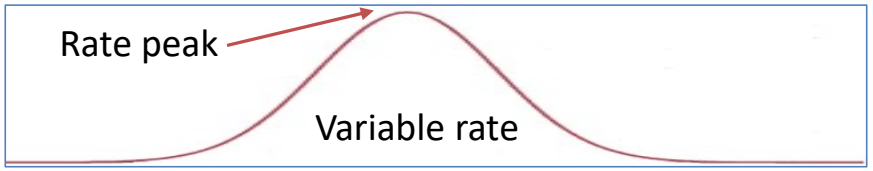
Practice vs theory

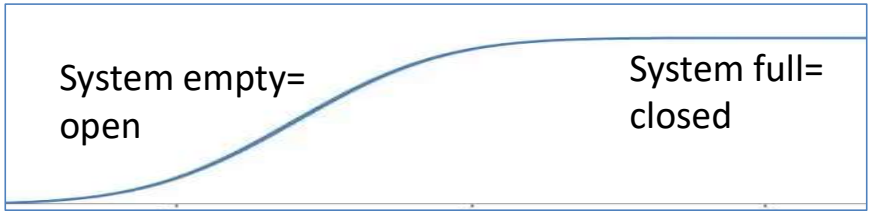
- About the queueing theory is much more important to know the concept and the methodology than to apply it in practical cases.
- The main problem is that having a stationary system and infinite source and exponential arrival is a too unrealistic condition. The reality is more adaptative in case of arrival rate and service rate since there is opportunities to simplify or accelerate and the rate is not constant.
- The M/M/1 system can be used just for services with high number of customers
- The M/M/S is mostly not used being shorter, inside the high approximation, to multiply the server rate by the number of servers and consider to be in M/M/1 system
- The M/M/S/N instead has a large range of application in the Building & utilities design.
- For the closed systems (e.g. a palletized assembly line) none of these Open System approaches can work.
- There are iterative algorithms that can reach a good level of approximation but can be used for specific cases only.
- The best is to use simulation with a better definition of the statistics.

The limit of the theory: nonstationary systems

Most of the real queuing systems are not stationary e.g. the rush hour traffic

1. There is a moment in which the system is effectively **open** but the rate is not constant.
2. Then, a period in which the System has few new arrival and is behaving as a **closed** system
3. Finally, sometime a second rush time as for the step 1





Backup

Simulation

- It has been said that the efficiency is a data given by experience. It is true but in the last 40 years a lot of simulation tools have been developed in order to identify with growing precision the efficiency of a complex system.
- If the operations are independent (like infinite warehouse in the middle) the total efficiency is the one of the bottleneck. But in real conditions there are interferences and the result is lower than the bottleneck.

A breakdown of the previous station or the followers stops the flow (shortage of feeding or queue) – **only simulation can identify**

Some time material not loaded or wrong - other random distribution

Efficiency of the single operation defined by the machine builder.
Distribution of defect exponential with possibly multiple root causes (scrap, jamming, breakdown, safety emergency)