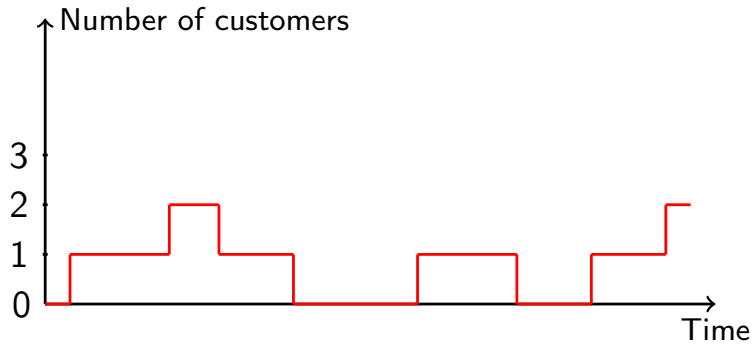
What is a queue?

Modeling the problem

Queuing systems problems

Mathematical modeling

- **the evolution over time of the number of customers** in the system (waiting or being served)
- is modeled by a **random process**
- of which we study the **steady-state regime**, in order to estimate **mean performance**

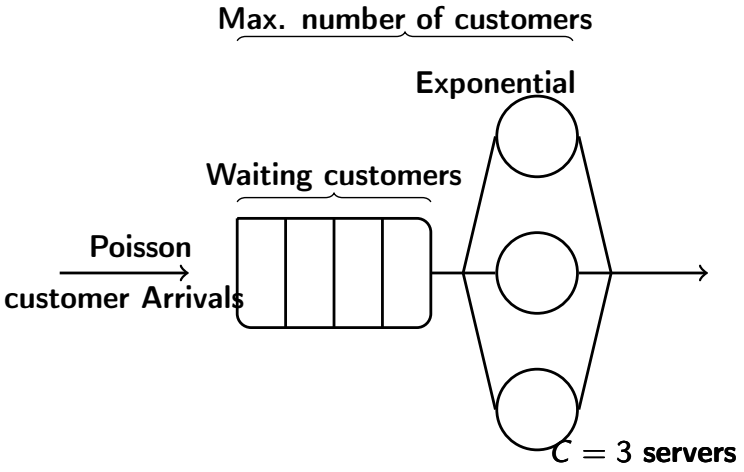


Queuing theory is mainly mathematics. The standard approach is to model the evolution over time of the number of customers in the system as a random process. The mean performance of the system is then obtained from an analysis of the steady-state regime of this random process.

Modeling the problem

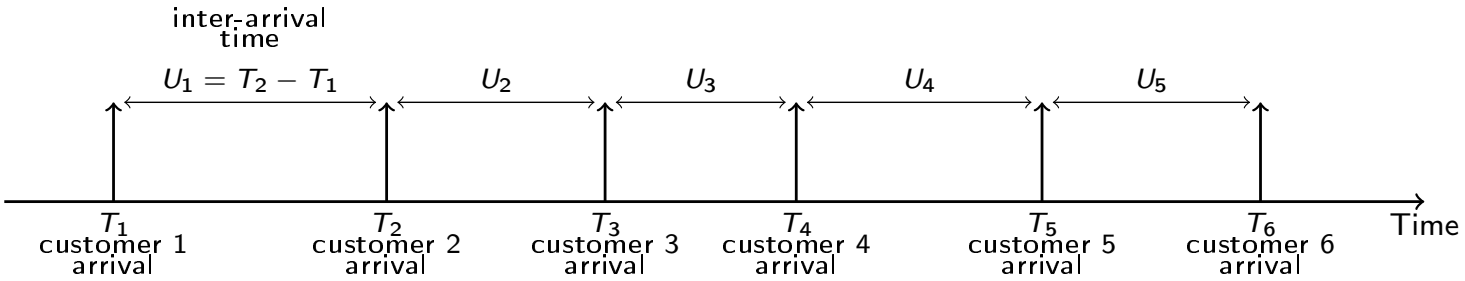
The system is characterized by

- customer arrival process (e.g., Poisson process)
- service duration law (e.g., exponential law)
- number of servers
- maximum number of customers in the system
- service discipline (e.g., FIFO=FirstInFirstOut) , possibly service classes (priorities), etc...



The first step of this modeling is to characterize the system. The system has different characteristics.
First of all we have to characterize the arrival process of customers, in other words, the way customers arrive in the system. The customer arrival process is often modeled as a Poisson process.
Then, we have to characterize the service duration, meaning the time which is required to serve each client. Note that the service duration is often modeled as an exponential law.
It is also necessary to specify the number of servers, since different servers can work in parallel and take on different customers.
Customers can wait in a buffer if all the servers are busy, so we have to define the total maximum number of clients in the system (including the clients that are being served and the waiting clients).
And finally we have to clarify the scheduling policy, or... the order in which customers access servers, the most common policy being FIFO (First In First Out), meaning that customers access the servers in the order in which they arrived. In other words, first come, first served.

Customer arrivals



T_i, **arrival time** of customer number i.

$U_i = T_{i+1} - T_i$, **interarrival time** between customers number i and $(i+1)$.

Let's go back to the characterization of customer arrivals. The arrival process is a point process, meaning a set of points on the timeline. Each point denoted by T_i on this graph represents the arrival time of customer number i.
The inter-arrival time denotes the time between two consecutive customers. It is denoted here by U_i which is equal to T_{i+1} minus T_i and represents the interval between the arrival of customer i and the arrival of customer $(i+1)$.

Customer arrivals

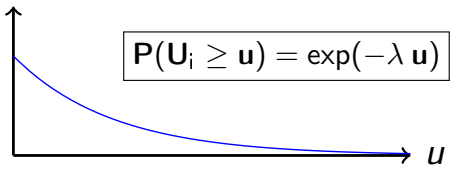
- In simple models, we assume that the **successive inter-arrival times** U_1, U_2, U_3, etc...
  - are **independent**
  - follow the **same probability distribution**
- The customer arrival process is then fully characterized by the **law of inter-arrival times**.

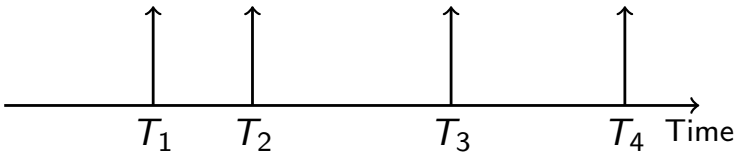
In simple models it is assumed that the successive inter-arrival times, or U_1, U_2, U_3, etc... are mutually independent and that they follow the same probability distribution. In this case the customer arrival process is therefore fully characterized by the law of inter-arrival times.

Poisson arrivals

The Poisson process is an important particular case of arrival processes.

- The successive inter-arrival times U_1, U_2, U_3, etc... are independent, and distributed according to the **exponential law with parameter** λ
- Equivalently, the arrival process is a **Poisson process with parameter** λ





Exponential interarrival times $\Leftrightarrow$ **Poisson arrival process**

It is very often assumed that the law of successive inter-arrival times is an exponential distribution, whose parameter is classically denoted by λ. The probability that the inter-arrival time U_i is greater than a given threshold u is thus equal to the exponential of minus λu.
In this case, when the inter-arrival times are independent and identically distributed according to an exponential distribution with parameter λ, the arrival process is said to be a Poisson process of rate λ.

Arrival rate λ

- The arrival rate λ is the **average number of arrivals per time unit** (e.g., customers/second).
- The average inter-arrival time equals $1/\lambda$ (time units, e.g. seconds).

λ is called the arrival rate. It represents an average number of customer arrivals per time unit. So its unit is the inverse of the time unit, for example customers per second.
Reciprocally the average time between two consecutive arrivals equals 1 divided by λ and is stated in, for example, seconds.