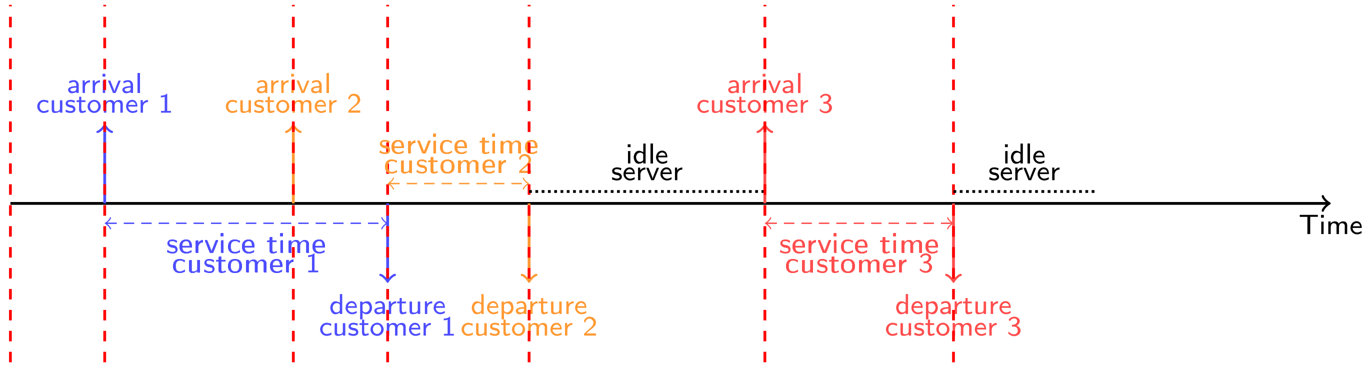
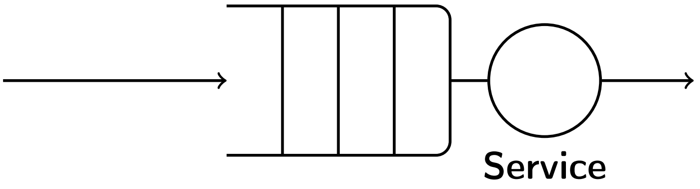
# What is a queue?

**Service durations**

## Service durations

The service time $S_i$ of customer $i$ is the time during which a server is busy serving customer $i$.





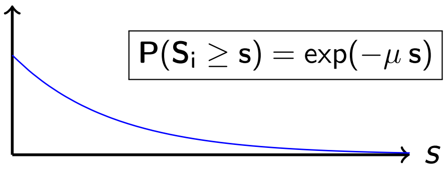
In order to characterize the system it is also necessary to properly define the service duration, which refers to the time it takes to serve a customer. Let's go back to our simple example of a queue with one server and an infinite waiting buffer and let's consider the evolution of this system over time. When the blue customer arrives, the server is free so the service of the blue customer starts immediately. A new, orange-colored customer arrives before the server has finished with the blue one, so the orange customer waits in the buffer. When the blue customer has been served, that customer leaves the system and releases the server for the orange one. The service duration of the blue client is here the time between arrival and departure (since the blue client did not wait to be served). Then the service of the orange customer starts. The service duration of the orange customer is the elapsed time between beginning of service and departure. It is represented here with a horizontal orange dashed line. As no customer has arrived in the meantime, the system is thus empty and the server is idle until a new customer arrives. And so on and so forth for further customers...

# Exponential service durations

In general, it is assumed that:

- service times are independent and identically distributed;
- they are characterized by their probability distribution.

The exponential distribution is an important particular case of service duration.



**Exponential service duration**

It is generally assumed that service durations are random independent and identically distributed variables. The service duration is fully characterized by its probability distribution. Again, the exponential distribution is a very important and classic particular case of service duration. In this case it is assumed that, if $S_i$ denotes the service duration for customer $i$, then the probability that $S_i$ is greater than a threshold $s$ equals the exponential of $-\mu s$.

## Service rate $\mu$

- The **service rate** $\mu$ is the average number of clients served per time unit if the server is always busy (e.g. customers/second).
- The **mean service duration** is $1/\mu$ (time units, e.g. seconds).

$\mu$ is classically used as the parameter of the service duration distribution. The unit of mu is the inverse of the time unit (for example second$^{-}$1). The average service duration is one divided by $\mu$.
Reciprocally, $\mu$ is the average number of clients that the server is able to serve (per time unit) if it is never idle.

# Offered load

## Notations

- The **arrival rate** $\lambda$ is the average number of arrivals per time unit.
- The **service rate** $\mu$ is the average number of customers a server is able to handle if it is always busy. Equivalently, $1/\mu$ is the average service duration.

Average performance depends on the offered load $\rho$, defined as

$$\boxed{\rho = \frac{\lambda}{\mu}(\text{unit: Erlang})}$$

The unit of $\rho$ is the Erlang, a dimensionless unit.

The offered load denoted by $\rho$ is an important parameter. Indeed the average system performance depends on it.
Assume that $\lambda$ is the arrival rate, meaning the average number of customer arrivals per time unit. And assume that $\mu$ is the service rate ; equivalently the average service duration equals one divided by $\mu$.
Then the offered load $\rho$ is defined as the ratio of $\lambda$ to $\mu$. Stated differently, $\rho$ is the product of the average number of customers per time unit and the average service duration per customer. It is a dimensionless quantity and it is stated in the Erlang unit.