Continuous-Time Markov Chains

PASTA property

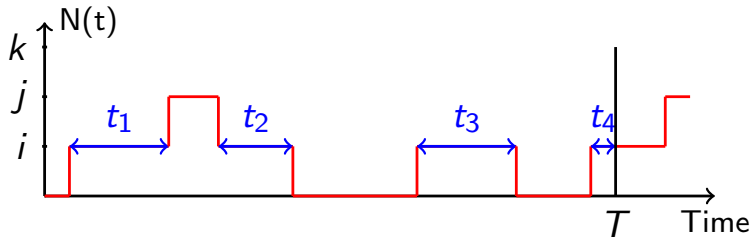
We finish this week with a crucial property satisfied only by Poisson processes.

Ergodic theorem

For a CTMC of stationary probability π

$$\text{Percentage of time in state } i \text{ up to time } T \xrightarrow{T\to\infty} \pi_i$$

$$\text{i.e. } \frac{1}{T}\sum t_k \xrightarrow{T\to\infty} \pi_i$$



Consider the evolution of the number of customers $N(t)$ when time goes by.
Look at the time intervals when $N(t)$ is equal to, say, a value i. In the picture, these intervals are of length t_1, t_2, t_3, etc.
The ergodic theorem for continuous time Markov chains says that for a Continuous Time Markov Chain with stationary probability π, the percentage of time in state i approaches π_i as the observation time tends to infinity. This means that $t_1 + t_2 + t_3 + ...$ etc divided by T, the observation time, tends to π_i when T goes to infinity. This is called the ergodic theorem.

M/M/1 again

- Stationary distribution : $\pi_i = \rho^i(1-\rho)$
- Idle time corresponds to 0 customer in the system
- Percentage of idle time up to time T approaches $\pi_0 = 1 - \rho$.

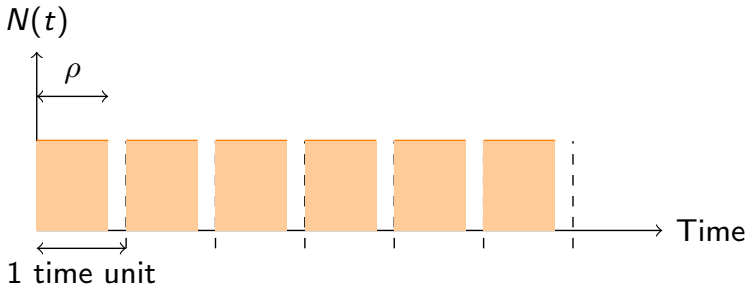
For instance, for the M/M/1 queue, π_i equals $\rho^i(1-\rho)$. What is called idle time corresponds to the period of time during which the system is empty. The ergodic theorem says that asymptotically, the idle time is almost equal to $\pi_0 = 1 - \rho$.

Blocking and loss probabilities

- **Blocking probability :** probability that *at time t*, all the resources are occupied, i.e. the system is blocked
- Also equal to the *percentage of time* during which the system is blocked
- **Loss probability :** probability that a customer finds the system blocked *at his arrival*

Having seen the ergodic theorem, a new problem arises for queueing analysts.
The blocking probability is the probability at steady state that the system is full, and therefore can't accept any further customers. According to the ergodic theorem, this probability is equal to the percentage of time during which the system is blocked.
On the other hand, what really interests the practitioners is the loss probability : the probability that a customer finds the system blocked upon his or her arrival. These two probabilities have no reason to coincide in general.

Counterexample: the D/D/1/1 queue



- **Blocking probability** = percentage of the orange area = ρ
- **Loss probability** = 0

Consider for instance, the D/D/1/1 queue. The inter-arrival and service times are deterministic, there is one server and no buffer. We choose the time unit that corresponds to the inter-arrival time. The service duration is ρ, which is assumed to be less than 1.
The time during which the system is blocked corresponds to the orange part of the picture. So, the percentage of time the system is blocked corresponds to the percentage of the orange area, which is equal to ρ. If we follow the previous definitions, this means that the blocking probability is equal to ρ. On the other hand, at each arrival the system is empty, so the loss probability is zero. So, there is a priori no relationship between blocking and loss probability.

PASTA property

- Poisson Arrivals See Time Averages
  For Poisson arrivals,

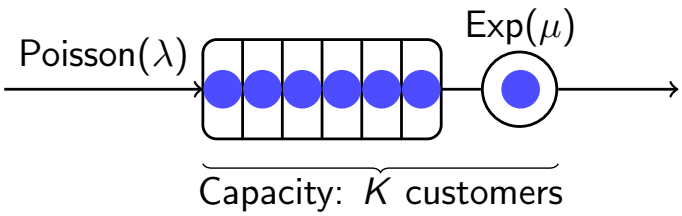
$$\text{blocking probability} = \text{loss probability}$$

- For a queue represented by a continuous time Markov chain,

$$\text{blocking probability} = \sum_{i \in \text{blocking states}} \pi_i$$

However, there may be such a relationship and this property is known as the PASTA property. PASTA is the acronym for "Poisson Arrivals See Time Averages" or, in other words, if the arrivals in a system occur according to a Poisson process, then the blocking and loss probabilities do coincide.
For a queue represented by a Continuous Time Markov Chain, the blocking probability can be computed from the stationary distribution. So, we have a way to really find the quantity which interests the practitioners, namely the loss probability.

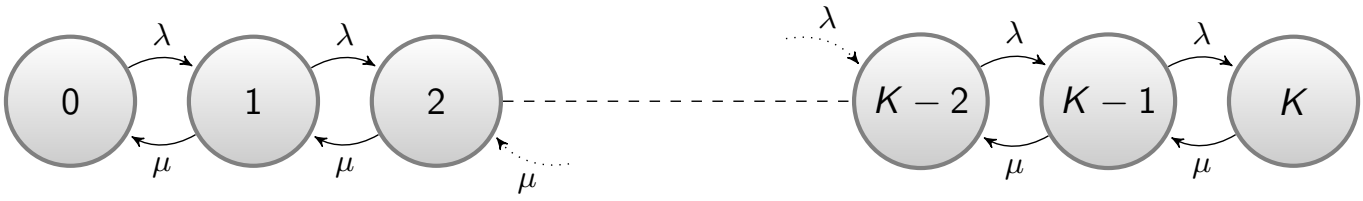
M/M/1/K



Given $\rho = \lambda/\mu$, find K such that the loss probability
is less than a given threshold

Let's have a look at the finite capacity M/M/1 queue, namely the M/M/1/K queue, meaning that the buffer has a finite capacity of size $K - 1$.
The practitioner's goal is to find the right value of K: if it's too big, it will maybe cost too much; if it's too small, there will be a huge loss probability, meaning bad quality of service and thus unhappy customers.
The quality of service is measured here by the loss probability, so knowing the load ρ, the problem is to find K such that the loss probability is less than a desired threshold.

Example : M/M/1/K



$$\mu\,\pi_1 = \lambda\,\pi_0$$
$$\mu\,\pi_2 = \lambda\,\pi_1$$
$$\mu\,\pi_3 = \lambda\,\pi_2$$
$$\vdots$$
$$\lambda\,\pi_{K-1} = \mu\,\pi_K$$

The transition diagram is the truncation to $\{0, 1, 2, ..., K\}$ of the M/M/1 diagram. It is a birth-death process so we can apply the results we saw before: the local balance equations should be satisfied.

$$\begin{aligned} \mu\pi_1 &= \lambda\pi_0, \\ \mu\pi_2 &= \lambda\pi_1, \\ &\text{? and so on and so forth until} \\ \mu\pi_K = \lambda\pi_{K-1}. & \end{aligned}$$

M/M/1/K continued

- Stationary distribution

$$\pi_i = \pi_0 \, \rho^i \quad \text{and} \quad \sum_{i=0}^{K} \pi_i = 1$$

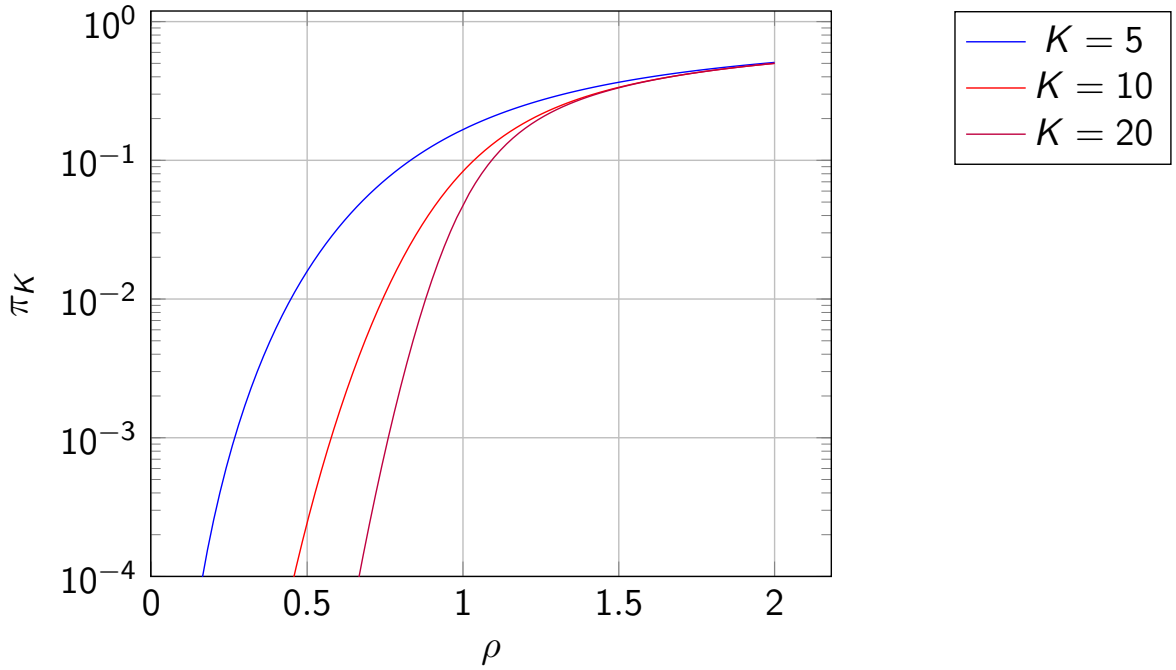
$$\Rightarrow \quad \pi_i = \frac{\rho^i}{1+\rho+\rho^2+\ldots+\rho^K} = \frac{1}{1-\rho^{K+1}} \, (1-\rho) \, \rho^i$$

- One blocking state: K
- Loss probability:

$$\text{Loss probability} = \pi_K = \frac{1-\rho}{1-\rho^{K+1}} \, \rho^K$$

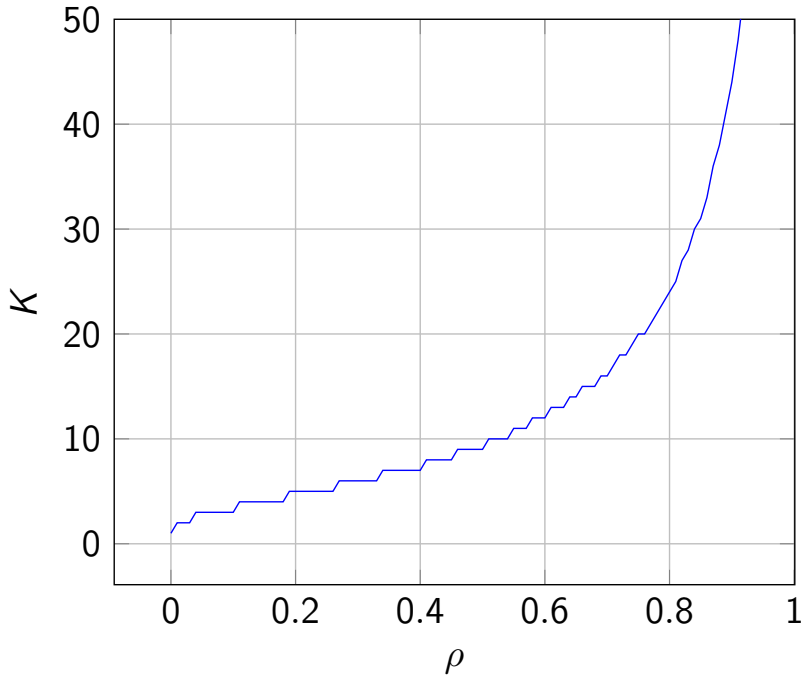
Then the stationary probability of state i turns out to be proportional to ρ^i. Using the normalization condition and the formula for the sum of the terms of a geometric sequence, we get: $\pi_i = \rho^i(1-\rho)/(1-\rho^{K+1})$.
We have one blocking state which corresponds to a full system, meaning that the number of customers is equal to K . And since the arrivals are Poisson, the loss probability is equal to π_K.

π_K vs ρ



We can plot this loss probability for K fixed against ρ. It is more meaningful to represent it in a semi-log scale since the values are fortunately rather small.
Three observations are worth mentioning. Doubling the size of the buffer does much more than dividing the loss probability by 2. For instance for $\rho = 0.5$, the loss probability for $K = 10$ is almost two orders of magnitude less than the loss probability for $K = 5$.
In practice, this is very convenient since it means that adding a few resources may drastically improve the performance of the system. On the other hand, doubling the load for a fixed capacity does much more damage than doubling the loss probability. For instance for $K = 5$, there is more than one order of magnitude between the loss probability for $\rho = 0.5$ and the loss probability for $\rho = 1$.
The third remark is that mathematically speaking, the system is always stable since we have a finite state space. However, as ρ gets bigger, the loss probability tends to 1. Of course the bigger K is, the bigger ρ has to be in order to reach a given threshold of loss probability but inevitably, it goes to 1.

K such that $\pi_K \leq 0.001$



The last plot is the evolution of K which guarantees the loss probability to be less than 0.001 (a rather severe requirement) when ρ varies. Once again, it is important to notice the non linearity of K. For low loads, K varies slowly whereas for higher traffic, the variations are considerable and the values soon become very high.