

Multi-Server Systems

Waiting Systems – Erlang-C and Application

Erlang-C– Waiting Probability

- The waiting probability of an $M/M/C/\infty$ system is referred to as the Erlang-C formula
- It is expressed as a function of the offered load ($\rho = \lambda/\mu$) and the system's capacity (C).
- It is obtained from the steady-state distribution

We will now see another famous formula proposed by Erlang, which has received the name “Erlang-C” formula. It represents the waiting probability of an $M/M/C$ system, and is expressed as a function of the offered load, namely ρ , and the capacity of the system, namely C . This formula is obtained from the steady-state distribution

Erlang-C– Waiting probability

- PASTA property \Rightarrow The waiting probability is given by $\sum_{j \geq C} \pi_j$

$$E_C(\rho, C) = \frac{\frac{\rho^C}{C!} \frac{C}{C-\rho}}{\sum_{k=0}^C \frac{\rho^k}{k!} + \frac{\rho^C}{C!} \frac{\rho}{C-\rho}}$$

Indeed, on the one hand, we know that an arriving customer will have to wait if, at arrival, all servers are busy. On the other hand, the PASTA property, seen in week 4, states that if we have Poisson arrivals, then an arriving customer sees the system in steady state. Thus, the probability of waiting is given by the sum of the probabilities of the system being at any state equal to or greater than C . Since we have already calculated the steady state distribution, we can compute this sum to obtain Erlang-C formula

Erlang-C – Mean Performance

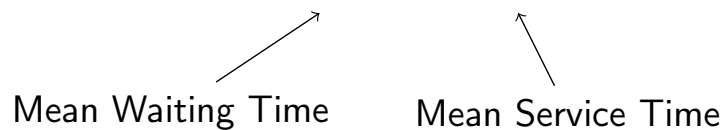
- Mean Number of clients in the system:

$$N = \frac{\rho}{C - \rho} E_C(\rho, C) + \rho$$

- Mean sojourn time:

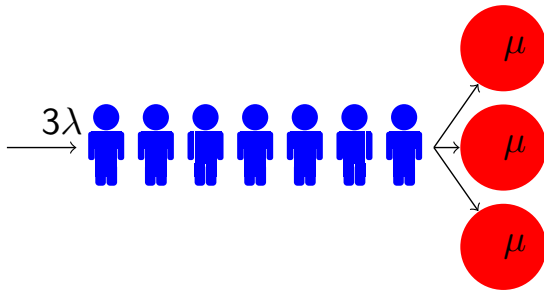
$$R = \frac{E_C(\rho, C)}{\mu(C - \rho)} + \frac{1}{\mu}$$

Mean Waiting Time Mean Service Time



Thanks to the steady state distribution and to Little's law we can deduce some mean performance metrics for the M/M/C system. For instance, we can compute the mean number of customers in the system and the mean sojourn time, where we can distinguish a component of the mean waiting time, and a component of the mean service time.

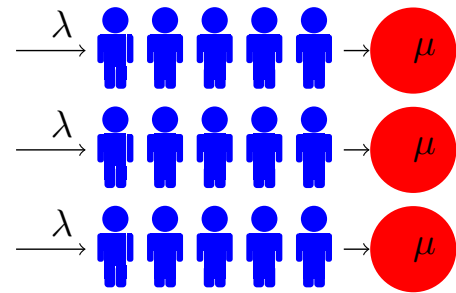
Which System Performs Better? $\lambda = 3 \text{ h/h}$, $1/\mu = 15\text{min} \Rightarrow \rho = 0.75$



One queue with three servers.

\Rightarrow
An M/M/3 system.

$$W = \frac{E_C(3\rho, 3)}{\mu(3 - 3\rho)} = 11.36\text{min}$$



Three queues with one server each.

\Rightarrow
Three M/M/1 systems.

$$W = \frac{\rho}{\mu - \lambda} = 45\text{min}$$

As a take away example, let's now analyze and compare the performance of these two systems. On the left hand side, we have one queue and three servers, on the right hand side, we have three servers and one single queue per server. We assume that service times for all servers are exponentially distributed, with the same parameter μ .

Let's assume that to the single queue, 3λ customers arrive per unit of time, and to each of the three queues at the right hand side, λ customers arrive per unit of time. We also assume that arrivals follow a Poisson process. With these assumptions we can model the left hand side system as an M/M/3 queue with offered traffic equal to 3ρ and the right hand side system as three M/M/1 queues with offered load equal to ρ .

So, which system do you think performs better?

Let's consider the mean waiting time as a metric, and put in some numerical values. We'll consider that on average 9 customers arrive per hour to the queue on the left, and 3 to each one of the queues on the right. Let the mean service time be equal to 15 minutes. So we have that the server utilization for both systems is equal to 0.75 Erlang. Computing the waiting time at each system, we obtain approximately 11 minutes for the multiserver system, and 45 minutes at each of the three single-server systems, which answers our question.

Beyond this particular example, thanks to the mutualization of resources, a multiserver system will always perform better than a single-server queue, for equal server utilization.