First batch of exercises - Software Performance and Scalability

Andrea Munarin (879607)

March 2023

1 Exercise 1: Conditioned response time on the M/M/1 system

Let us consider an M/M/1 queue. Assume that although the system does not know in advance the service time of a job, the user sending the request knows it.

Find the expression of the expected response time of a job whose service time is exactly L.
SOLUTION
The response time of a job is defined as the sum of the waiting time and the service time for that job: $r = w + s$. In order to compute the expected response time, we need to consider the average number of jobs in the queue. Then we have to sum

- The service time required for the job (that we know is L).
- The mean of the service time for each job in the queue. If we have an average of $\overline{N}$ jobs in the system, we know that there are $\overline{N} - 1$ jobs in the queue waiting.
- The service time of job currently served. Thanks to the PASTA property (we can use it because our system in modelled with Poisson arrivals) when we arrive in the system we know that we have to wait, on average, $\frac{1}{\mu}$ until the current job is complete

So the expression of the expected response time of the job is:

$$\overline{R} = L + (\overline{N} - 1) \cdot \frac{1}{\mu} + \frac{1}{\mu} = L + \frac{\overline{N}}{\mu}$$

Now, assume that $\lambda = 1$ job/s and $\mu = 1.5$ job/s, what is the expected response time of a job whose service time is 4 seconds?
SOLUTION
In order to compute the expected response time, we need to use the formula

derived in the previous point. To evaluate the results, we need to retrieve the expected number of jobs. Since $\lambda \leq 1 \cdot \mu$ our system is stable and therefor we can apply the Little's Law:

$$\overline{N} = \lambda \cdot \overline{R}$$

Now substituing this we obtain an equation with only one variable and we can easily resolve it:

$$\overline{R} = L + \frac{\lambda}{\mu} \cdot \overline{R}$$

$$\overline{R} - \frac{\lambda}{\mu} \cdot \overline{R} = L$$

$$\overline{R} \cdot (\frac{\mu - \lambda}{\mu}) = L$$

$$\overline{R} = L * \frac{\mu}{\mu - \lambda} = 4 * \frac{1.5}{1.5 - 1} = 12s$$

2 System with vacation

A system processes jobs with a FCFS order. Arrivals follow a Poisson process and the service time is exponentially distributed. Every 1 hour, the system runs a self-diagnostic of 2 minutes. During this times, it drops all arriving requests and does not serve any job. Once the self-diagnostic terminates, the system begins to work for another hour and the cycle restarts.

What percentage of jobs are dropped (use PASTA)?
SOLUTION
Since the arrivals are distributed with a Poisson Process and the inter-arrival times follow an Exponential Distribution (memoryless property) we can apply the PASTA property: the probability of the state as seen by an outside random observer is the same as the probability of the state seen by an arriving customer. For a random observer the probability to see the system while it is performing the self-diagnostic operation is $\frac{2}{60+2}$ so we can write the probability of dropping for an arriving customer as:

$$P_{drop} = \frac{2}{60+2} = \frac{1}{31} = 0.0323 \rightarrow 3.23\%$$

What is the throughput of the system if the arrival rate is 3 job/s and the system is stable?
SOLUTION
If we know the probability of dropping, the throughput of the system is easy to compute. We have that 0 job/s leaves the system while the self-diagnostic is operating and 3 job/s while the system is normally working. So, using the probability obtained before, we can write:

$$X = P_{drop} \cdot \lambda_{self-diagnostic} + (1 - P_{drop}) \cdot \lambda_{operational}$$

$$= \frac{1}{31} \cdot 0 + \frac{30}{31} \cdot 3 = 2.903 \ job/s$$

Hard part: now, assume that the self-diagnostic begins at random times after an average of 1 hour of operations. The operating time periods are exponentially distributed (independent by everything). The length of the maintenance is always deterministically 2 minutes. Compute the expected response time. [Tips: you need PASTA at the beginning of the maintenance and Little's law]

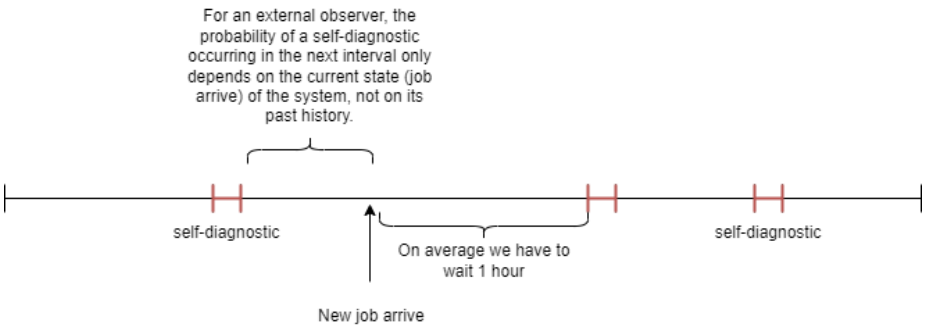

Figure 1: Sketch system with vacation

SOLUTION

When we arrive at the system, we know that the average time before maintenance is one hour[1], thanks to the PASTA and memoryless property (see Figure1). Thus, we need to add two minutes, on average, for every 60 minutes of system working. The expected response time is the sum of the average self-diagnostic time and working time, which can be expressed as follows:

- Self-diagnostic time: $\frac{\frac{\overline{N}-1}{\mu}+\frac{1}{\mu}+\frac{1}{\mu}}{3600} \cdot 120$
- Working time: $\frac{\overline{N}-1}{\mu}+\frac{1}{\mu}+\frac{1}{\mu}$

The expected response time is therefore:

$$\overline{R} = \frac{\frac{\overline{N}-1}{\mu}+\frac{1}{\mu}+\frac{1}{\mu}}{3600} \cdot 120 + \frac{\overline{N}-1}{\mu}+\frac{1}{\mu}+\frac{1}{\mu}$$

$$\overline{R} = \frac{\overline{N}+1}{\mu} \cdot \frac{31}{30}$$

Since the system is stable, we can use Little's Law to relate the average number of jobs in the system to the average response time:

$$\overline{N} = X \cdot \overline{R}$$

[1]It is important to note that if we arrive at the system during the maintenance period, on average, we need to wait one minute before the self-diagnostic ends (as it is deterministic). However, computing the expected response time does not make sense since our jobs are dropped and never enter the system.

Substituting the expression for $\overline{R}$ above, we obtain:

$$\overline{R} = \frac{31 \cdot X}{30 \cdot \mu} \cdot \overline{R} + \frac{31}{30 \cdot \mu}$$

$$\overline{R} \cdot \left(\frac{30 \cdot \mu - 31 \cdot X}{30 \cdot \mu} \right) = \frac{31}{30 \cdot \mu}$$

$$\overline{R} = \frac{31}{30 \cdot \mu} \cdot \frac{30 \cdot \mu}{30 \cdot \mu - 31 \cdot X} = \frac{31}{30 \cdot \mu - 31 \cdot X}$$

$$\overline{R} = \frac{31}{30 \cdot \mu - 90} s$$

Since the system is stable, we know that μ must be greater than or equal to 3 jobs per second. Thus, we can represent how the expected response time varies based on μ.