MTL 106 (Introduction to Probability Theory and Stochastic Processes)
Assignment 2 Report

Name: Arpit Saxena Entry Number: 2018MT10742

1. Basic Probability
2. Random Variable/Function of a Random Variable
3. Stochastic Processes
4. Stochastic Processes
5. DTMC
6. DTMC
7. CTMC
8. CTMC
9. Queueing Models

   A company has one 16-core machine, two 8-core machines and two 4-core machines. They want to use them as servers. The inter arrival time of queries is exponentially distributed with mean 0.1ms. They estimate that the time taken by a core per query would be exponentially distributed with mean time 3, 2, 4 milliseconds for the 16-core, 8-core and 4-core machines respectively. They want to set up a simple static load balancer in front of these machines, which will schedule the queries on a core of a machine with a probability to assign the query to each core.

   Determine the probabilities by which the load balancer should schedule queries on each type of core to minimise the maximum expected waiting time for a query. Note that one query would occupy the core on which it's running for the entire time it's running.

   Answer

   Basically, what we want to do here is to divide the incoming queries into 3 different types of cores (they are different since they have different service time distributions) We can tabulate the information as follows:

Number of machines	Number of cores	Total number of cores	Mean service time(ms)
1	16	16	3
2	8	16	2
2	4	8	4

   Number the types of cores given in the above table as 1, 2, 3 and let p_1, p_2, p_3 denote the probabilities by which a query will be sent to a core of that type by the load balancer.

Given that the incoming queries form a Poisson process with parameter $\frac{1}{0.1 \text{ ms}} = 10 \text{ ms}^{-1}$ and the load balancer is decomposing this Poisson process into separate streams. Then, the queries going to cores of type 1, 2, 3 will form a Poisson process with parameters $10p_1, 10p_2, 10p_3$ respectively and $16p_1 + 16p_2 + 8p_3 = 1$ since each query will be routed to one of the given cores.

We now model each core as a M/M/1 queue. Generically, let's take the arrival process parameter as λ and the service time parameter as μ.

Let W be the waiting time for a query, then $P(W < 0) = 0$ and $P(W = 0) = \rho$ where $\rho = \frac{\lambda}{\mu}$.

For $W > 0$, there has to be atleast one person in the system. Assuming there are n people in the system, we have $W = \widetilde{S_1} + S_2 + \cdots + S_n$, where $\widetilde{S_1}$ is the time left for query that is already running, and $S_2 \cdots S_n$ are the running times for the remaining queued queries.

Since service times are exponentially distributed, which has memoryless property, $\widetilde{S_1}$ is also exponentially distributed with parameter μ

Therefore, $W/N = n \sim Gamma(n, \mu)$ since it is a sum of n independent exponentially distributed random variables with parameter μ.

For $t > 0$,

$$\begin{aligned} P(W \le t) &= \sum_{n=1}^{\infty} P(W \le t/N = n)P(N = n) \\ &= 1 - \rho + \sum_{n=1}^{\infty} \int_0^t \frac{\mu^n x^{n-1} e^{-\mu x}}{(n-1)!}\, dx\, (1-\rho)\rho^n \end{aligned}$$

Now, taking summation inside the integral, we get

$$\begin{aligned} P(W \le t) &= 1 - \rho + \int_0^t \sum_{n=1}^{\infty} \frac{(\mu x \rho)^{n-1}}{(n-1)!} e^{-\mu x}\, dx\, (1-\rho)\mu\rho \\ &= 1 - \rho + \int_0^t \mu\rho e^{\mu x \rho} e^{-\mu x}\, dx\, (1-\rho) \\ &= 1 - \rho + \mu\rho(1-\rho)\frac{e^{-\mu(1-\rho)t} - 1}{\mu(\rho - 1)} \\ &= 1 - \rho + \rho(1 - e^{-(\mu-\lambda)t}) \\ &= 1 - \rho e^{-(\mu-\lambda)t} \end{aligned}$$

Therefore, the CDF of W in the steady state is given by

$$P(W \le t) = \begin{cases} 0 & t < 0 \\ 1 - \rho & t = 0 \\ 1 - \rho e^{-(\mu-\lambda)t} & 0 < t < \infty \end{cases}$$

where $\rho = \frac{\lambda}{\mu}$ and the steady state solution is only possible when $\rho < 1 \implies \lambda < \mu$

Then the pdf of W is given by $f_W(t) = \rho(\mu - \lambda)e^{-(\mu-\lambda)t}$ when $0 < t < \infty$ and 0 otherwise.

$$
\begin{aligned}
E(W) &= \int_0^\infty t\rho(\mu - \lambda)e^{-(\mu-\lambda)t}\, dt \\
&= \frac{\rho}{\mu - \lambda}\int_0^\infty e^{-(\mu-\lambda)t}(\mu - \lambda)t\, d[(\mu - \lambda)t]
\end{aligned}
$$

Since $\mu - \lambda > 0$, $(\mu - \lambda)t \to \infty$ as $t \to \infty$

$$
\begin{aligned}
E(W) &= \frac{\rho}{\mu - \lambda}\int_0^\infty te^{-t}\, dt \\
&= \frac{\rho}{\mu - \lambda} \\
&= \frac{\lambda}{\mu^2 - \lambda\mu} && \left(\text{Since } \rho = \tfrac{\lambda}{\mu}\right)
\end{aligned}
$$

Now substituting the values of μ as $\frac{1}{3}, \frac{1}{2}, \frac{1}{4}$ for cores 1, 2, 3 respectively and also using the query process parameters as calculated above, we get the expected waiting times for cores 1, 2, 3 respectively as:

$$\frac{90p_1}{1 - 30p_1}, \frac{40p_2}{1 - 20p_2}, \frac{160p_3}{1 - 40p_3}$$

We also need to have $\rho < 1$ for all the cores, i.e. $\frac{10p_i}{\mu_i} < 1$ for all the cores, which gives $p_1 < \frac{1}{30}, p_2 < \frac{1}{20}, p_3 < \frac{1}{40}$

So our problem reduces to the following optimisation problem:

$$\text{Minimise } max\left\{\frac{90p_1}{1 - 30p_1}, \frac{40p_2}{1 - 20p_2}, \frac{160p_3}{1 - 40p_3}\right\}$$

$$\text{In the domain } 16p_1 + 16p_2 + 8p_3 = 1, p_1 < \frac{1}{30}, p_2 < \frac{1}{20}, p_3 < \frac{1}{40}$$

Solving the equations with the aid of computational tools available, we find that the minimum expected waiting time is approximately 4.77 ms, when $p_1 \approx 0.020, p_2 \approx 0.035, p_3 \approx 0.013$

So, we have the probabilities by which the load balancer should send the queries to cores of type 1 as $16p_1 \approx 0.327$, cores of type 2 as $16p_2 \approx 0.564$ and cores of type 3 as $8p_3 \approx 0.109$ for minimisation of the maximum expected waiting time for each query.

10. Queueing Models

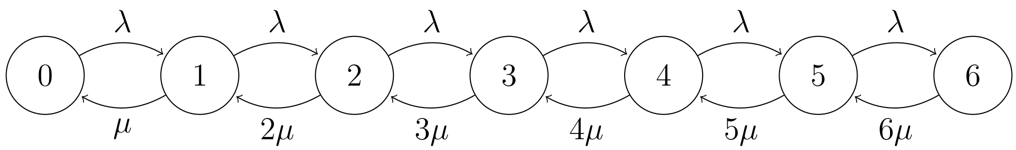
A restaurant on a very busy main road has a parking capacity of 6 cars. People whose car does not get a parking space can't go to eat in the restaurant due to the busyness of the road and the unavailability of any other parking nearby. People in a car are binomially distributed as B(5, 0.5) (0 included since some people just park their car and not eat). Arrival of cars follows a Poisson process with average inter arrival time 10 minutes. Time spent in the restaurant by one cars passengers is exponentially distributed with average as 40 minutes.

Draw the state diagram of the underlying process, and derive the steady state equations. What is the expected number of people in the restaurant in steady state?

Answer

By the description, we can observe the cars follow a M/M/6/6 queue. Let λ be the parameter of the arrival process and μ be the parameter of the service time exponential distribution. Then, by the given information, $\lambda = \frac{1}{10}$ min^{-1} and $\mu = \frac{1}{40}$ min^{-1}

The state diagram of the underlying birth death process is then given by:



Here the numbers in each state indicate the number of cars parked in the parking lot. Let $\{X_t \,|\, t \in \mathbb{R}\}$ be the underlying stochastic process where X_t denotes the number of cars parked in the parking lot at time t.

Define $\pi_n = \lim_{t\to\infty} \text{Prob}(X_t = n)$. Then using Kolmogorov, forward equations and letting $t \to \infty$, we get:

$$
\begin{aligned}
0 &= -\lambda\pi_0 + \mu\pi_1 \\
0 &= \lambda\pi_{i-1} + (i+1)\mu\pi_{i+1} - (\lambda + i\mu)\pi_i, 1 \le i \le 5 \\
0 &= \lambda\pi_5 - 5\mu\pi_6
\end{aligned}
$$

From the first equation, we get $\pi_1 = \frac{\lambda}{\mu}\pi_0$. For $i = 1$:

$$
\begin{aligned}
0 &= \lambda\pi_0 + 2\mu\pi_2 - (\lambda + \mu)\pi_1 \\
\implies 0 &= \lambda\pi_0 + 2\mu\pi_2 - (\lambda + \mu)\frac{\lambda}{\mu}\pi_0 \\
\implies 2\mu\pi_2 &= \frac{\lambda^2}{\mu}\pi_0 \\
\implies \pi_2 &= \frac{\lambda^2}{2\mu^2}\pi_0
\end{aligned}
$$

By induction, we can show that

$$\pi_i = \frac{\lambda^i}{i!\,\mu^i}\pi_0,\ 1 \le i \le 5$$

And then using the last equation, we get:

$$\begin{aligned} 0 &= \lambda\pi_5 - 5\mu\pi_6 \\ \implies 0 &= \lambda\frac{\lambda^5}{5!\,\mu^5}\pi_0 - 6\mu\pi_6 \\ \implies 6\mu\pi_6 &= \frac{\lambda^6}{5!\mu^5}\pi_0 \\ \implies \pi_6 &= \frac{\lambda^6}{6!\mu^6}\pi_0 \end{aligned}$$

Define $\rho = \frac{\lambda}{\mu}$ Using the fact that the total probability must be 1, which is also called the normalising condition, we get:

$$\begin{aligned} \sum_{n=0}^{6} \frac{\rho^n}{n!}\pi_0 &= 1 \\ \implies \pi_0 &= \frac{1}{\sum_{n=0}^{6} \frac{\rho^n}{n!}} \end{aligned}$$

Let X be the random variable denoting the number of cars in the parking lot at steady state. Then,

$$\begin{aligned} E(X) &= \sum_{n=0}^{6} n \cdot \frac{\rho^n}{n!}\pi_0 \\ &= \sum_{n=0}^{6} n \cdot \frac{\frac{\rho^n}{n!}}{\sum_{n=0}^{6} \frac{\rho^n}{n!}} \\ &= \rho \cdot \frac{\sum_{n=1}^{6} \frac{\rho^{n-1}}{(n-1)!}}{\sum_{n=0}^{6} \frac{\rho^n}{n!}} \\ &= \rho \cdot \frac{\sum_{n=0}^{5} \frac{\rho^n}{n!}}{\sum_{n=0}^{6} \frac{\rho^n}{n!}} \\ &= \rho\left(1 - \frac{\frac{\rho^6}{6!}}{\sum_{n=0}^{6} \frac{\rho^n}{n!}}\right) \end{aligned}$$

Using $\lambda = \frac{1}{10}, \mu = \frac{1}{40}$ we get $\rho = 4$. Then we can solve the solution using a calculator and get $E(X) \approx 3.531$

Now, we have the expected number of cars in steady state, but we need to calculate the expected number of customers in the restaurant in steady state.

Let N be the random variable denoting the number of customers in the restaurant in steady state. Let N_i be the number of passengers in car i, then we have $N = \sum_{i=0}^{X} N_i$, where X is the random variable denoting the number of cars in steady state. We know that N_i's are iid random variables which are distributed as $B(5, 0.5)$ from which we get $E(N_i) = 5 \times 0.5 = 2.5$ Using Wald's equation since N_i's are iid and also independent from X, we get

$$\begin{aligned} E(N) &= E(N_1) \times E(X) \\ &\approx 2.5 \times 3.531 \\ &= 8.828 \end{aligned}$$

Therefore, the expected number of customers in the restaurant in steady state is approximately 9.