**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 $\lambda$ and the service time parameter as $\mu$.

Using the result derived in class, let $W$ be the waiting time for a query, then its CDF in the steady state is given as

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

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 $\mu$ 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