~~Chapter 12~~ ~~Question 8~~ Tutorial 8, Question 2

To find the service demands of an HTTP request at the CPU:

The throughput of the server is

$$= \frac{10800}{3600}$$

$$= 3 \text{ HTTP requests/s}$$

By service demand law

service demand of CPU

$$= \frac{\text{utilisation of CPU}}{\text{server throughput}}$$

$$= \frac{0.3}{3}$$

$$= 0.1 \text{ s}$$

To find the throughput:

The easiest is to write a computer program and plug the values in.

$X(0) = 0$

$X_0(1) = 6.25$ requests/s

$X_0(2) = 8.16$ requests/s

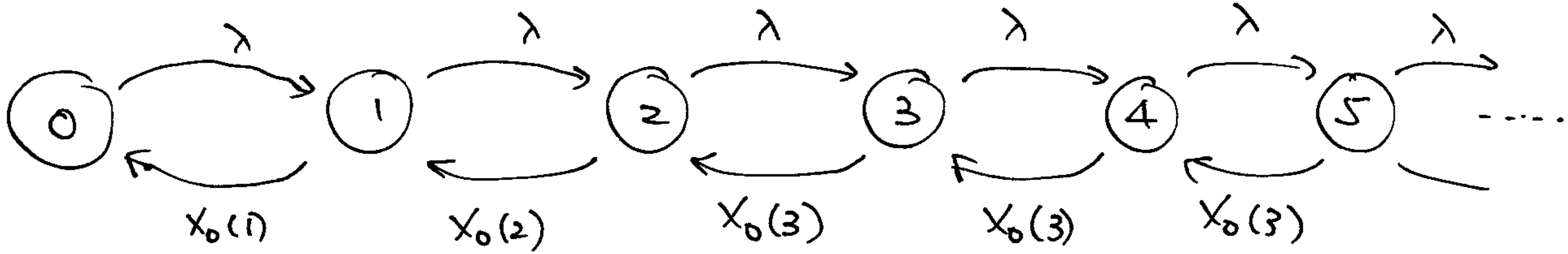
$X_0(3) = 9.01$ requests/s

Part 3 of the question: To find the average response time of HTTP requests when $\lambda = 5$ and the server can only process up to 3 requests at a time:

This can be modelled as a ~~generalised~~ Markov Chain ~~birth-death model~~.

Let state k $(k = 0, 1, 2, \ldots)$ be the number of requests in the web server. Note that the number of requests in the web server includes those that are being served (up to three) and those that are in the processing queue.

The state space diagram is in the following page:



* The transition rate from State k to State $(k+1)$ (for $k=0,1,\ldots$) is the arrival rate of the request.

* The transition rate from State $(k+1)$ to State k (for $k=0,\ldots$) is the rate at which requests are completed.

  - For State 1 to State 0, this is the same as the throughput of the web server when there is only one client. (Note that throughput is effectively the number of requests completed in an unit time.)
  - For state 2 to State 1, the request completion rate is $X_0(2)$.

- For state 3 to state 2, the request completion rate is $X_0(3)$
- For state $(k+1)$ to state k (where $k \geq 3$), the request completion rate is always $X_0(3)$ because only 3 requests are being processed by the server. The others requests are waiting in the queue.

In order to find the response time, we need to solve the model. Using the trick given in the notes, we know that

$$P(1) \quad X_0(1) = \lambda \; P(0)$$

$$P(2) \quad X_0(2) = \lambda \; P(1)$$

$$P(3) \quad X_0(3) = \lambda \; P(2)$$

$$P(4) \quad X_0(3) = \lambda \; P(3)$$

$$P(5) \quad X_0(3) = \lambda \; P(4)$$

- - . . .

Expressing $P(1), P(2), \ldots$ in terms of $P(0)$, we have

$$P(1) = \frac{\lambda}{X_0(1)} P(0)$$

$$P(2) = \frac{\lambda}{X_0(2)} \frac{\lambda}{X_0(1)} P(0)$$

$$P(3) = \frac{\lambda}{X_0(3)} \frac{\lambda}{X_0(2)} \frac{\lambda}{X_0(1)} P(0)$$

$$P(4) = \left(\frac{\lambda}{X_0(3)}\right)^2 \frac{\lambda}{X_0(2)} \frac{\lambda}{X_0(1)} P(0)$$

$$P(5) = \left(\frac{\lambda}{X_0(3)}\right)^3 \frac{\lambda}{X_0(2)} \frac{\lambda}{X_0(1)} P(0)$$

observing the pattern, we've

$$P(k) = \left(\frac{\lambda}{X_0(3)}\right)^{k-2} \frac{\lambda}{X_0(2)} \frac{\lambda}{X_0(1)} P(0)$$

for $k \geq 3$

Define
$$\begin{cases} \rho_1 = \dfrac{\lambda}{X_0(1)} \\ \rho_2 = \dfrac{\lambda}{X_0(2)} \\ \rho_3 = \dfrac{\lambda}{X_0(3)} \end{cases}$$

we have

$$P(1) = \rho_1 \; P(0)$$

$$P(2) = \rho_2 \rho_1 \; P(0)$$

$$P(k) = \rho_3^{k-2} \rho_2 \rho_1 \; P(0) \quad \text{for } k \geq 3$$

Since the sum of all probabilities must be 1,

$$P(0) + P(1) + \cdots + \cdots = 1$$

$\cancel{\rho_1 P(0)}$

$$P(0) + \rho_1 P(0) + \rho_2 \rho_1 P(0) + \rho_3 \rho_2 \rho_1 P(0) + \rho_3^2 \rho_2 \rho_1 P(0) + \cdots = 1$$

(geometric progression) (Note: you can use the formula given in the question)

$$\Leftrightarrow \quad P(0) + \rho_1 P(0) + \frac{\rho_2 \rho_1}{1 - \rho_3} P(0) = 1$$

(Note that $\rho_3 < 1$, so the geometric progression converges)

$$\Rightarrow \quad P(0) = \frac{1}{1 + \rho_1 + \frac{\rho_2 \rho_1}{1 - \rho_3}}$$

In order to calculate the response time, we need to compute the throughput and ~~queue length first~~ the mean # requests in the server.

Throughput

$$= X_0(1)\, P(1) + X_0(2)\, P(2) + X_0(3)\left(P(3) + P(4) + \cdots\right)$$

$$= X_0(1)\cdot \rho_1\, P(0) + X_0(2)\cdot \rho_2 \rho_1\, P(0) + X_0(3)\left(\rho_3\rho_2\rho_1\, P(0) + \rho_3^2\rho_2\rho_1\, P(0) + \cdots\right)$$

$$= X_0(1)\, \rho_1\, P(0) + X_0(2)\, \rho_2\rho_1\, P(0) + X_0(3)\, \rho_3\rho_2\rho_1\, P(0)\cdot \frac{1}{1-\rho_3}$$

$$= \left(X_0(1)\,\rho_1 + X_0(2)\,\rho_2\rho_1 + X_0(3)\,\frac{\rho_3\rho_2\rho_1}{1-\rho_3}\right)\cdot \frac{1}{1+\rho_1+\frac{\rho_2\rho_1}{1-\rho_3}}$$

You can plug the values of ρ_1, ρ_2, ρ_3, $X_0(1)$, $X_0(2)$ and $X_0(3)$ into the expression.

The mean # requests in the server is

$$0\,P(0) + 1 \cdot P(1) + 2\,P(2) + 3\,P(3) + \dots$$

$$= \rho_1 P(0) + 2\rho_2\rho_1 P(0) + 3\rho_3\rho_2\rho_1 P(0) + 4\rho_3^2\rho_2\rho_1 P(0) + 5\rho_3^3\rho_2\rho_1 P(0)$$

$$= \rho_1 P(0) + \rho_2\rho_1 P(0)\left[\underbrace{2 + 3\rho_3 + 4\rho_3^2 + 5\rho_3^3 + \dots}\right]$$

Need to recognise that this is an arithmetic-geometric progression. You can ~~look up a~~ use the given formula ~~book~~ or you can derive the result. Let us derive the result:

Let $z = 2 + 3\rho_3 + 4\rho_3^2 + 5\rho_3^3 + \dots$ — ①

then $\rho_3 z = 2\rho_3 + 3\rho_3^2 + 4\rho_3^3 + \dots$ — ②

Subtracting ② from ①, we have

$$(1-\rho_3)z = 2 + \rho_3 + \rho_3^2 + \dots$$

$$= 2 + \frac{\rho_3}{1-\rho_3}$$

$$\Rightarrow \quad Z = \frac{\rho_3}{(1-\rho_3)^2} + \frac{2}{1-\rho_3}$$

(Alternatively, substitute $p=2$, $m=\rho_3$, $\xi=1$ in the given formula)

Thus the mean # requests in the web server is

$$= \rho_1 P(0) + \rho_2 \rho_1 P(0) \cdot \left[\frac{\rho_3}{(1-\rho_3)^2} + \frac{2}{(1-\rho_3)} \right]$$

You can now plug the values of ρ_1, ρ_2, ρ_3 and $P(0)$ to find the mean # requests

Finally, to obtain the mean response time, we use Little's Law

$$\text{Mean response time} = \frac{\text{mean \# requests in the server}}{\text{mean throughput}}$$