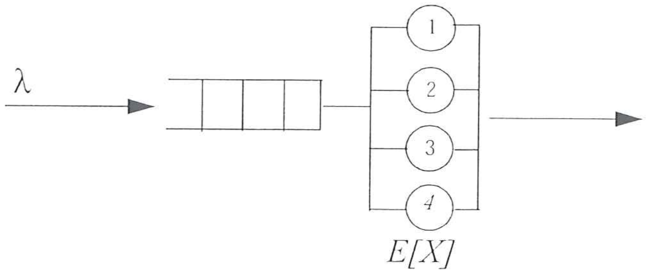
Attila

Solutions manual
for
“Queuing Systems”

Solutions for the exercises in Chapter 3 (Basic queueing theoretic formulas)

1. ✓ The telephone support can be modelled as the queueing system below:

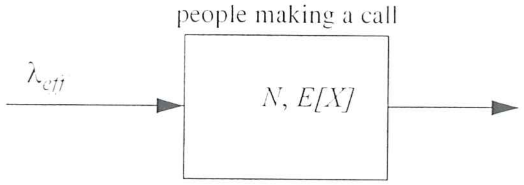


The average arrival rate of calls *per hour* can be estimated as: $\lambda = \frac{1000}{5 \cdot 9} \approx 22\ s^{-1}$, h^{-1}

The maximum processing capacity of the system is given by $\frac{c}{E[X]} = \frac{4}{E[X]}$.

The stability condition is given by : $\lambda < \frac{c}{E[X]} \Rightarrow E[X] < \frac{c}{\lambda} \approx 0,18$ *hours*, which is the same as 10 minutes and 48 seconds.

2. We can describe the number of active telephone calls in the town with the following model:



N is the average number of ongoing telephone calls, and $E[X]$ is the average length of a telephone call.

The average arrival rate of new calls *per minute* can be calculated as $\lambda = \frac{50000 \cdot 3}{24 \cdot 60} \approx 104$.

Since no customers are blocked we can assume that the *effective arrival rate*, λ_{eff}, is equal to λ.

The average length of a telephone call is $E[X] = 6$ minutes.

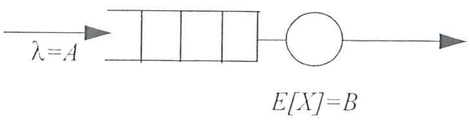
Little's formula can now be used to determine the average number of ongoing telephone calls: $N = \lambda_{eff} \cdot E[X] = 104 \cdot 6 \approx 625$

3. Average arrival rate: A jobs per second; Average service time: B seconds.

(a) One server means that the stability condition is: $\lambda < \frac{c}{E[X]} \Rightarrow \begin{bmatrix} \lambda = A \\ c = 1 \\ E[X] = B \end{bmatrix} \Rightarrow A < \frac{1}{B}$.

(b) m servers means that the stability condition is: $\lambda < \frac{c}{E[X]} \Rightarrow \begin{bmatrix} \lambda = A \\ c = m \\ E[X] = B \end{bmatrix} \Rightarrow A < \frac{m}{B}$.

(c) The system is modelled as below:



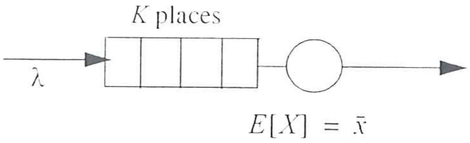
The throughput is always equal to the effective arrival rate, λ_{eff}. In this case no jobs are blocked, which means that $\lambda_{eff} = \lambda = A$ jobs per second.

(d) The interarrival times and service times must be exponentially distributed. Also, both the interarrival times and the service times must be independent between jobs and uncorrelated to each other. Finally, the servers must be chosen randomly.

(e) The interarrival times must be exponentially distributed. Also, both the interarrival times and the service times must be independent between jobs and uncorrelated to each other. Finally, the servers must be chosen randomly.

(f) T=average response time for a job. Little's formula gives: $T = \dfrac{\overline{N}}{\lambda_{eff}} = N/A$ seconds.

4. We model the web site as below:



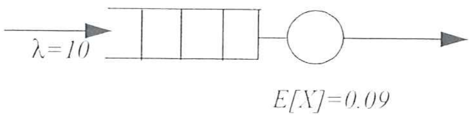
(a) The response time for a particular job always consists of the waiting time in queue plus the service time. The average response time, T, can be determined as $T = W + \bar{x}$, where W is the average waiting time. When the arrival rate is very low ($\lambda \to 0$), $W \to 0$, which means that $T \to \bar{x}$.

In this case: $T \to 0,2$ when $\lambda \to 0$, which means that $\bar{x} = 0,2$ seconds.

(b) When $\lambda \to \infty$, the queue will almost always be full. A job that enters the system will take the last place in the queue. This means that the average number of jobs in the system, $N=K+1$. Each time a job is completed, a new job can enter, which means that the effective arrival rate will be $\lambda_{eff} = \dfrac{1}{\bar{x}}$. Little's formula then gives: $T = \dfrac{\overline{N}}{\lambda_{eff}} = (K+1) \cdot \bar{x} \Rightarrow K = \dfrac{T}{\bar{x}} - 1$.

By studying the diagram we can see that when $\lambda \to \infty$, $T \approx 4 \cdot \bar{x}$ which means that $K=3$.

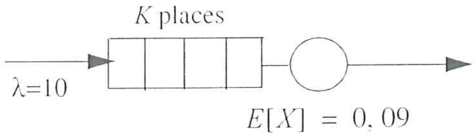
5. First, we model the system as below:



(a) The carried load is defined as $\rho_c = \lambda_{eff} \cdot E[X]$. In this case we have an infinite queue, which means that no jobs are blocked and that $\lambda_{eff} = \lambda = 10$. Therefore, the carried load is given by $\rho_c = 10 \cdot 0,09 = 0,9$ Erlangs.

(b) The throughput is always equal to the effective arrival rate, in this case 10 jobs per second.

(c) New model:



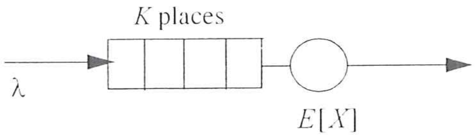
The throughput is 8 jobs per second, which means that $\lambda_{eff} = 8$.

The probability that a job is blocked is therefore: $P_b = \dfrac{\lambda - \lambda_{eff}}{\lambda} = \dfrac{2}{10} = 0,2$.

(d) The carried load is given by $\rho_c = \lambda_{eff} \cdot E[X] = 8 \cdot 0,09 = 0,72$ Erlangs.

(e) In the first system, all customers are admitted to the system. However, this can result in a longer response time and thereby a lower user-perceived QoS. In the second system some customers are blocked, which means that the admitted customers are guaranteed a short response time. However, the blocked customers are probably not too happy.

6. The node is modelled as below:



(a) When $\lambda \to \infty$ the system is almost always full. Every time a job is completed, a new job enters the system. This means that the average service time is $E[X] = \lim\limits_{\lambda \to \infty} \dfrac{1}{\lambda_{eff}} = \dfrac{1}{10}$ seconds.

(b) When $\lambda \to \infty$ the system is almost always full. This means that the average number of jobs in the system becomes $N = K + 1$. Studying diagram *(b)* we see that $\lim\limits_{\lambda \to \infty} N = 6$, which means that K=5.

(c) By looking in the diagrams we find that when λ=40: $\lambda_{eff} \approx 10$ and $N \approx 5,7$. By using Little's theorem we get : $T = \dfrac{N}{\lambda_{eff}} \approx 0,57$ seconds.