SOLUTION NOTES

Computer Systems Modelling 2001 Paper 8 Question 14 (TLH)

The initial section of the question concerns a simple M/M/1 queue. The Markov chain is, in this case, a birth-death process. The states should be labelled with the sum of the number of customers in the queue and in service. Arrivals occur at rate $p_1\lambda$, departures occur at rate μ_1.

The state residence probabilities and mean response time are as calculated on slides 108-111 (attached) after substituting these arrivals and departure rates.

Suppose that the system administrator wishes to ensure that customers receive the same mean response time irrespective of which server they visit. Express p_1 in terms of λ, μ_1 and μ_2. Qualitatively, when is it reasonable to consider dispatching work to both servers to maintain an equal mean response time? How will the system behave outside this interval?

For equal response times we require

$$\begin{aligned}
\frac{1}{\mu_1 - \lambda p_1} &= \frac{1}{\mu_2 - \lambda p_2} \\
\Rightarrow \quad \frac{1}{\mu_1 - \lambda p_1} &= \frac{1}{\mu_2 - \lambda(1 - p_1)} \\
\Rightarrow \quad 2\lambda p_1 &= \lambda + \mu_1 - \mu_2 \\
\Rightarrow \quad p_1 &= \frac{\lambda + \mu_1 - \mu_2}{2\lambda p_1}
\end{aligned}$$

For equal response times to be achievable we require $p_1 \in (0, 1)$. Qualitatively:

- If the load is too high then the system is not stable and the response time will tend to infinity on both servers.
- If the load is too low, and the service rates are skewed, then the entire load can be handled by the faster server. This will occur if the response time of the faster server, when receiving the full load, is less than the service time of the slower server.

Stochastic balance (3)

Since the sum of state probabilities must be unity,

$$
\begin{aligned}
p_0 + \sum_{k=1}^{\infty} p_k &= 1 \\
p_0 + \sum_{k=1}^{\infty} p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} &= 1 \\
p_0 \left[1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right] &= 1
\end{aligned}
$$

so that

$$
\begin{aligned}
p_0 &= \left[1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right]^{-1} \\
p_k &= p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}
\end{aligned}
$$

which are known as the **general flow balance equations**

The M/M/1 queue

The BDP maps well onto our domain of study – queueing systems

Births represent arrivals to queue, deaths represent departures as customers finish service

The M/M/1 queue is an infinite customer system, with infinite waiting room, and a state independent service rate

This means that $\lambda_i = \lambda$ and $\mu_i = \mu$ for all i and we can solve the balance equations:

$$
\begin{aligned}
p_k &= p_0 \prod_{i=0}^{k-1} \frac{\lambda}{\mu} \\
&= p_0 \left(\frac{\lambda}{\mu} \right)^k .
\end{aligned}
$$

The M/M/1 queue (2)

Writing $\rho = \dfrac{\lambda}{\mu}$

$$\begin{aligned} p_0 &= \frac{1}{1+\sum_{k=1}^{\infty}\rho^k} \\ &= \frac{1}{1+\rho\sum_{k=0}^{\infty}\rho^k} \\ &= \frac{1}{1+\rho\left(\frac{1}{1-\rho}\right)} \\ &= 1-\rho \end{aligned}$$

Consequently, the number in the system is geometrically distributed

$$p_k = (1-\rho)\rho^k, \qquad k = 0, 1, 2, \ldots$$

If $\rho > 1$, i.e. if $\lambda > \mu$ the system will not reach equilibrium

The M/M/1 queue (3)

What is the average number of customers, $\overline{N}$ in the system?

$$\begin{aligned} \overline{N} &= \sum_{k=0}^{\infty} k p_k \\ &= \sum_{k=0}^{\infty} k(1-\rho)\rho^k \\ &= (1-\rho)\rho\frac{\partial}{\partial\rho}\left(\sum_{k=0}^{\infty}\rho^k\right) \\ &= (1-\rho)\rho\frac{\partial}{\partial\rho}\left(\frac{1}{1-\rho}\right) \\ &= \frac{\rho}{1-\rho} \end{aligned}$$

The M/M/1 queue (4)

An arriving customer will find, on average $\overline{N}$ in the system, and will spend a time, say $\overline{T}$, in the system. During $\overline{T}$ there will be, on average $\lambda\overline{T}$ arrivals, leaving $\overline{N}$ customers in the queue. Thus

$$\overline{N} = \lambda\overline{T}$$

which is Little's result restated. In our case

$$\begin{aligned}\overline{T} &= \frac{\overline{N}}{\lambda} \\ &= \frac{\frac{1}{\mu}}{1-\rho}\end{aligned}$$

which is the M/M/1 average response time.

Note that

- $\frac{1}{\mu}$ is the average service time
- ρ is the utilization

Performance at high load

At high utilizations ρ approaches one and the response time and queue lengths are unbounded.