Question 6

a)
- Applying the M/M/1 delay formula to router B where the average service time $= \frac{1}{\mu_B} = \frac{L}{R_{AB}}$, we get $T_{BA} = \frac{1}{\mu_B - \lambda_{all}}$.
- Similarly, applying the M/M/1 delay formula to router A where the average service time $= \frac{1}{\mu_A} = \frac{L}{R_{AS}}$, we get $T_{AS} = \frac{1}{\mu_A - \lambda_{all}}$.
- Thus, our model approximates $T_{response}$ as

$$T_{response} = T_{serve} + \frac{1}{\mu_B - \lambda_{all}} + \frac{1}{\mu_A - \lambda_{all}}$$

b) Assuming that $L = 850000$ bits, $\lambda_{all} = 16$ requests/sec, $T_{serve} = 3$ sec, $R_{SA} = 100$ Mbps, and $R_{AB} = 15$ Mbps, we get the following approximate values:

- $\mu_B = 17.647$, $T_{BA} = 0.607$ sec
- $\mu_A = 117.65$, $T_{AS} = 0.01$ sec
- $T_{response} = 3 + 0.607 + 0.01 = 3.617$ sec

c) After installing the cache server, on average 60% of the requests are served from the server, and 40% of the requests are served from the Internet servers. Thus,

$$T_{response} = 0.6 \times T_{response\_internal} + 0.4 \times T_{response\_internet}.$$

where

- $T_{response\_internal} = \frac{L}{R_{AS}}$ (this term may be ignored)
- $T_{response\_internet} = T_{serve} + \frac{1}{\mu_B - 0.4\lambda_{all}} + \frac{1}{\mu_A - 0.4\lambda_{all}}$

d) Assuming the parameter settings in part (b), we get the following approximate values:

- $T_{response\_internal} = 8.5$ msec
- $T_{response\_internet} = 3.089$ sec
- $T_{response} = 0.6 \times 0.0085 + 0.4 \times 3.089 = 1.24$ sec

---