Chapter 2, Exercise 1 page 31

$M/M/1$ queue.

Mean interarrival time 1.25 minutes.

Mean service time 1 minute.

Arrival rate λ is 4 arrivals per 5 minutes. Service rate μ is 5 service completions in 5 minutes. (The choice of "5 minutes" is convenient and arbitrary.) Therefore, server utilization $\rho = \lambda / \mu = 0.8$. This is a dimensionless number. The server, in the long run, is busy 80% of the time.

$L = \rho / (1 - \rho) = 0.8/0.2 = 4$ (dimensionless, represents average number of entities in system).

From page 24, $L = \lambda W$ (Little's law) or $W = L / \lambda$. We want W to be measured in minutes, so take λ as 0.8 arrivals per minute.

Then $W = 4 / (0.8/\text{minute}) = 5$ minutes. On average, an entity spends 5 minutes in the system, counting both waiting time and service time.

Also from page 24, we have $W = W_q + E(S)$. So $W_q = W - E(S) = 5$ minutes – 1 minute = 4 minutes. On average, an entity spends 4 minutes waiting in queue.

Also from page 24, $L_q = \lambda W_q = 0.8/\text{minute} * 4$ minutes = 3.2 entities (the "minutes" units cancel). On average, the waiting line contains 3.2 items.

Now, it is easy to believe, as is indeed the case, that, on average, total time in system = time in queue + time in service.

A bit less obvious: Average number in queue = 3.2. Average number in system = 4 (not 4.2 = average number in queue + one in service). Often, the number in system will be one more than the number in queue, because one is in service. That is true even when one entity is being served but the queue is empty. But it is *not* true when the system is completely empty and quiescient.