Single Server Queue

Let $\overline{r}$ be the Mean Interarrival Time, $\overline{s}$ be the Mean Service Time:

$$\overline{r} = \frac{1}{n}\sum_{i=1}^{n} r_i \quad \text{and} \quad \overline{s} = \frac{1}{n}\sum_{i=1}^{n} s_i. \tag{0.1}$$

The Mean Arrival Rate (λ) and Mean Service Rate (μ) are defined as:

$$\lambda = \frac{1}{\overline{r}} \quad \text{and} \quad \mu = \frac{1}{\overline{s}}. \tag{0.2}$$

Definition 0.1. The Traffic Intensity (or Server Utilization) is defined as:

$$\rho = \frac{\lambda}{\mu}. \tag{0.3}$$

At a given time t, let

$q(t)$ be the number of jobs in the queue (waiting for service),

$x(t)$ be the number of jobs at the server (0 or 1),

$l(t)$ be the number of jobs in the system (service node).

Note that q, x and l are all functions of time t and

$$l(t) = q(t) + x(t). \tag{0.4}$$

Let $\overline{l}, \overline{q}$ and $\overline{x}$ be the mean number of jobs in the system (the service node), the mean number of jobs in the queue (waiting for service) and the mean number of jobs with the server respectively.

Lemma 0.2. *If the simulation runs from time 0 to time T,*

$$\begin{aligned}
\overline{l} &= \frac{\textit{Area under the curve } l(t)}{T} = \frac{\int_0^T l(t)dt}{T}, \\
\overline{q} &= \frac{\textit{Area under the curve } q(t)}{T} = \frac{\int_0^T q(t)dt}{T}, \\
\overline{x} &= \frac{\textit{Area under the curve } x(t)}{T} = \frac{\int_0^T x(t)dt}{T}, \\
\overline{l} &= \overline{q} + \overline{x}.
\end{aligned}$$

Let $\overline{w}, \overline{d}$ and $\overline{s}$ be the mean waiting time for each job, the mean delay time for each job and the mean service time for each job. Thus:

$$\overline{w} = \frac{1}{n}\sum_{i=1}^{n} w_i, \quad \overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i, \quad \overline{s} = \frac{1}{n}\sum_{i=1}^{n} s_i.$$

If n is the total number of jobs (arrivals) and T is the length of the simulation (assume that $T = a_n$). Then,

(a) Total waiting time $=\sum_{i=1}^{n} w_i$ = Area under the curve $l(t)$;

(b) Total delay time $= \sum_{i=1}^{n} d_i$ = Area under the curve $q(t)$;

(c) Total service time $= \sum_{i=1}^{n} s_i$ = Area under the curve $s(t)$.

Theorem 0.3. *(**Little' Results**)*
Consider the mean number of jobs in the system (service node) $\overline{l}$ and the mean waiting time for each job $\overline{w}$, we have

$$\overline{l} = \lambda\overline{w}; \quad \overline{q} = \lambda\overline{d}; \quad \overline{x} = \lambda\overline{s}$$

here, λ is the mean arrival rate.

Remark 0.4. From Little's Result, we can have

$$\overline{x} = \lambda\overline{s} = \lambda\frac{1}{\mu} = \rho.$$

So, traffic intensity or server utilization is the mean number of jobs at the server in one time unit. It is the percentage of server busy time over the length of simulation.

Remark 0.5. Little's result is true for arbitrary scheduling discipline and queueing systems.