Important Concepts in SSQs

We use two "arrival" measures, **they are inverses of each other**:

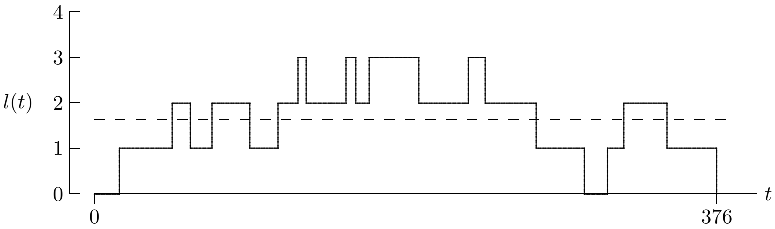
Arrival Rate	(jobs per unit time)	$\lambda = \frac{n}{a_n}$
Average interarrival time	(time between jobs)	$\bar{r} = \frac{a_n}{n}$

Utilization, the time domain average of the "server component", $\bar{x}$ can also be considered the probability that there is at least one job in the SSQ:

$$Pr(\ l(t) > 0\ )$$

Integration of NES or DES measures is simply the sum of rectangles, **by the very nature of DES**.

- All three functions are *piece-wise constant*



- Figures for $q(\cdot)$ and $x(\cdot)$ can be deduced

$$q(t) = 0 \text{ and } x(t) = 0 \text{ if and only if } l(t) = 0$$

Little's Theorem

How are job-averaged and time-average statistics related?

Theorem (Little, 1961)

If (a) queue discipline is FIFO,
(b) service node capacity is infinite, and
(c) server is idle both at $t = 0$ and $t = c_n$
then

$$\int_0^{c_n} l(t)dt = \sum_{i=1}^{n} w_i \qquad \text{and}$$

$$\int_0^{c_n} q(t)dt = \sum_{i=1}^{n} d_i \qquad \text{and}$$

$$\int_0^{c_n} x(t)dt = \sum_{i=1}^{n} s_i$$

Proof

We want to show:

$$\int_0^{c_n} l(t)\,\mathrm{d}t = \sum_i^n w_i$$

Define an indicator function:

$$\psi_i(t) = \begin{cases} 1 & a_i < t < c_i \\ 0 & \text{otherwise} \end{cases}$$

$\psi_i(t)$ "lights up" when job i is in the SSQ at time t. So we can write $l(t)$ (the number of jobs in the SSQ at time t) as

$$l(t) = \sum_{i=1}^n \psi_i(t) \quad 0 < t < c_n$$

Proof . . .

. . . taking the integral and rearranging the RHS:

$$\begin{aligned}
\int_0^{c_n} l(t)\,\mathrm{d}t &= \int_0^{c_n} \sum_{i=1}^{n} \psi_i(t)\,\mathrm{d}t && (1)\\
&= \sum_{i=1}^{n} \int_0^{c_n} \psi_i(t)\,\mathrm{d}t \quad \text{integral of sums is the sum of integrals} && (2)\\
&= \sum_{i=1}^{n} 1(c_i - a_i) \quad \text{each } \textstyle\int \psi_i(t) \text{ is a rectangle of width } c_i - a_i \text{ (height 1 job)} && (3)\\
&= \sum_{i=1}^{n} w_i \qquad w_i = c_i - a_i \quad \square && (4)
\end{aligned}$$

Little's Equations

- Using $\tau = c_n$ in the definition of the time-averaged statistics, along with Little's Theorem, we have

$$c_n \bar{l} = \int_0^{c_n} l(t)dt = \sum_{i=1}^{n} w_i = n\bar{w}$$

- We can perform similar operations and ultimately have

$$\bar{l} = \left(\frac{n}{c_n}\right)\bar{w} \quad \text{and} \quad \bar{q} = \left(\frac{n}{c_n}\right)\bar{d} \quad \text{and} \quad \bar{x} = \left(\frac{n}{c_n}\right)\bar{s}$$

Traffic Intensity

- *Traffic intensity*: ratio of arrival rate to service rate

$$\frac{1/\bar{r}}{1/\bar{s}} = \frac{\bar{s}}{\bar{r}} = \frac{\bar{s}}{a_n/n} = \left(\frac{c_n}{a_n}\right)\bar{x}$$

- Assuming c_n/a_n is close to 1.0, the traffic intensity and utilization will be nearly equal

Traffic Intensity provides a single metric by which we can classify FIFO SSQs:

$$\text{Traffic Intensity} = \frac{1/\bar{r}}{1/\bar{s}} = \frac{\bar{s}}{\bar{r}} = \begin{cases} > 1 & \textbf{???} \\ \approx 1 & \textbf{???} \\ < 1 & \textbf{???} \end{cases}$$

Traffic Intensity provides a single metric by which we can classify FIFO SSQs:

$$\text{Traffic Intensity} = \frac{1/\bar{r}}{1/\bar{s}} = \frac{\bar{s}}{\bar{r}} = \begin{cases} > 1 & q(t) \uparrow \quad l(t) \uparrow \\ \approx 1 & \bar{x} \approx 1 \\ < 1 & q(t) \approx 0 \quad \bar{x} < 1 \end{cases}$$

Traffic Intensity provides a single metric by which we can classify FIFO SSQs:

$$\text{Traffic Intensity} = \frac{1/\bar{r}}{1/\bar{s}} = \frac{\bar{s}}{\bar{r}} = \begin{cases} > 1 & \textbf{(queue blows up)} \\ \approx 1 & \text{very } \textbf{efficient} \\ < 1 & \text{waste of resources?} \end{cases}$$

Traffic Intensity provides a single metric by which we can classify FIFO SSQs:

$$\text{Traffic Intensity} = \frac{1/\bar{r}}{1/\bar{s}} = \frac{\bar{s}}{\bar{r}} = \begin{cases} > 1 & \textbf{(queue blows up)} \\ \approx 1 & \text{very } \textbf{efficient} \\ < 1 & \text{waste of resources?} \end{cases}$$

It appears that for “well behaved” FIFO SSQs:

$$\text{Traffic Intensity} \approx \bar{x}$$

Let's look at the cases that don't blow up...

$$\text{Traffic Intensity} = \frac{1/\bar{r}}{1/\bar{s}} = \frac{\bar{s}}{\bar{r}} = \frac{\bar{s}}{a_n/n} = \frac{n\bar{s}}{a_n}$$

Substituting the "average" Little's equations (author slide 23/1):

$$\bar{x} = \left(\frac{n}{c_n}\right)\bar{s} \quad \Rightarrow \quad c_n\bar{x} = n\bar{s}$$

And we have

$$\text{Traffic Intensity} = \frac{n\bar{s}}{a_n} = \left(\frac{c_n}{a_n}\right)\bar{x}$$

And we have proved the somewhat non-intuitive equation on author's slide 26/1 …
Hint: it must be a well-behaved FIFO SSQ **"at steady state"**.

What is c_n? What is a_n?

So **if** $\frac{c_n}{a_n} \approx 1$ we can surmise

$$\text{Traffic Intensity} \approx \bar{x}$$

Looking a little deeper into when $\frac{c_n}{a_n} \approx 1 \ldots$

Suppose traffic intensity is well behaved ($\approx$ or < 1) and service times are bounded $s_i < B$, a first job arriving at $a_1 = 3$ would have

$$\frac{c_1}{a_1} < \frac{3+B}{3} = 1 + \frac{B}{3}$$

which is possibly quite large.

The same fraction for job $n = 1000$ is arguably closer to 1:

$$1 + \frac{B}{1000}$$

Since $s_i < B$, **eventually for well behaved SSQs** (**queue not steadily growing**) the difference between c_n and a_n becomes very very small relative to the total SSQ lifetime. λ doesn't matter when we consider **efficient, well-behaved, long-lived SSQs**.

(Sounds like we need a $\lim_{n \to \infty}$)

Looking a little deeper into when $\frac{c_n}{a_n} \approx 1$ ...

$$
\begin{aligned}
\lim_{n\to\infty} \frac{c_n - a_n}{n} &= 0 \\
\lim_{n\to\infty} \frac{c_n}{n} - \frac{a_n}{n} &= 0 \\
\lim_{n\to\infty} \frac{c_n}{n} &= \lim_{n\to\infty} \frac{a_n}{n}
\end{aligned}
$$

And in the “steady state” ($n \to \infty$),

$$\frac{c_n}{n} = \frac{a_n}{n} = \bar{r} = \frac{1}{\lambda} \qquad (\lambda \text{ the } \textit{arrival rate}\text{, jobs per unit time})$$

... and Little’s “average” equations

$$\bar{l} = \left(\frac{n}{c_n}\right)\bar{w} \qquad \bar{q} = \left(\frac{n}{c_n}\right)\bar{d} \qquad \bar{x} = \left(\frac{n}{c_n}\right)\bar{s}$$

become

$$\bar{l} = \lambda\bar{w} \qquad \bar{q} = \lambda\bar{d} \qquad \bar{x} = \lambda\bar{s}$$

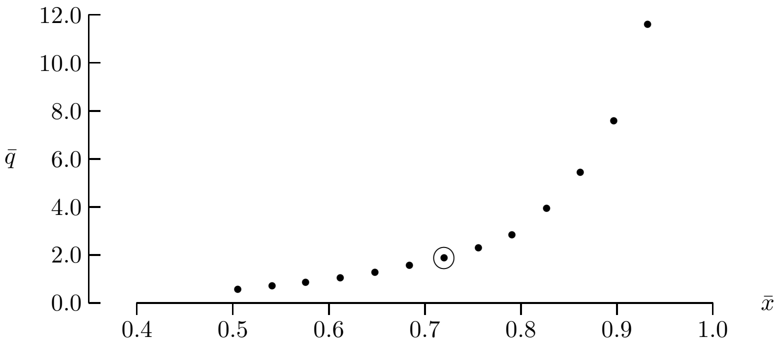
Case Study

Sven and Larry's Ice Cream Shoppe

- owners considering adding new flavors and cone options
- concerned about resulting service times and queue length

Can be modeled as a single-sever queue

- `ssq1.dat` represents 1000 customer interactions
- Multiply each service time by a constant
  - In the following graph, the circled point uses unmodified data
  - Moving right, constants are 1.05, 1.10, 1.15, ...
  - Moving left, constants are 0.95, 0.90, 0.85, ...



- Modest increase in service time produces significant increase in queue length
  - Non-linear relationship between $\overline{q}$ and $\overline{x}$
- Sven and Larry will have to assess the impact of the increased service times

Another Way to Think of this Approach

The authors generate the graph of $\bar{q}$ vs $\bar{x}$ by multiplying service times by a constant factor.

When they generated the data point for $0.95s_i$, what **traffic intensity** is represented by the data point?

Another Way to Think of this Approach

The authors generate the graph of $\bar{q}$ vs $\bar{x}$ by multiplying service times by a constant factor.

When they generated the data point for $0.95 s_i$, what **traffic intensity** is represented by the data point?

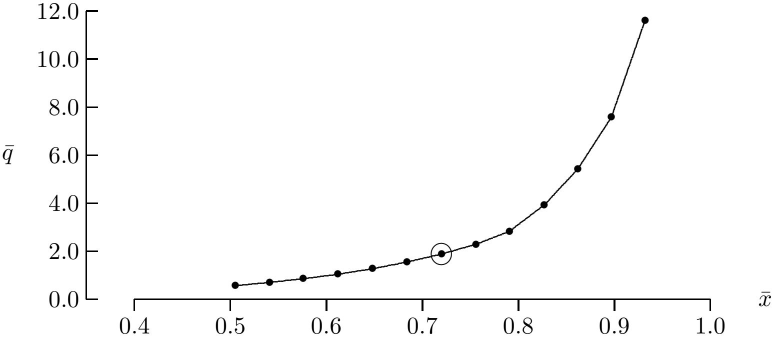
$$\text{Traffic Intensity} = \frac{1/\bar{r}}{1/\bar{s}} = \frac{\bar{s}}{\bar{r}}$$

So 95% service rate is a traffic intensity of

$$\frac{95\%\bar{s}}{\bar{r}} = 95\%\frac{\bar{s}}{\bar{r}}$$

or 95% of the real world (trace) data.

Note: they didn’t change the trace data arrival times, just the service times.



- Since both $\overline{x}$ and $\overline{q}$ are continuous, we could calculate an "infinite" number of points
- Few would question the validity of "connecting the dots"

Guidelines

- If there is essentially no uncertainty and the resulting interpolating curve is smooth, connecting the dots is OK
  - Leave the dots as a reminder of the data points
- If there is essentially no uncertainty but the curve is not smooth, more dots should be generated
- If the dots correspond to uncertain (noisy) data, then interpolation is not justified
  - Use approximation of a curve or do not superimpose at all
- Discrete data should *never* have a solid curve