Lecture 4 — Notes (Little's Theorem)

This lecture concerns one of the most important (and simplest) theorems in Queuing Theory, Little's Theorem. More information can be found in the course book, Bertsekas & Gallagher, section 3.2 (note that there are some minor errors and notational inconsistencies in their version) and in Kleinrock I, section 2.1. Little's Theorem is, in some ways, obvious. First we will introduce the theorem in a handwaving manner and then make the details of the definition more precise. Little's Theorem states:

$$N = \lambda T \tag{1}$$

Where N is the average number of customers in a queue, T is the average time a customer spends queuing and λ is the average rate of arrivals to the queue. In many ways this theorem represents an obvious truth — if a lot of people are in a queue (N is large) then they will have long delays (T is large); if few people arrive in a queue (λ is small) then the average number of people in the queue is small (N is small). This lecture will be spent making this intuitive idea more rigorous.

Let us first make precise the definitions of N, λ and T and then make clear the assumptions on which the theorem rests.

Let us first make some definitions:

$N(\tau)$ is the number of customers in the system at time τ.

$\alpha(\tau)$ is the number of customers who arrived in the interval $[0, \tau]$.

$\beta(\tau)$ is the number of customers who have departed in the interval $[0, \tau]$.

t_i is the time at which the ith customer arrived.

$T(i)$ is the time spent queuing by the ith customer.

If N_t is the mean value of $N(\tau)$ taken over the interval $[0, \tau]$ then it is clear that:

$$N_t = \frac{1}{t} \int_0^t N(\tau) d\tau \tag{2}$$

Let us assume:

$$N = \lim_{t \to \infty} N_t \tag{3}$$

(Note that this limit is *not* guaranteed to exist — imagine, for example, a queue which keeps growing.) If the limit exists, N is the *steady state time average* of $N(\tau)$.

We next define the average arrival rate over the time period $[0, t]$.

$$\lambda_t = \frac{\alpha(t)}{t} \tag{4}$$

and, again, we assume that the following limit exists:

$$\lambda = \lim_{t \to \infty} \lambda_t \tag{5}$$

Finally, the average delay experienced by those customers who enter the system at times in $[0, t]$ is given by:

$$T_t = \sum_{i=1}^{\alpha(t)} \frac{T(i)}{\alpha(t)} \tag{6}$$

And, for a third time, we assume that the following limit exists:

$$T = \lim_{t \to \infty} T_t \tag{7}$$

Little's Theorem Proof assuming FIFO

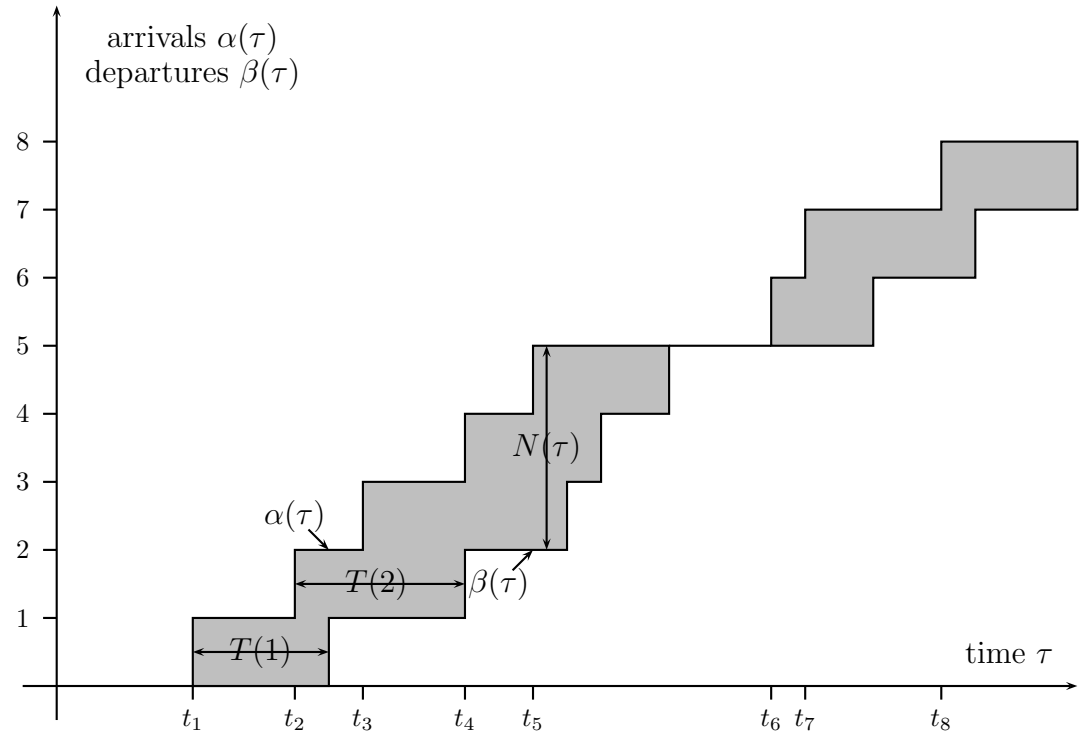

Figure 1: Little's Theorem in a FIFO System

At first, we will prove Little's Theorem with the restrictive conditions that, in addition to the limits above existing, the queue is empty at time 0 ($N(0) = 0$), that queuing is FIFO and that the queue becomes empty infinitely often beyond any given time τ. Figure 1 shows a FIFO queue which is initially empty.

We note that at any time τ:

$$N(\tau) = \alpha(\tau) - \beta(\tau) \tag{8}$$

It is clear that if we choose a time t when the system again becomes empty then we can calculate the area of the shaded area $A(t)$:

$$A(t) = \int_0^t N(\tau) d\tau \tag{9}$$

However, equally, we can consider the shaded area to be composed of horizontal strips of height 1 and width $T(i)$ (for the ith customer). In this case, we have:

$$A(t) = \sum_{i=1}^{\alpha(t)} T(i) \tag{10}$$

Setting these equations equal and dividing each side by t gives us:

$$\frac{1}{t}\int_0^t N(\tau)d\tau = \frac{1}{t}\sum_{i=1}^{\alpha(t)} T(i) = \frac{\alpha(t)}{t}\frac{\sum_{i=1}^{\alpha(t)} T(i)}{\alpha(t)} \tag{11}$$

Therefore we have:

$$N_t = \lambda_t T_t \tag{12}$$

which, if we take the limit as $(t \to \infty)$ becomes Little's Theorem as required. Note that some of the assumptions made here are, in fact, unnecessary as we shall see. In fact, Little's Theorem only requires the following:

1. The limit $\lambda = \lim_{t\to\infty} \alpha(t)/t$ exists
2. The limit $\delta = \lim_{t\to\infty} \beta(t)/t$ exists
3. The limit $T = \lim_{t\to\infty} T_t$ exists
4. $\delta = \lambda$

Little's Theorem Proof not assuming FIFO

Figure 2 shows the situation if we don't assume FIFO. Now, it is still clear that the shaded area in the interval $[0, t]$ is given by:

$$A(t) = \int_0^t = N(\tau)d\tau \tag{13}$$

even though, in the diagram, $N(\tau)$ is no longer necessarily a continuous vertical slice of the area (consider, for example, the situation at time t_4 in the diagram).

Now, we define the following:

$D(t)$ is the set of customers who have departed the system by time t.

$\overline{D}(t)$ is the set of customers who are still in the system at time t.

The delay experienced up to time t by a customer still in the system at time t is $t - t_i$. Therefore we can say:

$$A(t) = \sum_{i\in D(t)} T(i) + \sum_{i\in\overline{D}(t)} (t - t_i) \tag{14}$$

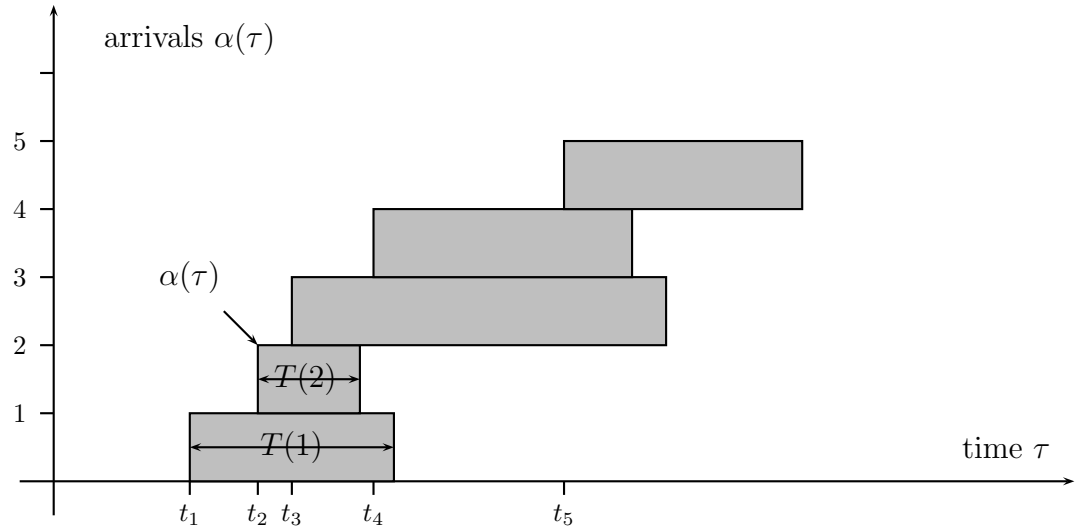

Figure 2: Little's Theorem in a non FIFO System

Therefore, equating these and dividing by t we get

$$\frac{1}{t}\int_0^t N(\tau)d\tau = \frac{1}{t}\sum_{i\in D(t)} T(i) + \frac{1}{t}\sum_{i\in \overline{D}(t)} (t - t_i) \tag{15}$$

Up to time t, the average arrival rate λ_t is given by:

$$\lambda_t = \frac{|D(t) + \overline{D}(t)|}{t} \tag{16}$$

Up to time t, the average waiting time T_t is given by the total waiting time of all the customers so far over the total number of customers entering the system so far (up to time t):

$$T_t = \frac{\sum_{i\in D(t)} T(i) + \sum_{i\in \overline{D}(t)} (t - t_i)}{|D(t) + \overline{D}(t)|} \tag{17}$$

Substituting these two equations and our previous equation (2) into (15) we therefore have:

$$N_t = \lambda_t T_t \tag{18}$$

which, in the limit as $t \to \infty$ gives us Little's Theorem again.

Probabalistic Form of Little's Theorem

Throughout this section we will assume that all relevant systems are ergodic. Let us define that at a time t the probability that there are exactly n customers in the system is $p_n(t)$. (Typically, in a probabalistic situation we would be given, or assume, the starting conditions — the distribution given by $p_n(0)$ for all n). From this, we can say that the mean number of customers in the system at time t is given by:

$$\overline{N(t)} = \sum_{n=0}^{\infty} n p_n(t) \tag{19}$$

Note that $\overline{N(t)}$ and $p_n(t)$ depend on both t and the initial distribution of $\mathbf{p}(0) = \{p_0(0), p_1(0), p_2(0) \dots\}$. If we assume that the system converges to some steady state distribution then we have:

$$\lim_{t\to\infty} p_n(t) = p_n \quad n = 0, 1, \dots \tag{20}$$

And therefore we can say that the average number of customers in the system is:

$$\overline{N} = \sum_{n=0}^{\infty} n p_n \tag{21}$$

Now, if we assume (as before) that this converges to a limit, we have:

$$\overline{N} = \lim_{t\to\infty} \overline{N(t)} \tag{22}$$

Similarly, let us assume that the average delay for the kth customer $\overline{T(k)}$ also converges as $k \to \infty$ to a steady state value:

$$\overline{T} = \lim_{k\to\infty} \overline{T(k)} \tag{23}$$

Now, since we assumed ergodicity, we can say that the time average of the number of customers N is (with probability 1) equal to the steady state state-space average $\overline{N}$:

$$N = \lim_{t\to\infty} N_t = \lim_{t\to\infty} \overline{N}(t) = \overline{N} \tag{24}$$

Similarly, we can say that the time average of customer delay T is equal to the steady state state-space average $\overline{T}$.

$$T = \lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} T(i) = \lim_{k\to\infty} \overline{T}(k) = \overline{T} \tag{25}$$

Little's theorem then holds with λ given by:

$$\lambda = \lim_{t\to\infty} \frac{E\{\text{Arrivals in interval } [0,t]\}}{t} \tag{26}$$

A Simple Example Using Little's Theorem

Take, as an example, the very simple model of an ethernet system shown in figure . Assume that the ethernet can (as is often the case) only transmit a single packet at a time (other packets to be sent must queue).

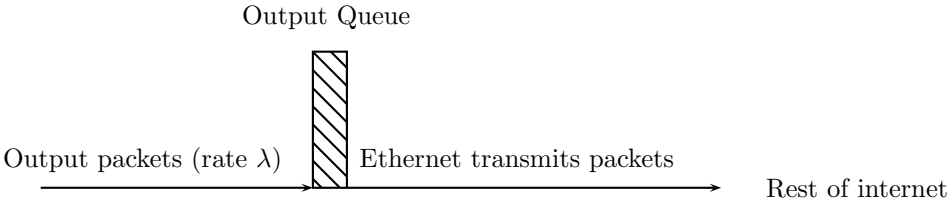

Figure 3: A very simple view of an ethernet system.

Note that we can apply Little's theorem twice even in this simple situation. First, if N_Q is the average number of packets queuing to leave the system and W is the average time a packet spends in the queue waiting to be transmitted then we have:

$$N_Q = \lambda W$$

Now, assuming no packets are ever dropped at this point (this would typically be the case unless the system is misconfigured) and the mean transmission time taken for the packet along this section of ethernet is $\overline{X}$ then we have:

$$\rho = \lambda \overline{X}$$

where ρ is the average number of packets on the outgoing ethernet section. Since, this must be either zero or one at any given time, ρ is also the *utilisation factor* — the proportion of the time which we expect the line to be busy.

Another Simple Little's Theorem Example

Consider a windowed flow control system such as TCP. At any given time, at most W packets (and ACKs) can be outstanding where W is the window size. Let us assume now that W has reached some constant maximum (in TCP, users advertise a maximum window size) W_m (this is the maximum size the window will reach if no packets are lost). Since the number of packets under transmission is less than or equal to W_m we can say that if packets depart at an average rate λ and, on average, have a round trip time of T then

$$W_m \geq \lambda T$$

This gives us an insight into the flow-control behaviour of TCP since, it is obvious from this equation that if delay increases (T goes up) then λ the sending rate will come down (or the window size will be decreased).

If the system is congested then λ cannot increase (the system is transmitting at its maximum rate). If no packets are lost then the window size W will become W_m and

$$W_m \approx \lambda T$$

Therefore, in a congested state, if packet loss does not yet occur, increasing the window size simply serves to increase delay with no advantages to throughput which shows the reason for setting a maximum window size.