Computer Networks
A gentle introduction to queuing theory

Saad Mneimneh
Computer Science
Hunter College of CUNY
New York

So how little is Little's theorem?

1 Introduction

Before we address other aspects of TCP or start discussing the network layer, we consider some theoretical treatment of the network as a system that provides service to customers. In this system, customers arrive at random times to obtain service.

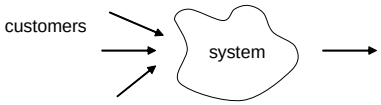

Figure 1: Customers arrive to the system for service

In the context of computer networks, customers represent packets (or frames, and we shall not make the distinction here), and service represents the assignment of packets to communication links. For instance, if a link (server) has a bandwidth of B bps (service rate), then the service time for a packet (customer) assigned to that link will be L/B, where L is the size of the packet in bits (including headers if frame).

Because the number of servers is usually finite, customers of this system are often modeled to be waiting for service in queues. Queuing theory is the field responsible for the study of such systems. The theory will help us gain some insight about buffer space, packet delays, and network utilization. This in turn could help us in the design of switching strategies (network layer) and congestion control mechanisms (e.g. TCP).

We are interested in answering questions like:

- What is the average number of customers in the system? (i.e. the "typical" number waiting in queue or undergoing service)
- What is the average delay per customer? (i.e. the "typical" time a customer waits in queue plus the service time)

These quantities are often obtained in terms of known information such as:

- The customer arrival rate (i.e. the "typical" number of customers entering the system per unit time)
- The customer service rate (i.e. the "typical" number of customers the system serves per unit time when it is constantly busy)

2 Preliminaries

Let us work out what we mean by average or "typical". We start with few definitions:

$$N(t) = \text{ Number of customers in the system at time } t$$

$$A(t) = \text{ Number of customers who arrive in the interval } [0, t]$$

$$T_i = \text{ Time spent in the system by the } i^{th} \text{ arriving customer}$$

A notion of "typical" number of customers observed up to time t is the time average

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

In many systems of interest, N_t converges to a steady state

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

Similarly, we define $\lambda_t = \frac{A(t)}{t}$, and the time average arrival rate $\lambda = \lim_{t \to \infty} \lambda_t$ (assuming the limit exists). We also define

$$T_t = \frac{\sum_{i=1}^{A(t)} T_i}{A(t)}$$

and the time average customer delay (assuming the limit exists)

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

It turns out that the quantities N, λ, and T above are related by a simple formula that makes it possible to determine one given the others. This result was proved by John Little at MIT in 1961, and is known as Little's theorem. Before that, the result was a "folk theorem" in operations research for many years.

3 Little's theorem

Little's theorem expresses the natural idea that crowded systems are associated with long delays. It has the following form:

$$N = \lambda T$$

What is remarkable about this theorem is that it applies to any system, regardless of the arrivals and what the system looks like inside. For instance, on a rainy day, traffic on a rush hour moves slower than average (large T), and the streets are more crowded (large N). Similarly, a fast food restaurant (fast service, small T) needs a smaller waiting room (e.g. drive through, small N) than a regular restaurant for the same customer arrival rate.

We will prove Little's theorem graphically under some simplifying assumptions (we relax those assumptions later on).

- assumption 1: the system is initially empty, i.e. $N(0) = 0$
- assumption 2: the system is FIFO
- assumption 3: the system becomes empty infinitely many times

Let $D(t)$ be the number of customers that depart the system in the interval $[0, t]$. $A(t)$ and $D(t)$ trace the arrivals and departures by time t respectively as shown in Figure 2 below:

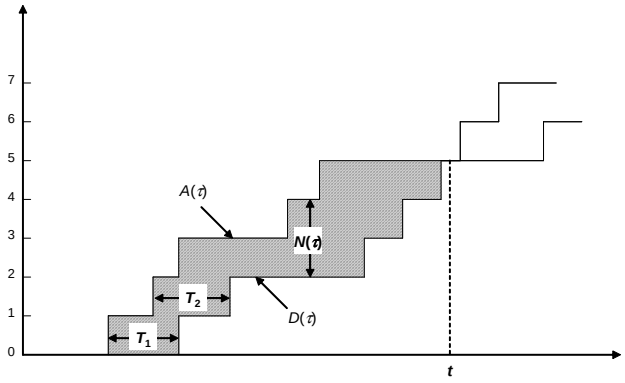

Figure 2: Graphical proof of Little's theorem

The system is empty at t implies that $\int_0^t N(\tau)d\tau = \sum_{i=1}^{A(t)} T_i = \frac{A(t)\sum_{i=1}^{A(t)} T_i}{A(t)}$. Dividing by t, we get $N_t = \lambda_t T_t$. Taking the limit as t goes to infinity (assuming λ and T exist), we get $N = \lambda T$.

The first assumption is not really necessary. In fact, the same argument works as long as we adopt the convention that, for customers initially in the system, the time T_i is counted starting at $t = 0$.

We now relax all assumptions. We only require that $\lambda = \lim_{t\to\infty} \frac{A(t)}{t}$ exists and $T = \lim_{t\to\infty} \frac{\sum_{i=1}^{A(t)} T_i}{A(t)}$ exists. Let t_i be the time of arrival of customer i. Then using the same graphical argument above:

$$\sum_{i:t_i+T_i\le t} T_i \le \int_0^t N(\tau)d\tau \le \sum_{i:t_i\le t} T_i$$

$$\sum_{i:t_i+T_i\le t} T_i \le \int_0^t N(\tau)d\tau \le \sum_{i=1}^{A(t)} T_i$$

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i:t_i+T_i\le t} T_i \le \lim_{t\to\infty} \frac{1}{t} \int_0^t N(\tau)d\tau \le \lim_{t\to\infty} \frac{A(t)}{t} \frac{\sum_{i=1}^{A(t)} T_i}{A(t)}$$

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i:t_i+T_i\le t} T_i \le \lim_{t\to\infty} N_t \le \lambda T$$

Therefore, to prove the result, we only have to show that:

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i:t_i+T_i\le t} T_i = \lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{A(t)} T_i$$

We first show that $\lim_{n\to\infty} T_n/t_n = 0$. The intuition behind this approach is that if T_n becomes very small compared to t_n, the average delay for customers who leave the system by time t (left hand side above) becomes very close to the average delay for customers who arrive to the system by time t.

$$\frac{T_n}{t_n} = \frac{T_n}{A(t_n)} \frac{A(t_n)}{t_n}$$

Therefore, $\lim_{n\to\infty} \frac{T_n}{t_n} = \lambda \lim_{n\to\infty} \frac{T_n}{A(t_n)}$.

$$\frac{T_n}{A(t_n)} = \frac{\sum_{i=1}^{n} T_i}{A(t_n)} - \frac{A(t_{n-1})}{A(t_n)} \frac{\sum_{i=1}^{n-1} T_i}{A(t_{n-1})}$$

Note that $A(t_n) = n$; therefore,

$$\frac{T_n}{A(t_n)} = \frac{\sum_{i=1}^{A(t_n)} T_i}{A(t_n)} - \frac{n-1}{n} \frac{\sum_{i=1}^{A(t_{n-1})} T_i}{A(t_{n-1})}$$

$$\lim_{n\to\infty} \frac{T_n}{A(t_n)} = T - 1 \cdot T = 0$$

Therefore, $\lim_{n\to\infty} T_n/t_n = 0$. This means that for any $\epsilon > 0$, there exists a finite time τ such that $T_i < t_i\epsilon$ for all $t_i > \tau$. Now we finish the proof. Consider a time $t > \tau$:

$$\sum_{i:t_i+T_i\le t} T_i = \sum_{i:t_i\le\tau, t_i+T_i\le t} T_i + \sum_{i:t_i>\tau, t_i+T_i\le t} T_i$$

Since $t_i > \tau \Rightarrow T_i < t_i\epsilon$, then $t_i > \tau$ also means that $t_i(1+\epsilon) \le t \Rightarrow t_i + T \le t$. Therefore,

$$\begin{aligned}\sum_{i:t_i+T_i\le t} T_i &\ge \sum_{i:t_i\le\tau,t_i+T_i\le t} T_i + \sum_{i:t_i>\tau,t_i(1+\epsilon)\le t} T_i \\ &= \sum_{i:t_i\le\tau,t_i+T_i\le t} T_i - \sum_{i:t_i\le\tau,t_i(1+\epsilon)\le t} T_i + \sum_{i:t_i\le t/(1+\epsilon)} T_i \\ &= \sum_{i:t_i\le\tau,t_i+T_i\le t} T_i - \sum_{i:t_i\le\tau,t_i(1+\epsilon)\le t} T_i + \sum_{i=1}^{A(t/(1+\epsilon))} T_i\end{aligned}$$

The first two terms are finite. Dividing by t and taking the limit as $t \to \infty$, we get

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i:t_i+T_i\le t} T_i \ge \frac{1}{1+\epsilon} \lim_{t\to\infty} \frac{A(t/(1+\epsilon))}{t/(1+\epsilon)} \frac{\sum_{i=1}^{A(t/(1+\epsilon))} T_i}{A(t/(1+\epsilon))}$$

But $\frac{1}{t}\sum_{i:t_i+T_i\le t} T_i \le \frac{1}{t}\sum_{i:t_i\le t} T_i$; therefore,

$$\frac{1}{1+\epsilon}\lambda T \le \lim_{t\to\infty} \frac{1}{t} \sum_{i:t_i+T_i\le t} T_i \le \lambda T$$

Since ϵ was arbitrary, we can take the limit as $\epsilon \to 0$ to obtain the result.

4 Probabilistic Little

In the above analysis, we relied on a single sample and computed averages over time. For almost every system of interest, the same applies if we replace time averages with ensemble average, i.e. under a probabilistic setting.

- N can be replaced by $\bar{N} = E[N]$, the expected number of customers in the system
- T can be replaced by $\bar{T} = E[T]$, the expected time spent by one customer
- λ can be replaced by $\lim_{t\to\infty} \frac{\text{expected number of arrivals in } [0,t]}{t}$

Therefore,

$$E[N] = \lambda E[T]$$

Usually, λ is given as a property of arrival process (we will see this when we model arrivals), and $E[N]$ can be computed by a simple analysis of p_n, the probability of having n customers in the system (later).

5 Little Examples

Little's theorem can be applied to any system or even parts of it. The following two examples illustrate the idea:

5.1 Example 1

Consider the following node where the arrival rate is λ packets per second and the link bandwidth is μ bps:

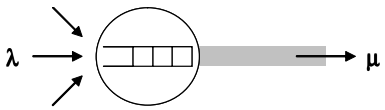

Figure 3: Simple system

- Looking at the node: $N = \lambda T$, where N is the average number of packets in the node and T is the average delay per packet
- Looking at the queue: $N_Q = \lambda W$, where N_Q is the number of packets in the queue and W is the average waiting time per packet
- Looking at the transmitter: $\rho = \lambda \frac{L}{\mu}$, where:
  - ρ is the average number of packets being transmitted (served), also known as link utilization, efficiency, or throughput
  - L/μ is the average transmission time per packet (L is the average packet size)
  - note that $N = N_Q + \rho$
- Looking at the link: $B = \lambda D$, where B is the number of packets in transit and D is the propagation delay of the link

5.2 Example 2

Consider the following system of nodes:

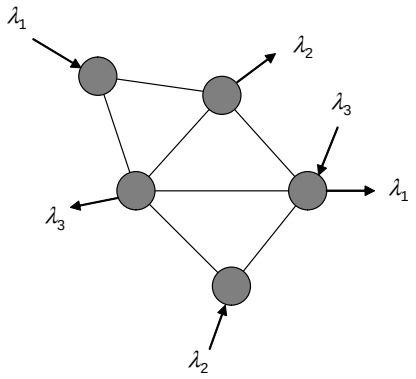

Figure 4: Combination of systems

- For each subsystem, $N_i = \lambda_i T_i$
- For the whole system, $N = \lambda T$, where $N = \sum_i N_i$ and $\lambda = \sum_i \lambda_i$
- Therefore

$$T = \frac{\sum_i \lambda_i T_i}{\sum_i \lambda_i}$$

(average weighted by $\lambda_i's$)

References

Dimitri Bertsekas and Robert Gallager, Data Networks
Shaler Stidham, A Last Note on $L = \lambda W$, JSTOR 1972.