CptS/EE 555
Basic Queuing Theory
Oct 29 & 31, 2001

The purpose of this section is for students to become familiar with the concepts of basic queuing theory, including Little's Theorem and the behavior of simple queuing models such as M/M/1 and M/M/n/m, including the ability to apply the concepts to analyze or predict the behavior of networks.

Background

Recall

Total Delay = Transmission time + Propagation Delay + Queuing Delay + Processing Time

We have discussed transmission time and propagation delay previously, and have observed that in many situations processing time is a minor contributor. We now want to get some further understanding of queuing delay.

In contrast to the simple characterizations of transmission time and propagation delay, queuing delay is considerably more complex. Indeed, queuing is a subject a large body of research and the models developed are often at best approximate in actual systems. Nevertheless, queuing theory can provide valuable quantitative and qualitative insights into the behavior of real systems.

Mental experiment: consider a link with capacity C bits/sec and a demand D bits/sec. If D is always less than C, no queue forms. If D is always greater than C, as time passes packets will wait longer and longer to use the link – there is no bound on how long they will wait. The interesting case is when D is on average less than C, but sometimes greater. In that case a queue will form and we'd like to know something about how long it is at different times, how long packets must wait, etc.

Definitions

N(t) = the number of customers (packets, jobs, etc.) in the system at time t. Note that "in the system" means either waiting or being served.

$\alpha(t)$ = the number of customers (packets, jobs, etc.) that have arrived in the time interval (0,t)

T_i = the time spent in the system by the i^{th} customer

t_i = the arrival time in the system of the i^{th} customer

Note that $T_i + t_i$ is the departure time of the i^{th} customer

With those definitions in mind we can now calculate

$N_t = \frac{1}{t}\int_0^t N(\tau)d\tau$ = the average number of customers in the system during [0,t]

If N_t tends to a constant as t increases we can define $N = \lim_{t\to\infty} N_t$, and call it the steady state time average number of customers in the system.

Similarly we write

$\lambda_t = \dfrac{\alpha(t)}{t} =$ the average arrival rate over $[0,t]$

and $\lambda = \lim_{t\to\infty} \lambda_t =$ the steady state arrival rate.

Likewise,

$T_t = \dfrac{\sum_{i=0}^{\alpha(t)} T_i}{\alpha(t)} =$ the average customer delay up to time t, and

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

Little's Theorem

Little's Theorem is simply $N = \lambda T$. That is the average number of customers (packets, jobs, etc.) in the system is the product of the average arrival rate and the average time each packet (customer, job, etc.) spends in the system.
Proof: omitted – see your in-class notes.

Application 1

Suppose that a point-to-point link has an arrival rate of λ, that there is a queue ahead of the transmitter for the link, that packets on average spend W_Q seconds in the queue and W_X seconds being transmitted. Then the average number in the queue will be $N_Q = \lambda W_Q$, and the average number being transmitted is $N_X = \lambda W_X$. Note that at any given time a point-to-point link can have at most 1 packet being transmitted, so $0 \le N_X \le 1$ and we could interpret N_X as the utilization of the link.

Additional notes

In class I defined a couple of additional functions that I used in the proof of Little's theorem. For completeness here are those definitions:

$D(t) =$ the set of customers who have left (Departed) the system at time t.

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

$\left|\overline{D}(t)\right| = N(t)$

$\beta(t) = \left|D(t)\right| =$ the number of customer who have left the system before time t.