Performance Engineering Cookbook
Ingredients for Performance and Capacity Success

7. Little's Law: How Long is the Wait

Peter van Eijk

This is a series of brief articles explaining the basic concepts of systems performance and capacity planning. Motivated by the Computer Measurement Group, these concepts are applicable to IT systems and beyond.

Little's law and measuring queues

If there is one thing that you need to know when measuring and predicting queues, it is Little's Law, which relates average arrival rate, waiting time and queue length.

The peculiar thing about Little's law is that it is trivially obvious to a lot of people and unintelligible to others. That might be because it is almost universally true.

First, let's look at the arrival rate, the speed at which 'customers' arrive. Little's law is applicable if the system is *stable*, meaning the arrival rate is equal to the departure rate, and every customer gets served eventually. Starting and stopping a system are typically not stable. Little's law applies no matter what the distribution of the arrival times is. It could be regular, random, Poisson or whatever.

Then the waiting time: this is the time between arrival and departure. Beware: we typically have a queue in front of a server. The waiting time includes the serving time so it is not just the time spent in the queue.

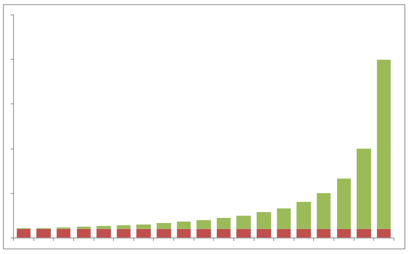
Finally the queue length. Again don't forget to add the customers that are already being served. Little's law is also valid when there are queue jumpers or priority queues.

And this leads us to

> Little's law: the average queue length equals the average arrival rate times the average waiting time.

Alternatively: the average waiting time equals the average queue length times the average time between two arrivals.

So whenever you are standing in a customer line, Little's law allows you to estimate how long you still have to wait. Estimate how many people are in front of you, estimate how long it takes for each customer to be served (more accurately, you would estimate the inverse of the departure rate), and multiply these.



The graph represents how the total time spent increases as the arrival rate increases to approach the maximum server capacity (i.e. the maximum departure rate).

Link farm

Wikipedia: http://en.wikipedia.org/wiki/Little's_law
Measureit: Again http://www.cmg.org/measureit/issues/mit99/m_99_3.pdf by Bob Wescott, gives some useful applications.