Performance Engineering Cookbook
Ingredients for Performance and Capacity Success

Peter van Eijk

6. Queues: waiting in line

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.

Queues

What happens in your local coffee shop if more customers arrive then there are staff members? They have to line up and wait, and unless they are impolite, they will do so.

What happens when the highway is very crowded, and you want to get on it?

What happens when you request a page from a web server?


If we abstract away from the differences, these three examples have a few things in common. There are users of a service. They are the customers of a coffee shop, the cars on the road, and the requests that your browser sends to the web server. In our business we call these 'arrivals'. Then there is handling those arrivals: the staff members, the highway and the webserver. These we call 'servers'. Finally, there is the number of arrivals that are not being served yet, which we often call the queue length.

I have started this series with the topics 'response time' and 'capacity'. Queuing theory gives us the tools to link these two. As the utilization of capacity goes up, so goes the response time. How much exactly? That depends on the characteristics of the arrival traffic, the number of servers, and a ton of other things. That is what the queuing theory is all about.

Of course, no real system is best modeled as a single queue, and accurately modeling the performance can be a bit tricky. For example, in the highway traffic case, the capacity is not constant but decreases at high utilization. This leads to congestion, but that is the topic of another issue in this series.

Link farm

Wikipedia: http://en.wikipedia.org/wiki/Queueing_theory which is mostly a list of definitions and pointers to theory.
Measureit: http://www.cmg.org/measureit/issues/mit99/m_99_3.pdf by Bob Wescott, which gives an introduction in quantitative queue modeling and analysis.