Performance Engineering Cookbook
Ingredients for Performance and Capacity Success

Response Time – How Fast Are You Being Served?

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

Response Time

Arguably, response time is the single most important measure of how IT systems perform. A lot of performance engineering and capacity planning is focused on keeping response times optimal. We will get into more detail later when we dive into the elements that drive response time.

According to ITIL, response time is "A measure of the time taken to complete an Operation or Transaction" (capitalized words here refer to other concepts for which an ITIL definition exists).

So, for example, response time is the time that elapses between submission of a batch or print job and its completion, or between clicking on a hyperlink and the appearance of the webpage. Response time is measured in seconds (or scales of it: milliseconds, hours).

Response times are often used as a quality indicator in the delivery of a service, and appear in service level agreements between a service provider and its customers.

However, defining response time is sometimes elusive, and misunderstandings arise. The main causes for that are confusion about the beginning and the ending of the operation or transaction. For example, in incident response, response time typically refers to the time between when a call is made and the point where incident management actually begins working on the incident, rather than complete its work. And when does the operation begin? When the request is submitted, or when the request is received by the service provider? And when is a webpage done? When the browser is finished loading, or when the user can act upon the information in the page?

What this hints at is that what looks like an atomic transaction often breaks down into sub transactions. In the incident response case, the transaction as perceived by the user as resolution of his or her problem. The sub transactions of that include: determining the appropriate incident responder, initiating contact with the incident responder, the time the incident responder takes to get started and to resolve the incident, and then 'mopping up' after that. The real end point for the users is when they can get back to their normal work.

Suggested Links:

Introductory:
http://en.wikipedia.org/wiki/Response_time_(technology)
http://www.webperformancetoday.com/2012/02/13/non-geeky-guide-to-performance-measurement/
Advanced:
http://www.cmg.org/measureit/issues/mit62/m_62_15.html (Dr. Neil Gunther)

Note to readers: Are there any concepts here that need further elaboration (i.e. service provider)? Volunteers wanted to find more link worthy pages in sources such as the CMG archives, Wikipedia,

and for linking back from Wikipedia to these pages. Please write to the author: Dr. Peter HJ van Eijk at pveijk@nlcmg.nl.