Performance Engineering Cookbook
Ingredients for Performance and Capacity Success

Availability – At your service

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.

Availability

When is a service available to its users, or consumers? At the face of it, this seems a simple question. But there are some complications.

According to ITIL, availability is the "Ability of a Configuration Item or IT Service to perform its agreed Function when required. Availability is determined by Reliability, Maintainability, Serviceability, Performance and Security. Availability is usually calculated as a percentage. This calculation is often based on Agreed Service Time and Downtime. It is Best Practice to calculate Availability using measurements of the Business output of the IT Service."

So, availability (or uptime) is the chance that a service is actually working. But when is it working? This is tied to response time. What is the difference between a very slow service and a non operational service? This always reminds me of Monty Python's iconic "dead parrot sketch", in which the shopkeeper claims that the parrot's lack of response is due to it being asleep rather than dead.

A pragmatic definition of unavailability therefore is: too slow to be useful for the business.

Typical service level agreements specify 99.5% or 99.9% ('3 nines') as acceptable uptime. This is comparable to what is experienced on the internet http://www.slideshare.net/pveijk/cloud-encounters-sept-2009-for-cmg-dec-6 .

SLA penalty clauses should specify the period over which downtime is computed. A downtime period of 8 hours is still within 3 nines, measured over a year. You may not want your provider to get away with that.

At very high availabilities, downtime is indistinguishable from slowness: 6 nines equals a maximum of 3 seconds of downtime per month.

How to measure availability? Do we measure the success of real transactions or do we regularly attempt synthetic transactions? Measuring synthetic transactions is more consistent, but it tends to underreport perceived downtime. The busier the service is, the more likely it is to fail. A service may stay within SLA limits, even as it slows unacceptably at peak periods.

Link farm

Wikipedia: http://en.wikipedia.org/wiki/Availability (though much wider than IT).
http://en.wikipedia.org/wiki/Dead_Parrot_sketch

Note to readers: are there any concepts here that need further elaboration? We want volunteers 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.