The New, New Math of Cloud

Lori MacVittie, 2009-26-02

It's been a long time since I had the (mis)fortune to sit in a math class, so bear with me while I figure this out. In order to determine my daily budget for the application I am hosting with Google's App Engine, I need to sum the results of the standard deviation of the derivative of yesterday's CPU utilization, multiplied by the bandwidth used divided by *pi* and then multiple the whole thing by the number of e-mail messages sent by the application. Got that? Go.

5…4…3…2…1 Pencils down. What? Not finished yet?

Okay, I'm being a bit (okay, maybe a lot) facetious, but the combination of variables necessary to determine what my daily budget should be in order to exceed my quota on Google's App Engine seems a bit overwhelming. Google's decision, driven by customer demand, to allow organizations and individuals to exceed the free quotas (thresholds) set for applications hosted on App Engine is a good one, but it may require a bit of work on the part of the application owner to figure out just what the "maximum amount you're willing to pay for computing resources each day" really is. The variables involved are many, it's not just a matter of determining the rate of user growth and comparing it against one or even two resource usage statistics.
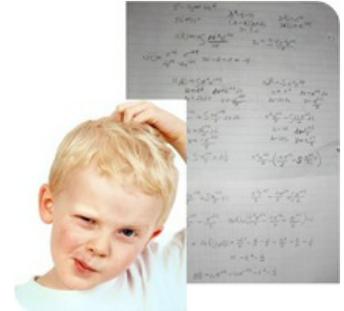
Sort of like trying to help a teenager with her math homework these days. It's *different*, it's *new* math. I'm not even sure where to start. For those of you who didn't see the announcement, here's the relevant excerpt:

> *You allocate this budget across CPU, bandwidth, storage, and email, and you pay for only what your app consumes beyond the free thresholds -- prorated up to the nearest penny.*

Capacity planning has always been more art than science, but there has been science in it, at least, and didn't require a PhD in calculus.

One of the hardest things to wrap your head around will be the fact that it's nearly impossible to determine what the concurrent user and request capacity of your application may be based on those free thresholds because, well, you probably haven't been able to load test it.

Google and other cloud providers can – and most do – offer up statistics on resource use (how else could the bill/track/manage those variables if they didn't have them available), but they're usually based on application, not necessarily *per request* or *per user* so capacity planning at this granularity may require some heavy math and good guestimates. So what *is* the maximum amount you're willing to pay on a daily basis given those parameters? And how much goes where? What if you don't allocate enough storage and over provision the CPU? Or vice-versa?

Feels like, if you ask me, bidding on e-Bay. It's a gamble, calculated to be sure, but a gamble nonetheless.

Remember, the cloud changes everything, including how we budget projects. It's likely that even most IT folks can't accurately predict the usage of these resources on a per request or per user basis across the applications which they *have* the ability to load test, let alone one in which they don't.

This is one of the reasons when we talk about dynamic infrastructure that we necessarily include applications in the "feedback loop". Applications know how much storage they need, they know how many e-mail messages they're likely to send. Sure, you can correlate all that information if you have the right management systems, but it's more likely that such correlation systems are still to be built and you'll have to find another way to figure out what the variables involved in computing a daily budget will be for any cloud-based application. It is this connectivity intelligence that enables collaboration in a dynamic infrastructure that will better enable administrators and other stakeholders to more accurately understand the complex web of resource use in a virtual or cloud-based environment.

We're going to have to learn to change how we think about capacity, and we may have to learn some new, new math to do it.