A common problem I see people running into when using a cloud computing service is the trap of under-provisioning. There’s a chain effect that leads to this result: 1) people don’t understand how virtualization works, and therefore 2) they don’t realize how much of a computing resource they’re really buying, so 3) they assume they are entitled to more than they really are, and 4) they under-provision. A few other causes and effects come into play here, too. For example, the choice to use the cloud is sometimes founded on economic assumptions that frequently turn out to be wrong. The cloud service looks more economically attractive than it really is, due to under-provisioning. Let’s get back to this idea that people under-provision. How do I know that’s happening? I’ll use anecdotal evidence to illustrate. Here’s a real quote from a recent engagement about database (MySQL) performance problems: Do you think it’s likely that the underlying hardware is simply worse than average? If you think this will be an ongoing problem, maybe we should try our luck with a new instance/storage cluster? The fundamental assumption here is that some clusters are overloaded and are giving poor quality of service. We’re trained to think this way because we are familiar with services such as shared hosting, where other users on your particular server might really be abusive and claim resources that should be yours. But this isn’t how virtualization works in the common cloud platforms. In these platforms, you aren’t sharing resources with other users. You are guaranteed to get what you deserve! No kidding — this actually works. If that’s true, then why does performance fluctuate so much? The answer lies in how resources are parceled out. Assume there are 10 units of computing resources, and you’re paying for one of them. You buy 1/10th of the machine’s power. But it just happens that you’re the only virtual instance running on that physical server. You fire up an intense job. How much power do you get? You paid for 1 unit, but you get 10, because no one else is using the other 9 units. This is the way most virtualization platforms work: they give you extra resources if they’re available and not being claimed by anyone else’s instance. This guarantees that you’ll never get less than you deserve, but it leaves open the possibility that you’ll get more than you deserve. (What would be the point of wasting that power, really?) Under-provisioning is the obverse of over-providing, which is what the virtualization platform does. First-generation hyperthreading gave the same illusion of more resources than are really available, by the way. It made you think there were multiple processors, when in fact there weren’t — there were multiple sets of registers. Hyperthreading is a form of virtualization, too. What typically happens is that people are running their cloud instances on machines whose underlying physical hardware is not fully utilized, and they get used to a certain level of performance they’re not really paying for. Alas, you can’t really know whether this is happening or not! But it surely is in many (most?) cases, which is why occasionally you get some resource that seems much slower than you’re accustomed to, and you think it’s “too slow.” Not so. Your other units are “too fast.” I have a theory that if you really knew the true capacity you were buying, you’d view the price-to-performance ratio much less favorably. But it’s almost impossible to know that, really; it doesn’t help that the cloud service providers are rather vague about how much power a certain instance size really gives you. (They aren’t being malicious; it’s just the way virtualization works.) Under-provisioning is almost forced on users because they have no alternative — you could plan for worst-case performance, and you’d be doing the right thing, but how will you ever know you’ve really hit rock bottom and the worst case is really no worse? How can you even benchmark and do proper capacity planning, if you don’t know what you’re benchmarking? This should really give you serious pause. You should be thinking “wait, I’m basing my capacity planning and provisioning on luck and the law of large numbers. What if my luck runs out and I get a Black
Swan event?” The question is not “what if,” but “when.” I also think that the lack of transparency encourages people to use cloud computing services for the wrong reasons altogether. I could write about this, but I think Theo Schlossnagle said it pretty well already.

Related posts:
- Drizzle stops the rain
- Sessions of interest at the Percona Performance Conference
- Why MySQL might not benefit from having a mother ship
- I moved this blog to pairLite with zero downtime, and it was easy
- How good is the new High Performance MySQL going to be?

Leggi: Planet MySQL