WPO: A New Wave of Performance Engineering?

Alex Podelko

Reading Steve Thair's post *Who should be on your WPO team?* (http://www.seriticonsulting.com/blog/) reminded me of so many things I read before like Connie Smith and Lloyd Williams' papers about creating a Software Performance Engineering teams, or Mercury/HP white papers about creating Performance Centers of Excellence.

It looks like we have a completely new area of performance engineering – Web Performance Optimization (WPO) -- with its own terminology, approaches, experts, tools, Web Performance meeting groups, and the Velocity conference. WPO actually was around for a while (looks like the first Velocity conference was in 2008), but only recently, after attending a couple of New York Web Performance events, I realized that it became a separate discipline. I guess that the appearance of this new movement concentrated on the web performance means that we get a pretty mature industry of very scalable web sites delivering sophisticated content.

Well, the history of performance engineering looks like a series of waves (at least to me). The Computer Measurement Group (CMG) was organized in 1975 as an organization of performance analysts and capacity planners. Dr. Connie Smith's book "Performance Engineering of Software Systems" was published in 1990 created the Software Performance Engineering movement.

Distributed systems brought new wave of performance engineering based around load testing. Perhaps because there was not much instrumentation available and only way to make sure that the system performs was to apply load. It looks like the first version of LoadRunner was shipped in 1989. But the first time I got involved into load testing in 1997 with SQL Bench (SilkPerformer's ancestor), it was still far from what we expect from load testing tools now. The latest wave was probably Application Performance Management with a large array of tools promising visibility in what is going on inside applications.

It is interesting that all these overlapping areas never completely merged. This is probably the reason why we have such discrepancy in performance terminology because every group often started their terminology from a scratch (while others still used old terminology).

And now we get WPO. Steve Souders, who coined the term, said in his recent interview: *For years when developers started focusing on the performance of their websites, they would start on the back end, optimizing C++ code or database queries. Then we discovered that about 10% or 20% of the overall page load time was spent on the back end. So if you cut that in half, you only improve things 5%, maybe 10%. In many cases, you can reduce the back end time to zero and most users won't notice. So really, improvement comes from the time spent on the front end, on the network transferring resources and in the browser pulling in those resources.*

While WPO looks like a separate discipline, I'd rather placed it as a part of overall performance engineering. You still have a back end in most cases – and while the back end is mentioned in the WPO presentations, it sometimes looks like authors mention it as something trivial. Well, it is not, even for most web sites, not to mention large banks and insurance companies with many tiers of sophisticated systems in the back – and for the end-user performance you need to consider it all together. Downplaying “back end” is probably as wrong as downplaying “front end” (which, working mostly with business applications, I am definitely guilty of this - well, historically load testing concentrated on the server performance). The importance of each component depends on the system. In my opinion, performance principles are much more generic that the details of specific technologies. Most of performance engineering experience may be applied to any technology (you, of course, still need to learn something about this new technology too).

If we see that "about 10% or 20% of the overall page load time was spent on the back end" under the maximal load, this statement is a great example of applying performance engineering to the problem analysis. It is definitely the first thing to do investigating any performance issue – find where time is spent. And, considering popularity of WPO, it is probably the case for most modern websites with rich web interface and no need of transactional processing behind the scene. But it is usually not the case for sophisticated business applications working with transactional data (although I even doubt that it is exactly the case for the moment you click on "confirm order" button when you buy something on the Internet).

So, while it is very promising and exciting that we get a new wave of people dedicated to performance, it is a little sad that it looks like it often gets started from a scratch inventing new terminology and ignoring what existed before. For me it would be better if we get all these waves together to enrich each other with the area of performance engineering they specialize in. Of course, there is some interaction – well, you need to work together in a way to ensure systems’ performance – but it still looks like every wave tends to stay somewhat separate, cultivating their own terminology, approaches, and events.

It is my personal view, of course. If you see it differently, write it up and submit it to Measure IT.