Brendan Gregg's Blog home

The Benchmark Paradox

03 May 2014

Benchmarks are often used for product evaluations, and they are often so inaccurate one may as well flip a coin. But it's even worse than that.

> If your product's chances of winning a benchmark are 50/50, you'll usually lose.

This seeming paradox can be explained by some simple probability...

Most of the benchmarks I've debugged were false or misleading for one reason or another. A long time ago I was explaining this to a sales person, whose prospects used benchmarks to evaluate his product, and he asked:

> "But it could be wrong either way, like a coin toss?"
> "Yes."
> "That gives me a 50/50 chance of winning!"

He was delighted. His logic was that if the benchmarks were usually wrong – producing random numbers for his product and his competitor's – then at least half the time they should be wrong, but in his favor. If he won half the benchmarks, he'd have great growth.

I was annoyed: his product *did* perform well, so he should have been winning more than 90%, not 50%. Reality was even worse: he didn't win 90%, 50%, or even 25%. It didn't make sense until I debugged some cases.

When buying a product based on performance, customers often want to be *really* sure it delivers. That can mean not running one benchmark, but several, and wanting the product to *win them all*.

One customer, who used a set of three benchmarks, found the product won two but lost the third. They'd only be happy if it won them all. Their technique was problematic, as is often the case, and basically became a coin toss. For each benchmark.

> Probability of winning all three benchmarks = 0.5 x 0.5 x 0.5 = 0.125 = 12.5%

The more benchmarks – with the requirement of winning them all – the worse the chances. And as I mentioned in my previous post about [compilers and benchmarks](), kitchen-sink benchmarks are now popular.

This failure mode wasn't always the case for the unlucky sales person, but it did happen at least a few times, and probably more that he didn't know of. This is a problem for anyone promising high performance, where a

single poor result can cast doubt on their claims. And when multiple benchmarks are conducted incorrectly, the chances of flipping tails becomes high.

Benchmark gambling is a sucker's game: play [active benchmarking](#) instead.

You can comment here, but I can't guarantee your comment will remain here forever: I might switch comment systems at some point (eg, if disqus add advertisements).