Cahners
MICROPROCESSOR REPORT
www.MPRonline.com
THE INSIDER'S GUIDE TO MICROPROCESSOR HARDWARE

THE EDITORIAL VIEW

BENCHMARKS ARE BUNK

By Keith Diefendorff {6/26/00-01}

The benchmark situation in the PC-microprocessor industry is a shambles. The existing ad hoc collection of benchmarks fails to shine much light on the important question of PC-processor performance. In fact, benchmarks seem to add more noise and confusion than anything else. The problem is twofold: most benchmarks don't do a very good job of measuring performance, and benchmarks are intentionally abused and misused by marketers to paint their desired picture of performance rather than the actual one.

Single-program benchmarks can sometimes give insight into a specific issue in a limited domain, but they often don't give much. Typical "floating-point" benchmarks, for example, are far more integer intensive than floating-point intensive, and they are usually more a test of the memory system and the compiler than of the processor's FPU. Benchmarks almost never measure just one thing, and they often don't measure what you think they're measuring.

Single-program benchmarks are even less useful for determining overall processor performance. Application-based system-level benchmarks fail because they don't simulate the way microprocessors get used in real systems. Measuring how fast a system rips through a script of "typical" user interactions on a variety of applications does not reflect the dynamics of the way systems get used by humans, nor does it evaluate the processor under the workloads that are the most critical to perceived performance. Moreover, these benchmarks are so overwhelmed by system features such as chip sets, memory size, and disk speed that the processor-performance picture can be badly distorted. In addition, such benchmarks rarely run cross-platform; thus, they offer no help at all in evaluating processors with different architectures, such as x86 versus PowerPC.

Using multiple-program benchmarks to compare overall performance is almost as hopeless. Anyone who has ever compiled a benchmark summary recognizes the futility of that exercise. Missing data, stale data, inconsistent data, conflicting data, different metrics, and incomparable data (from different system configurations) easily defeat even the cleverest spreadsheet jockey. Rather than gaining a good overall picture of performance, the user usually winds up confused and bewildered—or worse, misguided.

Benchmark suites consisting of component programs focused on specific areas give perhaps the best measure of overall performance, but even these are flawed. These suites are roundly criticized because they don't reflect typical usage patterns and workloads, and because there is no fair way to weight the relative importance of the various components. Even if there were, there is no agreed-upon or mathematically sound method of merging into an overall score components with widely different run times—which is what consumers want and need.

In the future, the situation could get worse. As systems continue their march toward lower cost, higher complexity, and a more media-rich environment, processor performance issues will change dramatically and rapidly. Benchmark

efforts are unlikely to keep pace. Already the benchmark community is far behind processors on multimedia issues. There are no good benchmarks, for example, that reveal the value of the various SIMD extensions that are now a part of all major processor architectures. And benchmarks that measure power consumption (for estimating battery life and thermal loads) are simply abysmal.

In part, the problems with benchmarks stem from the fact that "performance"—whatever that means to you—is an inherently difficult thing to measure and quantify. But the larger issue is that the interests of processor vendors are not the same as those of consumers. Consumers want truth; vendors want an illusion. Since benchmarks are highly vulnerable to manipulation, misuse, and misinterpretation, the illusion is what consumers usually get.

Even though the efforts of benchmarking organizations such as ZDBOp and SPEC are honorable and well intentioned, the problems they face are bigger than the organizations are. Developing high-quality benchmarks and performing thorough, independent tests is a massive job that requires the cooperation of both processor and system vendors. There is simply no compelling business model capable of funding the level of effort required to mount a good benchmarking effort and to stay ahead of wealthy processor vendors gunning to defeat any benchmark that threatens to become popular.

Gaining the cooperation of processor and system vendors is unlikely. It is simply not in the best interests of these vendors for a powerful independent benchmarking organization to exist. Intel, for example, believes it can deliver the highest-performance microprocessors—*most* of the time. But it recognizes that in a few instances it will fail. In these cases, the company would prefer to sweep the evidence under the rug with a little marketing. This would not be possible if a reliable, independent benchmarking organization existed.

Other vendors, especially Intel competitors, have even less enthusiasm for independent benchmarking organizations. The last thing most companies want is for someone to shine a bright light on the real performance of their products. Apple, for example, would have had to forgo a successful marketing campaign if a credible benchmark organization had been around to call B.S. on its ridiculous assertion that PowerPC processors were twice as fast as Pentium IIIs, a claim Apple based on the self-serving Bytemark benchmark. As long as independent benchmarking is not in the best long-term interests of the wealthy parties to the struggle, it is not likely to meet with much success.

Independent benchmarking would be more affordable if the industry would just agree on one standard benchmark or suite. But the likelihood of that happening is vanishingly small. Relying on vendor-developed benchmarks is absurd on its face, and asking vendors to run third-party benchmarks and report their results just deteriorates into the situation we have now: vendors avoid reporting results that put their product in an unfavorable light. Expecting vendors to play fair and to police themselves in a game whose outcome can determine their market valuation is just incongruous with reason.

It is a sad situation indeed, but as poor a measure of performance as frequency is, it may be the best performance metric the industry has available. The correlation between frequency and our subjective evaluations of performance is not all that bad. It is actually quite good among processors with the same microarchitecture (e.g., P6s) and not even that bad between similar processors with different microarchitectures (e.g., P6s and Athlons). Overall, frequency appears to be about as reliable as any other benchmark we know of.

Regardless of how well frequency reflects performance, however, consumers have clearly latched onto it as their preferred benchmark. This fact is due in large part, I'm sure, to the undecipherable noise generated by the current cornucopia of benchmarks. At least, frequency is easily and objectively measurable. Maybe consumers aren't as stupid as we think they are—at least on this issue.

The danger with frequency as the official "benchmark," of course, is that it motivates vendors to optimize their processors for frequency over performance. Intel's upcoming Willamette, with its breathtakingly long (20-stage) pipeline, is perhaps the most obvious example, but others have also quietly admitted to us that they are considering sacrificing parallelism, and even bottom-line performance, for frequency.

Similar design distortions would surely occur even if the industry used a standard benchmark instead of frequency. Just as engineers once added the "Dhrystone" instruction and today tune their caches and compilers for SPECint, any potentially popular benchmark is bound to face similar attacks. Benchmark programs have historically proved relatively easy to subvert. Changing benchmarks rapidly enough to preclude their being compromised is one way to avoid the problem, but it makes them useless for comparing processor performance over time, which is an important benchmark function. Frequency is, if nothing else, more immune to this problem.

Now, before I get a flood of email accusing me of heresy, let me make it clear that I believe benchmarks do have a place. A good independent processor benchmark or benchmark suite, such as SPEC2000, in the hands of a knowledgeable processor architect or compiler writer, is a very valuable tool. Users who intend to use a processor primarily for one specific application, such as Photoshop, can benefit from a focused application-specific benchmark. And because embedded processors tend to be used in narrow domains, we may see more success with benchmarks in this arena. The situation in the embedded space has not yet gotten as far out of control as it has in the PC space, and the credible efforts of EEMBC (see

MPR 5/1/00-02, "EEMBC Releases First Benchmarks") will, we hope, be in time to prevent it from doing so.

But as a method for consumers to evaluate PC-microprocessor performance, benchmarks are a disappointing failure. Aside from providing fodder for geek-oriented magazine articles and Web sites, they provide little real value to most consumers. Unfortunately, the outlook for improvement is not bright. So, unfortunately, for the foreseeable future, we appear to be stuck with frequency.