THE EDITOR'S VIEW

The Trouble with Benchmarks

By Michael Slater

Computer system benchmarking has always been a troublesome issue. At first, native MIPS—actual millions of instructions per second—was used, as the most natural measure. It is fraught with problems, however, since it varies depending on the instruction mix and does not allow meaningful comparisons among architectures with different instruction sets. VAX MIPS then started to take over—a performance rating of 10 VAX MIPS simply meant 10 times the speed of a VAX 11/780.

This was better than native MIPS, but not much, since the question became "on what program?" The Dhrystone benchmark quickly became the most widely used, and today, VAX MIPS most commonly means performance on Dhrystone relative to a VAX 11/780. The problem with Dhrystone is that compiler-writers found ways to optimize Dhrystone performance that did not have much effect on most other applications, so "Dhrystone MIPS" tends to overstate performance.

Of course, this is just what many marketeers seem to want—whatever will give them the biggest number. As a further example of this attitude, many companies still quote Dhrystone 1.1 in most cases, despite the fact that version 2 was released over five years ago and is preferred because it defeats some of the more meaningless compiler "cheats." So why do major companies continue to quote Dhrystone 1.1 ratings, and VAX MIPS ratings based on them? Because it gives bigger numbers than Dhrystone 2.0.

Now, native MIPS is back in fashion, since it gives great numbers for superscalar processors. The three-issue, 40-MHz SuperSPARC can deliver 120 peak, native MIPS, but don't expect this number to correlate to anything meaningful. A three-issue machine is not three times as fast as a single-issue machine, because of dependencies and other issue limitations.

When SPEC came on the scene three years ago, it looked like the answer to many of these problems. As a suite of 10 programs, it did not seem easily susceptible to compiler "cracking." To get a copy of the SPEC benchmarks, users have to sign a license agreement that sets some rules, including requiring complete disclosure of the system configuration and the 10 individual benchmark numbers whenever a SPECmark is quoted. It soon became popular to quote the geometric means for the integer and floating-point parts separately, and this later became part of the official reporting format.

Unfortunately, the SPEC rules simply aren't followed. On many occasions, composite SPEC numbers are provided for new chips or systems, without the detailed breakdown (SuperSPARC and hyperSPARC are two recent examples). It has also become common to quote simulated numbers with no system configuration information, hyperSPARC being a notable recent example. In some cases SPEC claims have been made without dividing them into integer and floating point. For example, HP claims that its PA-7100 will deliver 120 SPECmarks, without noting that the SPEC integer performance will be dramatically lower.

SPEC first ran into serious distortions when some users found that the "KAP" preprocessor, using supercomputer techniques, was able to increase performance on one benchmark (matrix300) by as much as an order of magnitude (see µPR 12/18/91, p. 3). This resulted in a gross inflation of SPECfp ratings.

Early this year, SPEC introduced a new suite, SPEC92, with twice as many programs—and no matrix300. This time, no composite metric is defined; integer and floating-point numbers are always separated. SPECfp92 produces much lower numbers than SPECfp89, primarily because of the elimination of matrix300. IBM's top-of-the-line system, for example, drops from 160.9 on SPECfp89 to 93.6 on SPECfp92. The number IBM chooses to highlight in the press releases, of course, is the composite SPEC89 figure of 100.3—conveniently ignoring the fact that the SPECint performance is below 50 and that even the floating-point rating reaches only 93.6 if the 1992 suite is used.

The SPEC92 suite is clearly better than SPEC89, but it is adding to the confusion. SPEC89 numbers are bigger, so vendors like the SPEC89 ratings. We're forced to use SPEC89 for some comparison tables because SPEC92 figures aren't available for all systems.

Sun is quoting only SPEC92 figures for the SPARCstation 10, which is laudable. (There is an ulterior motive. Sun never managed to get as big a boost on matrix300 as other vendors, so Sun looks better on SPEC92, relative to other vendors' SPEC92 numbers.) Beware of sloppy comparisons that mix the two—comparing one system's SPEC92 ratings to another's SPEC89 figures can be quite misleading. There are no simple answers, and for the moment, we will quote both figures when possible.

An even more fundamental problem with the SPEC suites is that they only measure computation speed. The workstation industry needs something like the new BAPCo suite for PCs (see p. 5), which measures full system performance—including disk and display—on real applications.♦