SPEC Data Helps Normalize Vendor mips-ratings for Sensible Comparison
OR
Your Mileage May Vary..But If Your Car Were A Computer, It Would Vary More

Issue 1.0 - 02/21/90

EXECUTIVE SUMMARY

John R. Mashey - MIPS Computer Systems

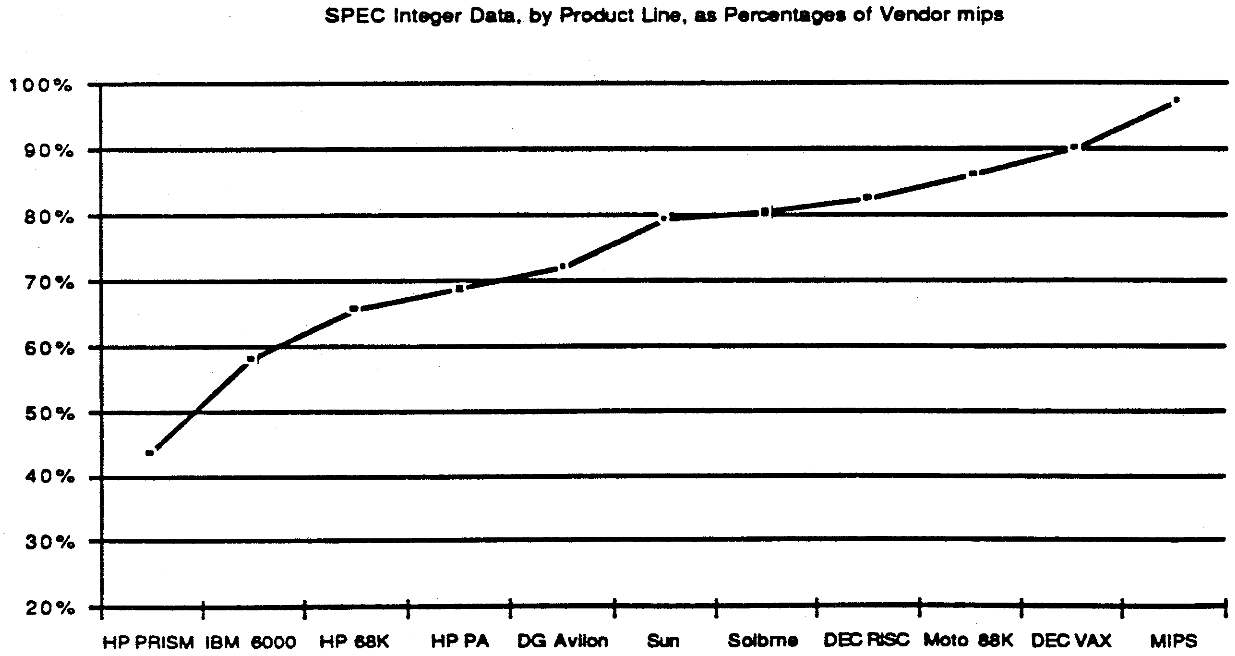
It is hardly news that the unadorned terms "mips" and "cost/mips" are essentially meaningless. In measuring floating-point performance, people have generally produced absolute measurements, based on benchmarks where people at least agree on the nature of the benchmark. However, for integer performance, everyone defines it differently. This leads to numerous apples-to-oranges comparisons and the user community remains suspicious, as users expect that everyone must inflate something that is as important as "mips", but is undefined.

Although there may not exist an "industry-standard defined mips", fortunately, more useful comparative data has just recently become available to help this problem. *The attached paper uses current data from the SPEC benchmarking effort to calibrate the confusing variety of mips-ratings now used.* The chart below summarizes the results, and tries to answer the questions:

> "If you know the claimed mips-rating for a computer, how does it perform on the SPEC integer benchmarks, as a percentage of the mips-rating?"

SPEC Benchmark Suite 1.0 consists of 10 large, heavily-scrutinized benchmarks, 4 Integer and 6 Floating Point. Results are reported as performance relative to a reference machine, a VAX-11/780 with current software. Thus, many vendors' systems are, finally, all compared to the same base machine, at about the same time, with the same benchmarks, yielding the most consistent measure of the elusive, but widely-used "VAX-mips" I could find.

For each product line below is shown a percentage that converts the vendor's published mips-rating into the Mean performance on the 4 SPEC Integer benchmarks. For example, for IBM the SPEC-measured performance is 58% of the vendor's mips-rating, while for MIPS Computer it is 97%. Thus a "34.5-mips" IBM RISC System/6000 and a "20-mips" MIPS RC3260 show very similar performance on these integer programs, and other vendors show similar variations. The paper's details show that vendors at least tend to use consistent ratings within product lines, but as this chart demonstrates, the wild variations among vendors render typical mips-ratings meaningless.



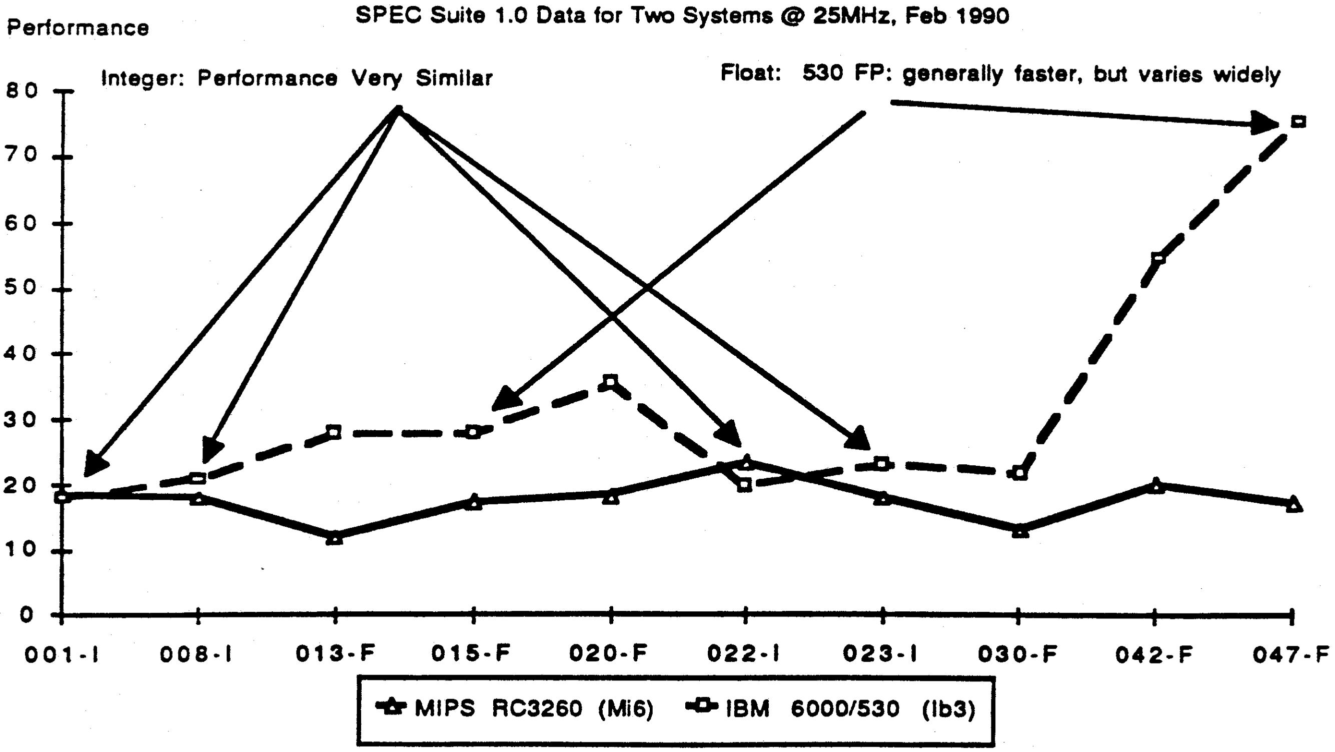
In the attached paper:

Section 1 reviews common different definitions of mips-ratings, showing the sources of confusion. It explains the approach of the SPEC benchmark effort, which involves many vendors in an effort to find more meaningful benchmarks over which to compete. This includes a brief description of the benchmark set of SPEC Release 1.0.

Section 2 analyzes the SPEC data that has been published by 8 vendors, covering many computers. It begins with the raw VAX-relative performance data for each of 10 benchmarks for 39 different hardware/software combinations. It then summarizes and charts these in different ways, to gain insight into the different measurement approaches, and nature of machines. It shows why SPEC insists that all numbers be published, not just summaries, so that people can see performance variations for themselves. An example is given in the chart below, which shows the VAX-relative performance for two RISC machines across 10 benchmarks (both integer and flaoting point).

Section 3 helps explain why so much work has been necessary to improve the quality of benchmarks. It explains why small benchmarks are notoriously unreliable in predicting the relative performance of systems on larger programs. It offers concrete examples to show the particular lack of correlation between the popular Dhrystone benchmark and larger integer programs. For example, Dhrystone predicts much better VAX-relative performance than systems actually show on real programs. It also poorly predicts the relative performance of machines from differing product lines. For example, for the two systems charted below, Dhrystone 1.1 predicts VAX-11/780-relative integer performance of 34.5X (IBM), and 24.5 (MIPS), but on the 4 large SPEC integer programs, the performance is about equal, 20X faster than that VAX-11/780.

All data is included, with detailed explanations. Some charts help show how people have arrived at very different mips-ratings for machines with similar performance, and others show the differing elationships between integer and floating point performance. Finally, readers who find the analysis convincing may be able to estimate performance on at least a few realistic applications just by knowing the vendor, product line, and mips-rating.


SPEC Suite 1.0 Data for Two Systems @ 25MHz, Feb 1990
Performance
Integer: Performance Very Similar
Float: 530 FP: generally faster, but varies widely
001-I
008-I
013-F
015-F
020-F
022-I
023-I
030-F
042-F
047-F
MIPS RC3260 (Mi6)
IBM 6000/530 (Ib3)