In my view, computer performance evaluation techniques are *means*, not an *end*. They are means that facilitates the rapid development of new computer designs by helping to separate the wheat from the chaff. For this role, it is important to distinguish *accuracy* and *precision*.

*Accuracy* is the difference between an estimate and the (usually unknown) true value being estimated. *Precision* gives the number of bits or digits available for specifying the answer. Given the radius of a circle estimated to be 5.00 meters, for example, many computers and calculators will report the circle’s area to be 49.3480220025 m$^2$. In high school science, however, most of us learned that, given the three significant digits of the radius’s estimate, a more appropriate answer for the circle’s area is 49.3 m$^2$.

The accuracy/precision question is pertinent computer performance evaluation. Often we seek simulators that mimic minute micro-architectural details, insist on benchmarks being simulated from start to finish, and report results via a speedup of 0.89999% (using 12,345,678,901 cycles for base and 12,456,789,012 cycles for enhanced). These results are highly precise.

But are they highly accurate? In many cases, the answer is “no.” Some inaccuracies are due to failing to model aspects of the target system that turn out to be important. Aspects omitted can include the operating system, virtual memory, threads, I/O, and multiprocessing. Other inaccuracies come from using benchmarks that are different from the target system’s likely workload. I have not yet seen a server or PDA run the SPEC benchmarks. Neither have I ever run a SPEC benchmark on any PC I own.

In my view, computer performance evaluation should strive for accuracy, not precision. When evaluating large multiprocessor servers, for example, we must estimate the effect of the operating system, virtual memory, threads, I/O, and multiprocessing. Furthermore, we must make
these estimates with realistic workloads (e.g., using commercial application servers and database management systems).

In some cases, we can perform full-system simulations, but brute force seems difficult. Assume that a user-level, uniprocessor simulation of the SPEC benchmark takes \( H \) hours. Full-system simulation (2x) of a 16-way multiprocessor (20x) running a full commercial workload (>100x) will take at least 4000 \( H \) hours.

For this reason, researchers studying servers must use other techniques to magnify simulation’s effectiveness. These include approximate processor models, approximate interconnect models, scaling workloads, careful use of sampling, selective use of analytic models, and hierarchical evaluation. Similar issues apply to other domains, such as networking, clustering, and, eventually, desktop and sub-desktop computing.

Going forward, I see three needed thrusts in computer performance evaluation.

- **Technical.** We need to develop the methodologies for building confidence when something other than completely-precise simulation is more accurate.

- **Tools.** We need to fund and value people and groups that develop tools that allow the community to leverage the methodological gains made elsewhere. I, for example, would like the community to have simulators for designing multiprocessor hardware, operating systems, and commercial applications that are as easy to use and accepted as Simplescalar.

- **Cultural.** We need to educate our community (and program committees) on methods beyond completely-precise simulation. I joke sometimes that papers with completely flawless simulation should be rejected as too incremental. Work on ideas that “push the envelope” will necessarily push methods and have educated estimates on machine parameters, workloads, etc.

To conclude, let us remember that evaluation methods are means to better computer designs. With these means, we should strive for accuracy, not precision. Moreover, let us remind ourselves that it is relative accuracy that matters. Papers demonstrating a 10% gain should be highly accuracy (e.g., less than 1% relative error) through excellent simulation, while papers examining a 10x change in technology or application area can be insightful with a back-of-the-envelope estimate that is accurate to about a factor of two. Finally, let us seek conferences with a mix of papers, where some focus on the near term and others target the longer term, while all use methods appropriate to their ends.