On the Road to Irrelevance?

David J. Lilja*

Abstract— **An examination of the papers published in a sampling of the proceedings of the International Symposium on Computer Architecture documents a profound shift in the techniques used to evaluate the performance of new computer architectures. The research presented in these papers has moved away from measurement of actual machines, analytical modeling, and descriptions of new ideas to an almost exclusive reliance on simulation. While it is impossible to prove precisely why this change has occurred, I speculate that there are three primary causes: 1) the "Hennessy and Patterson effect," 2) the development and widespread distribution of the SimpleScalar simulator, and 3) the wide availability of high-performance desktop computers capable of running these simulations. The use of simulations has encouraged the development of numerous new mechanisms for improving the performance of computer systems, many of which have been adopted by the computing industry and incorporated into new processor designs. These changes, in addition to improvements in semiconductor technology, have led to impressive gains in the performance of computer systems. However, it has been demonstrated [3] that the errors that exist in current simulations are often larger than the performance gains reported for the new ideas being evaluated. Computer architects have become too reliant on this single form of performance evaluation. We need to develop and encourage a diversity of techniques for evaluating the potential of new architectural ideas. Without this change, computer architecture researchers run the very real risk of moving into an intellectual backwater and becoming irrelevant to the computer industry.**

I. Trends in Performance Evaluation

As we think about the future of computer systems performance evaluation techniques, it is enlightening to first look at where we have been. An examination of a sampling of the proceedings of the International Symposium on Computer Architecture (ISCA) shows a clear trend towards simulation as the only acceptable approach for evaluating new architectural ideas. As shown in Table I, twenty-two of the twenty-five papers presented at ISCA in 2001 were based on simulation. One of these papers augmented the simulation results with measurements obtained on a real system, while only a single paper relied solely on measurement. The remaining two papers used neither measurement nor simulation. They were instead qualitative descriptions of some novel ideas or approaches. This emphasis on simulation repeated in 1997 and 1993, although several of the simulation-based papers in 1993 augmented or validated the simulation results with measurements on actual machines, or with analytical modeling.

In contrast to the heavy emphasis on simulation in these years, only 28 percent of the 43 papers presented at the conference in 1985 used simulation, while 33 percent used analytical modeling. Again only a single paper used actual measurements from a real system. However, in a special session of nonreferred papers, eight manufacturers described their existing systems. Of these papers, three presented actual performance measurements. The sixteen papers listed in the *Other* column in Table I qualitatively described new architectural ideas but presented no actual performance evaluations.

At the very first ISCA in 1973, 75 percent of the 28 papers were qualitative descriptions of new machines or ideas. Five of the papers included some analytical modeling with only two papers using any form of simulation. None of the papers reported actual measurements. The preface to the proceedings suggested that authors were encouraged to provide explanations of why specific architectural features were incorporated into a new design instead of simply describing the architecture. This encouragement to explain the design process was further emphasized by including five papers that discussed techniques for teaching computer architecture.

The remainder of this paper speculates on the causes of this shift towards simulation as the primary validation tool for computer architecture research and on where this trend may be taking us.

II. Proximate Causes

The proximate causes for a change or an event are those items and conditions that enable or encourage the change or event. The ultimate causes are those conditions that allowed the proximate causes to develop. For example, it has been argued [4] that one of the proximate causes for a community of humans to shift from hunting and gathering to farming is the existence of wild plants and large animals that can be domesticated. The ultimate causes for this shift from wild food sources to food production are the climatic and geographic conditions that allowed these progenitor species to develop and flourish in the given environment.

Some of the important ultimate causes or drivers for performing research in new computer architectures include the perceived needs of industry, the basic curiosity of the individual researchers, and the intellectual challenge provided by developing and evaluating new architectural ideas and mechanisms. While it is probably impossible to prove, I conjecture that there are three primary proximate causes for the shift to simulation as the primary performance evaluation technique in the computer architecture research community.

1. *The Hennessy and Patterson effect.* A quick perusal of ISCA proceedings from the 1970s into the mid-1980s suggests that early computer architecture research tended to be more descriptive than quantitative. There are a significant number of paper designs that proposed new architectures and new mechanisms, for instance, with little actual evaluation other than

*Department of Electrical and Computer Engineering, and Minnesota Supercomputing Institute, University of Minnesota, 200 Union St. SE, Minneapolis, MN 55455. E-mail: lilja@ece.umn.edu

TABLE I

THE PERFORMANCE EVALUATION METHODOLOGIES USED IN THE PAPERS THAT APPEARED IN A SAMPLING OF THE PROCEEDINGS OF THE INTERNATIONAL SYMPOSIUM ON COMPUTER ARCHITECTURE (ISCA). *Note:* THE TOTAL NUMBER OF PAPERS DOES NOT NECESSARILY ADD UP TO THE SUM OF THE NUMBER OF PAPERS ACROSS THE COLUMNS SINCE SOME PAPERS USED MORE THAN ONE EVALUATION TECHNIQUE.

Year	Number of papers*	Simulation	Measurement	Mathematical modeling	Other
2001	25	22	2	0	2
1997	30	24	6	0	0
1993	32	23	9	6	1
1985	43	12	1	14	16
1973	28	2	0	5	21

a qualitative summary of the idea's perceived advantages. Beginning in the 1980s, however, there began a definite shift towards more quantitative evaluations of new ideas. Due to the complexity and expense of actually constructing an entirely new computer system, these evaluations relied on simulations of benchmark programs to quantitatively compare the proposed ideas to existing mechanisms.

This notion of quantitatively evaluating new architectural ideas, particularly through simulations driven by real application programs, was strongly advanced by the 1990 publication of Hennessy and Patterson's now-classic textbook, "Computer Architecture: A Quantitative Approach" [8]. The widespread adoption of this textbook for advanced computer architecture courses has institutionalized this simulation-based approach to performance evaluation throughout the computer architecture research community. Even instructors who do not choose this particular textbook tend to encourage this quantitative mindset among their students since most of the available textbooks recognize the effectiveness of this approach and have incorporated it into the teaching of both basic and advanced concepts.

2. *SimpleScalar.* One of the inhibitors to performing good simulations is the difficulty of building an accurate, reliable simulator and all of the necessary support tools, such as compilers, assemblers, linkers, and loaders. In 1996, Burger *et al* [2] made the SimpleScalar simulation toolset freely available to the computer architecture research community.[1] Due to its relative simplicity of use, and its robustness, this simulator has been quickly and widely adopted by computer architecture researchers. Of the twenty-four simulation-based papers presented at the 1997 ISCA, the year after SimpleScalar was made widely available, only three used SimpleScalar. By 2001, however, 55 percent of the simulation-based papers used SimpleScalar. This tool is rapidly becoming an almost *de facto* standard for performing simulations in computer architecture research.
3. *Fast, inexpensive desktop computers.* In addition to a robust simulator, simulation-based research requires a computing system with enough memory and processing capability to run the simulations within a reasonable amount of time. Prior to the widespread availability of inexpensive desktop computers, this need for fast computers with large memories limited simulation studies to researchers with access to expensive systems. Today, however, the cost of high-speed processors and memory has been reduced so dramatically that almost any research group can purchase a desktop system that is capable of running a simulator such as SimpleScalar. Complete studies of large design spaces still may require access to large computational resources, such as those available at national and regional supercomputer centers. However, the combination of SimpleScalar and fast, inexpensive desktop systems has substantially lowered the barrier for performing rigorous, high-quality simulation studies.

[1] See `www.simplescalar.org` for more information about the current release.

III. LIMITATIONS OF SIMULATION

The "Hennessy and Patterson effect" has produced a widespread appreciation for quantitative comparisons when performing computer architecture research. Additionally, the availability of a free, robust simulator, combined with inexpensive high-performance computing systems, has made simulation the most common technique for evaluating new ideas in computer architecture. However, it is too easy for researchers to lose sight of the fact that a simulation really is only a model of an actual system. Common sources for errors and inaccuracies in simulations of computer systems include specification errors, modeling errors (bugs in the simulator [7]), and a lack of sufficient detail in the simulation model [1]. Furthermore, it is been shown [3] that the improvement in performance reported in many papers can actually be smaller than the experimental error in the simulator used to obtain these results.

In addition to these limitations of simulation, the time required to simulate new architectures is increasing dramatically as the architectural ideas being evaluated become more and more complex. Researchers have proposed techniques to control this increase in simulation time, such as *sampling* [10], [12] and reducing the run-time of standard benchmark programs [9]. However, these simulation time-

reduction techniques are only stop-gap measures that introduce their own sources of errors. It is clear that traditional simulation alone will not be able to keep pace with the demands for evaluating ever more complex systems.

IV. Prescription for Diversity

The trend towards evaluating new architectural ideas using comparative simulations has introduced a level of scientific rigor into computer architecture research that has led to many innovative new ideas. The convergence on the SimpleScalar toolset further allows a direct comparison of new ideas across studies using a common basis. However, this lack of diversity in evaluation techniques can lead to a narrowing of perspectives.

Computer architecture research has become too focused on simulations as the only acceptable approach for evaluating new architectural ideas. As a community, we need to encourage (and teach) the use of other methods of performance evaluation [11], especially for ideas that open up entirely new research directions. For instance, analytical modeling is an important technique for obtaining a feel for trends as parameters are changed, and for rapidly evaluating a large design space to find the most interesting points [6]. This fundamental technique is too often ignored or even denigrated in computer engineering curricula, however. It also would be appropriate to evaluate actual hardware implementations of specific point designs using FPGA prototypes, for instance, and to use simulations based on hardware description language models to obtain more precise information about actual costs and delays associated with a new idea.

Medical research studies often begin with experiments using mice as substitutes for humans. Results from these animal-based studies are used to filter out potentially dangerous or ineffective treatments before moving to human trials. Can we develop our own "mouse model" for computer architecture research? For instance, can we extrapolate from measurements of existing systems to validate simulations and suggest new research directions? Whatever model may be appropriate, it is clear that we will have to develop entirely new approaches for performance evaluation that cleverly combine analytical modeling, simulation, and measurement [5], [13], [14].

V. Conclusions

The history of the computer industry is littered with corpses of companies that over-specialized and failed to adapt to changing market needs and new technologies. As computer architecture researchers and educators, we have become over-specialized on simulation-based studies. University-based researchers need to teach a more complete range of performance evaluation techniques. Simultaneously, the overall research community needs to develop and accept a broader view of what constitutes "proof" of a good new idea. We seem to have lost sight of the fact that simulations are only models with limited fidelity, after all. They are not an end in and of themselves. Without a new respect for a diversity of performance evaluation approaches, we run the very real risk of becoming irrelevant to the industrial research and development community.

Acknowledgements

The computer architecture research work that has led me to these observations and speculations has been supported by a number of sources, including the National Science Foundation (currently under grants EIA-9971666 and CCR-9900605), Sun Microsystems, the IBM Corporation, and the Minnesota Supercomputing Institute. Their support is greatly appreciated.

References

[1] Bryan Black and John Paul Shen. Calibration of microprocessor performance models. In *IEEE Computer*, volume 31, pages 59–65, May 1998.

[2] D. Burger, T. M. Austin, and S. Bennett. Evaluating future microprocessors: The SimpleSclar tool set. In *University of Wisconsin-Madison Computer Science Department Technical Report no. CS-TR-96-1308*, July 1996.

[3] Rajagopalan Desikan, Doug Burger, and Stephen W. Keckler. Measuring experimental error in microprocessor simulation. In *International Symposium on Computer Architecture*, pages 266–277, 2001.

[4] Jared Diamond. Guns, germs, and steel: The fates of human societies. New York, 1997. W. W. Norton and Company.

[5] L. Eeckhout and K. De Bosschere. Hybrid analytical-statistical modeling for efficiently exploring architecture and workload design spaces. In *International Conference on Parallel Architecture and Compilation Techiques*, pages 25–34, 2001.

[6] Michael J. Flynn. *Computer Architecture: Pipelined and Parallel Processor Design.* Jones and Bartlett Publishers, Boston, 1995.

[7] Bob Glamm and David J. Lilja. Automatic verification of instruction set simulation using synchronized state comparison. In *Annual Simulation Symposium*, 2001.

[8] John L. Hennessy and David A. Patterson. *Computer Architecture: A Quantitative Approach.* Morgan Kaufmann, Inc., San Francisco, 1990.

[9] AJ KleinOsowski, John Flynn, Nancy Meares, and David J. Lilja. Adapting the SPEC 2000 benchmark suite for simulation-based computer architecture research. In *Workshop on Workload Characterization, International Conference on Computer Design*, September 2000.

[10] S. Laha, J. H. Patel, and R. K. Iyer. Accurate low-cost methods for performance evaluation of cache memory systems. In *IEEE Transactions on Computers*, volume 37, pages 1325–1336, November 1988.

[11] David J. Lilja. *Measuring Computer Performance: A Practitioner's Guide.* Cambridge University Press, Cambridge, United Kingdom, 2000.

[12] D. B. Noonburg and J. P. Shen. A framework for statistical modeling of superscalar processor performance. In *International Symposium on High-Performance Computer Architecture*, pages 298–309, 1997.

[13] S. Nussbaum and J. E. Smith. Modeling superscalar processors via statistical simulation. In *International Conference on Parallel Architecture and Compilation Techiques*, pages 15–24, 2001.

[14] M. Oskin, F. T. Chong, and M. Farrens. HLS: Combining statistical and symbolic simulation to guide microprocessor designs. In *International Symposium on Computer Architecture*, pages 71–82, 2000.