



Computer architecture research: Shifting priorities and newer challenges

PRADIP BOSE

pbose@us.ibm.com

..... CMOS technology roadmaps today are marked with “problem” markers and footnotes that are becoming all too familiar: power and power density overruns, intra- and interchip parametric variabilities, significant slowdown in delivered frequency growth, possible increase in hardware failure rates, lower effective yield because of chip leakage variability, and so on. What does this all mean to computer architecture researchers, especially to those focused on microprocessor design? Is the processor core a relevant focus for microarchitects and chip designers any more, and is performance enhancement still the primary goal and driver of research in this field?

Assuming that basic Moore’s law scaling will continue to yield increased transistor counts per chip die (except the average transistor might not grow in speed with each generation at quite the same rate as before), how do we use those transistors effectively and efficiently to meet chip-level performance growth targets, without exceeding affordable power budgets? Will we use most of those extra transistors to boost metrics other than just raw performance, such as error tolerance, self-testing, and fault recovery?

In the realm of high-end microprocessors, the trend toward lower-frequency multicore architectures is now clearly gaining ground. This is a definite paradigm shift away from single-core chip microarchitectures, which have witnessed an ever-increasing complexity in terms of pipeline depth, issue width, and speculative processing. Until now, power and power den-

sity limits have largely driven this shift. But, variability and increased failure rates might further enhance this trend. Architectures can provide on-chip redundancy through spare cores. Designs that support efficient task migration across multiple processing engines can help reduce average temperatures. It might be easier to implement power-gating to reduce leakage at the granularity of cores and cache banks rather than weaving it pervasively into the core design itself. Defective or over-leaky cores can be disabled to make use of partially good chips, boosting effective yield. Separate clocks for the multiple cores, with asynchronous interconnect interfaces, might help control clock distribution problems and provide further levers for chip-level power management through clock speeds that are adjustable as post-silicon tweaks, or even dynamically, during computation. This means, in effect, that we may continue to get the benefit of transistor growth through scaling for some time, but we will not be able to use all of them at the same time for useful work; and many of those additional transistors may need to be used for overhead tasks like fault tolerance, security, power management, etc.

OK; but what about growth in basic performance and functionality at the chip level? Does the trend in favor of multicore integration (as opposed to raw frequency growth) imply that chip-level throughput measured across multiple (largely independent) threads will be the sole performance metric of interest in the future? Maybe, but it is too early to write off growth in single-thread performance. The

exploding variety and size of application programs, many with real-time latency constraints (like those in the game and multimedia space) will drive the quest for balanced single- and multithread performance growth for quite some time. For example, power-gateable, special-purpose accelerators that provide an energy-efficient, on-demand “turbo” boost to performance comes to mind. The bottom line is: chip microarchitects must work in conjunction with circuit, technology and tools groups to ensure net (balanced) growth in single-thread and chip-level throughput performance and/or functionality, at affordable power, in order to justify the continuance of the technology scaling game

In other words, the future of computer architecture research remains bright; arguably, it is even brighter than before, as microarchitects take on the challenge to keep chip-level performance growing at historical levels in the wake of the frequency growth slowdown brought on by technological constraints. This particular year-end issue of *IEEE Micro* is a specially crafted one. Like last year, the committee has tried to select the most promising research ideas and evaluations from the past year of conference publications and asked the authors to present them in a form that the practicing microarchitect, designer, and researcher can easily digest. *IEEE Micro* is grateful to editorial board member David Albonesi and the team of expert reviewers whom he drew upon for taking the time to put together this final selection of articles after an intense and very careful review process.