- K. Lim et al., "Disaggregated Memory for Expansion and Sharing in Blade Servers," Proc. 36th Ann. Int'l Symp. Computer Architecture (ISCA 09), 2009, pp. 267-278.
- S. Li et al., "System-Level Integrated Server Architectures for Scale-Out Datacenters," Proc. 44th Ann. IEEE/

ACM Int'l Symp. Microarchitecture, 2011, pp. 260-271.

 J. Chang et al., "Totally Green: Evaluating and Designing Servers for Lifecycle Environmental Impact," Proc. 17th Int'l Conf. Architectural Support for Programming Languages and Operating Systems, 2012, pp. 25-36. **Steve Scott** is a principal engineer in the Platforms group at Google, where he is working on the design of future Google systems. Scott has a PhD in computer architecture from the University of Wisconsin at Madison. He is a fellow of IEEE and the ACM.



2013 International Symposium on Computer Architecture Influential Paper Award

MARGARET MARTONOSI Princeton University

•••• Each year at the International Symposium on Computer Architecture (ISCA), the Influential Paper Award recognizes the paper from the ISCA proceedings 15 years earlier that has had the most impact on the field (in terms of research, development, products, or ideas) during the intervening years. Candidate papers for the award are selected by the current year's ISCA Program Committee (PC). The final award selection was made by the current ISCA PC Chair (Margaret Martonosi), the ACM Special Interest Group on Computer Architecture (SIGARCH) Chair (David Wood), and the IEEE Computer Society Technical Committee on Computer Architecture (TCCA) Chair (David Kaeli). The award includes an honorarium for the authors and a certificate.

Although ISCA 1998 offered many worthy papers to consider, the 2013 award was given to the paper "Pipeline Gating: Speculation Control for Energy Reduction" by Srilatha Manne, Artur Klauser, and Dirk Grunwald. All the authors were at the University of Colorado at the time of publication.

The paper explores methods to estimate the confidence of branch predictor outcomes, with a goal of speculating past branches with low-confidence predictions less aggressively. In superscalar processors, branch prediction has long been used to allow instruction fetch, decode, and execute to continue speculatively after a conditional branch with a not-yet-known outcome has been encountered. If the branch predictor causes speculation to continue down an incorrect path, then all such speculated instructions must be squashed so that execution can instead continue down the correct path. Although branch speculation is almost always good for performance, this pipeline gating paper was the first to acknowledge the power or energy cost of speculation down incorrect paths. Instead, their technique only speculated past branches if the confidence of the branch prediction was high enough. In this way, speculation could

help performance for the cases when it was most likely to be correct, while limiting the power and energy impact of incorrect predictions.

The paper is recognized as the first ISCA paper—and indeed the first paper in the computer architecture research community overall-to focus directly on power issues. In particular, the pipeline gating idea does not seek to improve performance, which had been ISCA's design metric of choice for many years. Rather, this paper's proposed techniques seek to improve CPU power and energy without unduly impacting performance. This shift in mindset was seminal. Prior to the late 1990s, the prevailing wisdom was that power was a metric for device and circuit researchers to worry about, but too low level to be on the radar screen of computer architecture researchers. Clearly, now power stands alongside performance as a first-order design metric for computer architects.

In the years since its publication, the work has had significant impact. With

.....

AWARDS

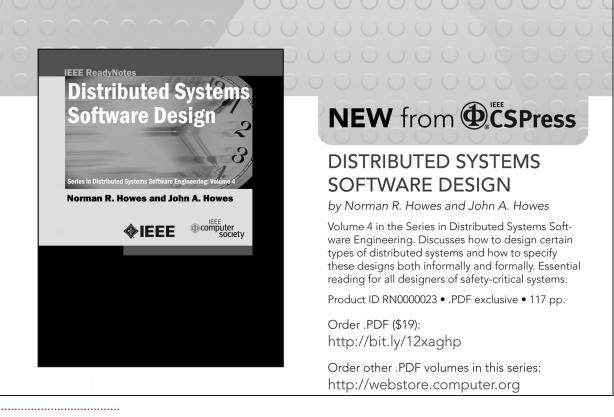
nearly 500 citations, the paper has clearly influenced a wide range of subsequent publications. At the time, my research group at Princeton University had also begun to think about power-oriented architecture techniques, and we viewed this paper as an encouraging sign that ISCA was recognizing the relevance and value of such work for computer architects. The paper's ideas had industry influence as well. While we know of no industry processor directly employing pipeline gating, the paper's philosophy of curtailing "extra work" predicted to be of lower payoff caught hold over the years. In a general sense, this work was part of a multiyear groundswell of interest in

power/performance tradeoffs that over time has shifted computer architecture.

n addition to the effectiveness of the ideas and the relative simplicity of their implementation, another strong point of this paper that aided its long-term influence is its well-written accessibility. The paper reports its results not in terms of watts or joules, but in terms of the relative numbers of instructions executed or squashed under different scenarios. This helped the significant instruction speculation savings to be accessible and understandable to a wide range of readers. Overall, we congratulate the authors on this fine work.

Margaret Martonosi is the Hugh Trumbull Adams '35 Professor of Computer Science at Princeton University. Her research interests include computer architecture and mobile computing, focusing on power-performance tradeoffs in systems ranging from smartphones to chip multiprocessors to largescale datacenters. Martonosi has a PhD in electrical engineering from Stanford University. She is a fellow of IEEE and the ACM.

Cn Selected CS articles and columns are also available for free at http://ComputingNow. computer.org.



IEEE MICRO