Perspective

Variability Expeditions: A Retrospective

Rajesh K. Gupta
UC San Diego

Puneet Gupta
UCLA

Subhasish Mitra
Stanford

Expeditions in Computing are among the largest and most ambitious projects sponsored by the National Science Foundation that seek to explore far out ideas in computing with potential for significant impact on computing and its industry. In 2009, we conceived of an Expeditions project in response to an alarming trend in the semiconductor industry: the manufactured chips were demonstrating significant variation in their performance and power consumption that was increasing at a rapid pace, leading designers to overdesign circuits with margins (e.g., margins on speed and voltage, also referred to as guard bands) that often exceeded 40% of the nominal target specifications. This overdesign threatened to make new process nodes ineffective by causing significant loss of yield, effectively wiping out gains due to scaling geometries. In addition to manufacturing, computing machines were experiencing increasing variation in operating conditions due to the proliferating use of these devices in mobile and wireless applications.

We envisioned an alternate universe where computing systems would sense their conditions and their environment and adapt to them. Software would drive such adaptation as the underlying components deviated from the normative manufacturing parameters or aged or just faced different operating conditions. Besides improved efficiency (e.g., in energy) due to reduced or eliminated overdesign, such systems will be inherently more reliable and available due to their continued operation through uncertain operating hardware or environments. The latter aspect is becoming critical as our reliance on autonomous systems (e.g., autonomous driving) continues to increase. The Expedition was an audacious attempt to rethink the computing universe, one worthy of an expedition to a new way of computing, where sensing circuits provide signatures that propagate through a new software stack, one that matches application needs to underlying physical capabilities, scaling one or the other appropriately. Software, appropriately enabled by sensing hardware, would ultimately provide an expanding source of new capabilities that tradeoff reliability, costs against the quality of computing, and/or storage.

Expedition took on a detailed characterization of uncertainty in computing from spatial dimensions such as manufacturing to temporal dimensions such as circuit aging effects, to dynamic variations caused by workload and operating environments. As we built circuits, microarchitecture, devised coding methods, and adaptive algorithms, the research accelerated the trends toward fault tolerance in programming languages from fringe efforts such as principled approximation and probabilistic accuracy bounds to more mainstream approximate computing. The research showed that some of the classical fault tolerance techniques had outlived utility in the new computing systems. Instead, resilience enabled by cross-layer measures carried a prominent role in building robust systems that relied upon new methods for online self-test, diagnostics, and self-repair. Techniques such as concurrent autonomous chip self-test using stored test patterns assisted

Digital Object Identifier 10.1109/MDAT.2018.2889103
Date of current version: 15 February 2019.

Copublished by the IEEE CEDA, IEEE CASS, IEEE SSCS, and TTTC

by software orchestration were demonstrated using OpenSPARC T2 design to be capable of autonomously self-repairing multiple faulty components with graceful degradation of system performance for an order of magnitude less cost than traditional redundancy-based techniques. Inspiration from such methods led to the creation of imperfection-immune design techniques that enabled practical "beyond-silicon" nanosystems using the emerging carbon nanotube devices.

With over 25 doctoral students spread among 14 principal investigators, a significant amount of characterization, simulation, and experimental techniques were explored and implemented. These techniques ranged from novel sensing circuits, error recovery circuits, and microarchitectural assists that worked in concert with software, error correcting codes that automatically and opportunistically recover using software-defined methods, techniques that estimate effects on application quality to building a complete programming environment with a set of programming abstractions and associated runtime and a context awareness toolkit that was recognized with a best paper award at the prestigious ACM Mobicom 2015. The new programming environment helps application developers compose inferences in a modular way and it can synthesize an inference pipeline that is partitioned across multiple devices. This is, of course, one of the series of tools and simulators that emerged from the project, notably VIP-Zone, a variability-aware memory allocator, VarEMU, a variability emulator, and so on. It is not the purpose of this recollection to be a comprehensive list of contributions. The Expeditions were at the source with multiple promising thoughtlines in a broad community of researchers directly leading to efforts that are currently underway in two of the six Semiconductor Research Corporation/Defense Advanced Research Projects Agency (SRC/DARPA) centers, Computing on the Network Edge (CONIX) and Center for Research on Intelligent Storage and Processing in Memory (CRISP), focused on distributed computation and memory systems, respectively.

Yet perhaps the most important outcome is the tremendous pool of talent we created for industry and academia. With over half a dozen academic researchers at universities such as CMU, MIT, U Chicago, U Minnesota, Villanova, and overseas at ETHZ and NTU Singapore, the expeditions' spark continues to drive the research agenda in areas such as approximate computing, brain-inspired hyperdimensonal computing, and nanosystems using post-CMOS computing and storage devices. Indeed, growth in approximate computing, especially in the programming language community, is a direct consequence of the research pursued in Variability Expeditions.

Nearly a decade since we put the team together to build the Variability Expeditions, we can undoubtedly state that technology, circuit, microarchitecture, and, to a lesser extent, system software have embraced handling hardware variations and their manifestations either as power/performance variabilities or as errors. Embedded hardware variation monitors and their use in various system-level adaptations (e.g., voltage/frequency scaling) are increasingly common; environmental variation (e.g., temperature) aware management of distributed compute resources is common; several commercial processors have tried implementing variation-aware microarchitectural adaptation; and novel error correcting codes and schemes for existing and emerging memories and storage are in use or under active investigation.

Going back to the original source of inspiration from microelectronics manufacturing, it is worth pointing out a few current trends in the semiconductor industry. Scaling silicon technology has become a very expensive exercise and vanishing few companies are able to afford it. Scaling enablers such as extreme ultraviolet (EUV) lithography have been delayed and industry is now revisiting "beyond silicon" technologies for complementing conventional silicon technologies. The countervailing trend has been a rise in hardware startups and a surprising increase in the number of design startups driven by new applications such as machine learning, cryptocurrencies, and computer vision. Beside scaling to advanced technology nodes (through the use of EUV, for example), semiconductor foundries are augmenting existing process nodes to support special features to extract the maximum value from these nodes. For example, new nonsilicon memory technologies such as Phase change RAM (PCRAM), Magnetic RAM (MRAMs), and Resistive RAM (RRAM) are coming online. The emerging memory research is also feeding into the growing interest in neuromorphic systems. Architectural and software support for these new memory systems is an area of active research and development. Another promising direction has been the scaling of "outside the chip," i.e., integration technologies. Adoption of interposers and 3D integration has promised to be disruptive and has caused a surge in research spanning the entire hardware–software stack.

At the same time, we acknowledge that affecting software layers has been very challenging with limited adoption in practice, despite the growing research agenda in reliability-aware software. The software/compiler researchers brought together by the Expeditions had a mixed reaction in the early years, partly because the mainstream software community was not yet attuned to software's role in mitigating hardware variabilities. The idea that software could be a solution to any problem in hardware was considered far-fetched and did not see as much traction in the mainstream publication venues as we had hoped. Indeed, our researchers in operating systems found very little motivation to go beyond the prevailing trends in consolidating operating systems and their growth toward cloud-centric computing stack, rather than looking toward hardware. Ironically, in a post-Expedition period, our software (especially programming language) colleagues have discovered the growing role of software in handling variability in execution environments through innovations such as low-level liquid (logically qualified) data types and their use in automated synthesis of programs, assertions, and adapters between program components to handle dependencies on the hardware and operating environments. A new generation of researchers has emerged, notably at EPFL, MIT, UW, and other schools that are pushing the state of the art in approximate computing. Some more recent work in applications of machine learning techniques also seeks to contribute to the vision of approximate computing. Furthermore, as the use of hardware acceleration increases, driven by performance, cost, and scaling reasons in all kinds of systems, evolution of the software and systems stack is likely to follow, along with the tools and methods that were prototyped in the Variability Expeditions. ■

Rajesh K. Gupta is a Distinguished Professor of computer science and engineering at the University of California San Diego, San Diego, CA, USA, where he currently directs the Halıcıoğlu Data Science Institute. He is a Fellow of the IEEE, the ACM, and the American Association for the Advancement of Science.

Subhasish Mitra is Professor of electrical engineering and of computer science at Stanford University, Stanford, CA, USA, where he directs the Stanford Robust Systems Group and co-leads the Computation focus area of the Stanford SystemX Alliance. He is also a Faculty Member of the Stanford Neurosciences Institute, Stanford University. His research interests include robust computing, nanosystems, VLSI design, validation, test and electronic design automation, and neurosciences. He is a Fellow of the ACM and the IEEE.

Puneet Gupta is currently a Faculty Member of the Electrical and Computer Engineering Department, University of California, Los Angeles, Los Angeles, CA, USA. His research interests include building high-value bridges across software-hardware-technology interfaces for lowered cost and power, increased yield, and improved predictability of integrated circuits and systems. Gupta has a BTech in electrical engineering from the Indian Institute of Technology, Delhi, India (2000) and a PhD from the University of California San Diego, San Diego, CA, USA (2007).

■ Direct questions and comments about this article to Rajesh K. Gupta, UC San Diego, La Jolla, CA, USA; gupta@eng.ucsd.edu.