'Approximate computing' improves efficiency, saves energy

December 18 2013, by Emil Venere

(Phys.org) —Researchers are developing computers capable of "approximate computing" to perform calculations good enough for certain tasks that don't require perfect accuracy, potentially doubling efficiency and reducing energy consumption.

"The need for approximate computing is driven by two factors: a fundamental shift in the nature of computing workloads, and the need for new sources of efficiency," said Anand Raghunathan, a Purdue Professor of Electrical and Computer Engineering, who has been working in the field for about five years. "Computers were first designed to be precise calculators that solved problems where they were expected to produce an exact numerical value. However, the demand for computing today is driven by very different applications. Mobile and embedded devices need to process richer media, and are getting smarter – understanding us, being more context-aware and having more natural user interfaces. On the other hand, there is an explosion in digital data searched, interpreted, and mined by data centers."

A growing number of applications are designed to tolerate "noisy" real-world inputs and use statistical or probabilistic types of computations.

"The nature of these computations is different from the traditional computations where you need a precise answer," said Srimat Chakradhar, department head for Computing Systems Architecture at NEC Laboratories America, who collaborated with the Purdue team. "Here, you are looking for the best match since there is no golden

answer, or you are trying to provide results that are of acceptable quality, but you are not trying to be perfect."

However, today's computers are designed to compute precise results even when it is not necessary. Approximate computing could endow computers with a capability similar to the human brain's ability to scale the degree of accuracy needed for a given task. New findings were detailed in research presented during the IEEE/ACM International Symposium on Microarchitecture, Dec. 7-11 at the University of California, Davis.

The inability to perform to the required level of accuracy is inherently inefficient and saps energy.

"If I asked you to divide 500 by 21 and I asked you whether the answer is greater than one, you would say yes right away," Raghunathan said. "You are doing division but not to the full accuracy. If I asked you whether it is greater than 30, you would probably take a little longer, but if I ask you if it's greater than 23, you might have to think even harder. The application context dictates different levels of effort, and humans are capable of this scalable approach, but computer software and hardware are not like that. They often compute to the same level of accuracy all the time."

Purdue researchers have developed a range of hardware techniques to demonstrate approximate computing, showing a potential for improvements in energy efficiency.

The research paper presented during the IEEE/ACM International Symposium on Microarchitecture was authored by doctoral student Swagath Venkataramani; former Purdue doctoral student Vinay K. Chippa; Chakradhar; Kaushik Roy, Purdue's Edward G. Tiedemann Jr. Distinguished Professor of Electrical and Computer Engineering; and

Raghunathan.

Recently, the researchers have shown how to apply approximate computing to programmable processors, which are ubiquitous in computers, servers and consumer electronics.

"In order to have a broad impact we need to be able to apply this technology to programmable processors," Roy said. "And now we have shown how to design a programmable processor to perform approximate computing."

The researchers achieved this milestone by altering the "instruction set," which is the interface between software and hardware. "Quality fields" added to the instruction set allow the software to tell the hardware the level of accuracy needed for a given task. They have created a prototype programmable processor called Quora based on this approach.

"You are able to program for quality, and that's the real hallmark of this work," lead author Venkataramani said. "The hardware can use the quality fields and perform energy efficient computing, and what we have seen is that we can easily double energy efficiency."

In other recent work, led by Chippa, the Purdue team fabricated an approximate "accelerator" for recognition and data mining.

"We have an actual hardware platform, a silicon chip that we've had fabricated, which is an approximate processor for recognition and data mining," Raghunathan said. "Approximate computing is far closer to reality than we thought even a few years ago."

 More information: Quality Programmable Vector Processors for Approximate Computing, IEEE/ACM International Symposium on Microarchitecture, Dec. 7-11

Abstract

Approximate computing leverages the intrinsic resilience of applications to inexactness in their computations, to achieve a desirable trade-off between efficiency (performance or energy) and acceptable quality of results. To broaden the applicability of approximate computing, we propose quality programmable processors, in which the notion of quality is explicitly codified in the HW/SW interface, i.e., the instruction set. The ISA of a quality programmable processor contains instructions associated with quality fields to specify the accuracy level that must be met during their execution. We show that this ability to control the accuracy of instruction execution greatly enhances the scope of approximate computing, allowing it to be applied to larger parts of programs. The micro-architecture of a quality programmable processor contains hardware mechanisms that translate the instruction-level quality specifications into energy savings. Additionally, it may expose the actual error incurred during the execution of each instruction (which may be less than the specified limit) back to software. As a first embodiment of quality programmable processors, we present the design of Quora, an energy efficient, quality programmable vector processor. Quora utilizes a 3-tiered hierarchy of processing elements that provide distinctly different energy vs. quality trade-off, and uses hardware mechanisms based on precision scaling with error monitoring and compensation to facilitate quality programmable execution. We evaluate an implementation of Quora with 289 processing elements in 45nm technology. The results demonstrate that leveraging quality-programmability leads to 1.05X-1.7X savings in energy for virtually no loss (< 0.5%) in application output quality, and 1.18X-2.1X energy savings for modest impact (