Keynote II

Approximate Computing for Energy-efficient Error-resilient Multimedia Systems

Kaushik Roy
Electrical & Computer Engineering
Purdue University
West Lafayette, Indiana, USA
kaushik@purdue.edu

The rapid advancement in scaled silicon technology has resulted in the influx of numerous consumer devices with a plethora of applications. Multimedia applications which use image and video processing, pattern or facial recognition, data mining and synthesis have seen a significant increase in user base. These applications not only demand complex signal processing of digital data to achieve quality requirements specified by the user, but also need to operate in an energy-efficient manner, posing a significant design challenge. It should also be noted that majority of these applications have an inherent error-resiliency. This arises from the fact that:

(a) these algorithms have to be noise tolerant to deal with real world input data,
(b) large data sets are processed frequently with significant redundancy,
(c) statistical or probabilistic computations are used in several cases,
(d) human perception does not discern a small amount of error in output.

Error resiliency implies that it is possible to introduce approximations in intermediate computations of these algorithms without causing any significant deterioration in the quality of the output. Such relaxation in computational accuracy can be introduced at various levels of abstraction, thereby reducing the complexity of computations in the algorithms involved. Such complexity reduction directly manifests into reduced power consumption of the system. This class of low-power design techniques, which are classified under the paradigm of "approximate computing", has received significant attention recently. These techniques relax the conventional requirement of perfect equivalence between the specification and hardware implementation, and translate this flexibility into energy efficiency. The introduction of approximate computation at various levels of abstraction is discussed in brief.

Algorithm-Architecture Level: In most of the multimedia signal processing, it is observed that there is a varying impact of different computations on the output quality. Based on this observation, at the algorithm-architecture level of abstraction, it is possible to classify the computations into "significant" and "non-significant" computations. The significant computations increasingly affect the overall output quality as compared to the non-significant computations. Using statistical or probabilistic techniques, the significant computations can be intelligently identified. It is imperative that these computations are performed in an error-free manner in order to maintain the required output quality. In contrast, the non-significant computations can be subjected to calculated approximations, thereby, exploiting the algorithm's error resiliency and hence potentially increasing energy efficiency. Subsequently, the underlying architecture can be modified to predict the critical path activation so as to compute the significant computations precisely under voltage over scaling (VOS). Further improvement in error resiliency can be obtained by changing the probability of activation of critical path depending on the degree of VOS or process variations. For instance, a significance driven approach for implementation of a tunable video motion estimator shows average power savings of 33% as compared to the conventional design in 90nm CMOS technology. The maximum quality loss in terms of PSNR was ~ 1dB. Depending on the process corner of the die, the statistical thresholds of algorithm can be varied to achieve energy efficient and variation tolerant operation with graceful degradation in quality.

Logic-Circuit Level: As seen in the previous section, it is possible to perform algorithm and architectural modifications in some class of applications without quality degradation. An alternative approach to utilize approximate computing for video and image processing application is logic complexity reduction. For instance, this can be achieved by reducing the number of transistor and internal node capacitances in a conventional mirror adder ensuring minimal errors in the full-adder truth table. Such reduction at the circuit level significantly reduces the power consumption of individual cell. Furthermore, multi-bit adders can be designed using a combination of accurate and approximate full adder cells. In order to keep a check on the propagated error induced by approximation, the approximate adders are used only to

This research was sponsored in part by National Science Foundation.

978-1-4673-6136-1/13/$31.00 ©2013 IEEE

operate on the least significant bits with minimal impact on the output quality. The most significant bits which heavily impact output quality are computed using accurate adders, thereby, providing overall error resiliency at reduced power consumption. Simulations show over 60% power savings and 37% area savings over existing implementation with negligible loss in output quality.

Cross-layer Optimization: It should be noted that introducing the approximate computations in single level of design abstraction is not the optimal technique for completely exploiting the error resiliency. In order to exploit the full potential of the inherent resiliency in the multimedia applications, it is essential to have a synergistic cross-layer optimization of approximate computation techniques across the various layers of design abstraction. For instance, it has been shown in the case of Support Vector Machine (SVM) classification application that, there is a 1.4X-2X improvement in energy savings when approximations were introduced by scaling the computations across the algorithm, architecture and circuit levels of abstraction as compared to any single level.