Progress Report (2018)

Self-Aware Computing for Cyber-Physical Systems

1. Personnel:

- *Mingoo Seok, Associate Professor, Columbia University*
- *Peter Kinget, Bernard J. Lechner Professor, Columbia University*
- *Doyun Kim, current graduate student, Columbia University*
- Jiangyi Li, past graduate student, currently with Apple
- Teng Yang, past graduate student, currently with Intel
- Seongjong Kim, past graduate student, currently with Intel

2. Project Period:

- Base: 1/Jan/2013-31/Dec/16 (4 years)
- Extend: 1/Jan/2017 – 31/Dec/2019 (3 years including the 1-year non-cost extension)
- Current: 1/Jan/2018 – 31/Dec/2018

3. Summary of the Progress in the Current Period

The goal of the project was to create a range of new techniques for making digital systems self-aware with *timing errors*, *temperature*, *device wear-out (aging)*, and other high-level parameters, with which the system can autonomously perform a *runtime framework* to determine the operating settings for optimal performance, energy-efficiency, robustness, reliability, and security, without the worst-case margin and over-design.

In the timing error (circuit delay) sensing thrust:

In [Li, ESSCIRC18], we have published our work on a novel load and power-management-unit (PMU) co-design method for emerging processors such as one consuming less than one microwatt. Such low power dissipation of load circuits makes the existing scheme (e.g., voltage regulation) less applicable. Therefore, we proposed an alternative regulation scheme based on timing error statistics, and this year we have further improved this so-called error-based regulation system in three aspects: (i) modulating supply voltage of a digital processor across Process, Voltage, Temperature (PVT) variations to remove the safety margin that is prohibitive for deep sub-Vt circuits; (ii) enabling continuous regulation by adding tunable replica circuits to the regulation system, which can set the upper and the lower bounds of supply voltage without imposing the worst-case design margin; (iii) optimizing its Power Conversion Efficiency (PCE) by automatically finding an optimal configuration for the DC-DC converter. Based on these new ideas, we prototyped a sub-μW Neural Spike Processor (NSP) integrated with a Power Management Unit (PMU) for a Brain-Computer-Interface (BCI) implant. This SoC demonstrates (i) among the highest level of integration including spike detection, sorting, the first half of decoding to reduce wireless data-rate by more than 4 orders of magnitude as well as PMU and (ii) the lowest power dissipation of 0.77µW for 96 channels, 21× lower than the state of the art at comparable/better accuracy.

In [Kim, TVLSI18], we also have submitted the journal paper based on our past work [Kim, VLSI16]. The focus is to develop a new EDAC technique that is applicable to a processor employing both specialized architecture and near and sub-threshold voltage circuits. The existing EDAC techniques have targeted only Von-Neumann architecture and nominal voltage circuits, it becomes non-trivial to apply them to the accelerators many of which do not base on Von-Neumann architecture. In particular, those accelerators often have no instruction, making it difficult to use the popular instruction-replay based error correction. To tackle this challenge, in this work, we propose a novel in-situ error detection and correction technique that utilizes dynamic, temporarily and spatially fine-grained body-swapping for error correction without instruction replay. Using the proposed technique, we prototyped a spiking neural network sorter in non-Von-Neumann architecture. The prototyped chip can successfully remove the worst-case margin and thus achieve 49.3% higher energy efficiency and 35.6% higher throughput compared to the baseline that operates with the worst-case margin. The proposed technique incurs only 4.1% silicon area overhead and requires no additional supply voltage.

We have continued to create techniques for energy-efficient but variation-tolerant digital circuits. In [Zhang, DAC19], we have investigated the design of digital standard cell library design for near/sub-threshold circuits. A commercial standard multi-threshold logic cell library is designed for nominal super-threshold voltage circuits. Therefore, if blindly used at a sub-threshold voltage, such library exhibits excessively coarse granularity in driving strength, resulting in sub-optimal logic synthesis and placement-and-routing results. To tackle this problem, we propose a methodology to design a logic cell library for sub-threshold voltage circuits which has sufficient granularity while ensuring low leakage and small cell area overhead. The proposed holistic methodology leverages the reverse short channel effect, the inverse narrow width effect, and the uniform forward body bias for p-type transistors. Based on the methodology, we developed a 65-nm multi-threshold, multi-length library and benchmarked it against the super-threshold library across several common circuits in the physical design level simulations. The results show that the proposed technique achieves 26.6% reduction in power-delay product and 28.1% reduction in energy-delay-product at 5.8% area overhead on average under the highest performance constraints for synthesis and layout.

<u>In the temperature sensing thrust</u>:

In [Kim, JLPEA18], we have published the journal version of our past work on the extreme miniaturized thermal sensor circuits for the on-chip dynamic thermal management [Kim, CICC16]. The proposed sensor circuit directly senses the threshold voltage of a transistor that contains temperature information using a single PMOS device. This simple structure enables the sensor to achieve an ultra-compact footprint. The sensor also exhibits high accuracy and voltage-scalability down to 0.4V, allowing the sensor to be used in dynamic voltage frequency scaling systems without requiring extra power distribution or regulation. The compact footprint and voltage scalability enable our proposed sensor to be implemented in a digital standard-cell format, allowing aggressive sensor placement very close to target hotspots in digital blocks. The proposed sensor front end prototyped in a 65nm CMOS technology has a footprint of 30.1μm2, 3σ-error of ±1.1°C across 0 to 100°C after one temperature point calibration, marking a significant improvement over existing sensors designed for dynamic thermal management.

In [Li, JSSC18], we also have published a conference and a journal article on our transformable microcontroller-system-on-chip (µC-SoC). The µC-SoC can have its 6T SRAM instruction cache to dynamically transform to an ambient temperature sensor and physically unclonable function (PUF). As those sensing and PUF operations are performed in a duty-cycled manner in targeted systems, the proposed transformation can save silicon area, e.g., by ~9.8X as compared to the dedicated sensor and PUF circuits that achieve comparable robustness with our temperature sensor and PUF.

We have been further exploring the SRAM reuse concept and devised the technique to transform a regular SRAM array into a stochastic ADC. This SRAM-ADC approach needs a minimal modification compared to a standard SRAM, namely only in the SRAM peripherals. In SRAM mode, the SRAM-ADC operates as a conventional SRAM; in ADC mode, it repurposes each 6T SRAM bitcell as the gain stage of a comparator to implement a stochastic ADC. We prototyped an 8-kb SRAM-ADC in a 65nm CMOS. The chip consists of eight 1-kb macros. The area overhead is only 740µm2 or 0.93% of the area of the 8-kb SRAM-ADC. As compared to the prior dedicated stochastic ADC designs, our SRAM-ADC requires ~240X less silicon area. The results of the work were summarized and submitted for 2019 IEEE Custom Integrated Circuits Conference (CICC).

<u>In the aging sensing thrust:</u>

In [Yang, TVLSI18], we have published a journal article based on our in-situ, in-field sensing and compensation of 6-transistor SRAM circuits [Yang, ISSCC15]. This work has pioneered the in-situ and in-field technique for SRAM. The in-situ and in-field capability can be instrumental to implement the periodic maintenance framework for chip reliability.

In [Seok, IRPS18], we have also published an invited paper on our in-situ and in-field transistor aging management scheme. It reviews the recent works on this topic, including multiple of our own works for pipeline and SRAM circuits.

<u>In the runtime framework thrust,</u>

In [Zhang, ISLPED18], we have published our work on the dynamic voltage frequency scaling (DVFS) stack that is designed for mitigating a recently-emerging software-based fault injection attack called CLKSCREW. Most modern computing devices make available fine-grained control of operating frequency and voltage for power management. These interfaces, as demonstrated by recent attacks, open up a new class of software fault injection attacks that compromise security on commodity devices. The CLKSCREW attack stretches the frequency of devices beyond their operational limits to induce faults. Statically and permanently limiting frequency and voltage modulation space, i.e., guard-banding, could mitigate such attacks but it incurs large performance degradation and long testing time. Instead, we propose a run-time technique which dynamically blacklists unsafe operating performance points using a neural-net model. The model is first trained offline in the design time and then subsequently adjusted at run-time by inspecting a selected set of features such as power management control registers, timing-error signals, and core temperature. We designed the algorithm and hardware, titled a BlackList (BL) core, which is capable of detecting and mitigating such power management-based security attack at high accuracy. The BL core incurs a reasonably small amount of overhead in power, delay, and area.

In [Kim, ISLPED18], we suggest a new methodology in co-designing an integrated switched-capacitor converter and a digital load. Conventionally, a load has been specified to the minimum supply voltage and the maximum power dissipation, each found at her own worst-case process, workload, and environment condition. Furthermore, in designing an SC DC-DC converter toward this worst-case load specification, designers often have been adding another separate pessimistic assumption on power-switch's resistance and flying-capacitor's density of an SC converter. Such worst-case design methodology can lead to a significantly over-sized flying capacitor and thereby limit on-chip integration of a converter. Our proposed methodology instead adopts the better than worst-case (BTWC) perspective to avoid over-design and thus optimizes the area of an SC converter. Specifically, we propose BTWC load modeling where we specify non-pessimistic sets of supply voltage requirement and load power dissipation across variations. In addition, by considering coupled variations between the SC converter and the load integrated on the same die, our methodology can further reduce the pessimism in power-switch's resistance and capacitor density. The proposed co-design methodology is verified with a 2:1 SC converter and a digital load in a 65 nm. The resulted converter achieves *more than one order of magnitude reduction in the flying capacitor size* as compared to the conventional worst-case design while maintaining the target conversion efficiency and target throughput.

4. Key Accomplishments

During the last year, we have actively published seven papers in the target research area. We also submitted three new papers, one of which is based on a new chip prototype that we carried out in 2018. Below is the list of the publications during the 2018 project period.

5. Reference

[Kim, TVLSI18] Seongjong Kim, Joao Pedro Cerqueira, Mingoo Seok, "A Near-Threshold Spiking Neural Network Accelerator with a Body-Swapping based In-Situ Error Detection and Correction Technique," IEEE Transactions of Very Large Scale Integration Systems (TVLSI), 2018, *submitted*

[Yang, TVLSI18] Teng Yang, Doyun Kim, Jiangyi Li, Peter R. Kinget, Mingoo Seok, ``In-Situ and In-Field Technique for Monitoring and Decelerating NBTI in 6T-SRAM Register Files," IEEE Transactions of Very Large Scale Integration Systems (TVLSI), 2018

[Kim, JLPEA18] Seongjong Kim, Mingoo Seok, A Sub-50µm2, Voltage-Scalable, Digital-Standard-Cell-Compatible Thermal Sensor Frontend for On-Chip Thermal Monitoring," Journal of Low Power Electronics and Applications - Special Issue on CMOS Low Power Design, 2018

[Li, JSSC18] Jiangyi Li, Teng Yang, Minhao Yang, Peter R. Kinget, Mingoo Seok, "An Area-Efficient Microprocessor based SoC with an Instruction-Cache Transformable to an Ambient Temperature Sensor and a Physically Unclonable Function," IEEE Journal of Solid-State Circuits (JSSC), 2018, invited to the special issue

[Zhang, DAC19] Hao Zhang, Weifeng He, Yanan Sun, Mingoo Seok, "A Design Methodology of Energy-Efficient Logic Cell Library with Fine-Granularity of Driving Strength" ACM/IEEE Design Automation Conference (DAC), 2019, *submitted*

[Kim, CICC19] Doyun Kim, Peter R. Kinget, Mingoo Seok, "SRAM-ADC: SRAM Circuits Transformable to a Stochastic ADC at Ultra-Low Area Overhead," IEEE Custom Integrated Circuits Conference (CICC), 2019, *submitted*

[Li, ESSCIRC18] Jiangyi Li, Pavan Kumar Chundi, Sung Justin Kim, Zhewei Jiang, Minhao Yang, Joonseong Kang, Seungchul Jung, Sang Joon Kim, Mingoo Seok, ``A 0.78-µW 96-Ch. Neural Signal Processor Integrated with a Nanowatt Power Management Unit based on Energy-Robustness Co-Optimization Control," IEEE European Solid-State Circuits Conference (ESSCIRC), 2018

[Zhang, ISLPED18] Sheng Zhang, Adrian Tang, Zhewei Jiang, Simha Sethumadhavan, Mingoo Seok, "Blacklist Core: Machine-Learning Based Dynamic Operating-Performance-Point Blacklisting for Mitigating Power-Management Security Attacks," ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), 2018

[Kim, ISLPED18] Dongkwun Kim, Mingoo Seok, "Better-Than-Worst-Case Design Methodology for a Compact Integrated Switched-Capacitor DC-DC Converter," ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), 2018

[Seok, IRPS18] Mingoo Seok, Peter R. Kinget, Teng Yang, Jiangyi Li, Doyun Kim, “Recent Advances in In-situ and In-field Transistor-Aging and Compensation Techniques,” IEEE International Reliability Physics Symposium (IRPS), 2018, invited

[Kim, VLSI16] Seongjong Kim, Joao Pedro Cerqueira, Mingoo Seok, “A 450mV Timing-Margin-Free Waveform Sorter based on Body Swapping Error Correction," IEEE Symposium on VLSI Circuits (VLSI), 2016

[Yang, ISSCC15] Teng Yang, Doyun Kim, Peter R. Kinget, Mingoo Seok, “In-situ Techniques for In-field Sensing of NBTI Degradation in an SRAM Register File," IEEE International Solid-State Circuits Conference (ISSCC), 2015