POLITECNICO DI TORINO

Dipartimento di Automatica e Informatica

PhD in Computer and Control Engineering

XXXI cycle

Supervisor

Prof. Andrea Calimera

Prof. Enrico Macii

Energy-Efficient Speculative Computing for IoT ICs

PhD Candidate: *Roberto Giorgio Rizzo*

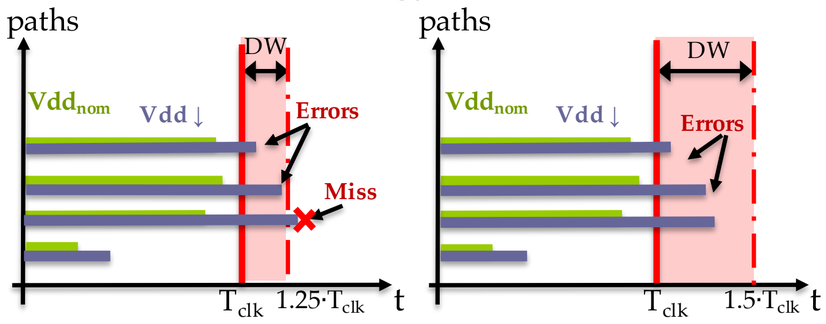
1. Context

The key success for the **Internet-of-Thing** (IoT) is the availability of always-on smart objects with embedded Integrated Circuits (ICs) that can process/transmit sensor data ceaseless. Due to limited budget of energy made available by small batteries, such ICs must show ultra-high energy efficiency thus to guarantee reasonable throughput. Overcoming classical low-power techniques, **Speculative Computing** trades Quality-of-Results (QoR) for energy exploiting error resilience of *Data-intensive* applications.

Unlike the state-of-the-art methodologies (like *Razor-based speculation*), two unconventional knobs for speculative computing were explored: **Timing-Speculation via Approximate Error Detection-Correction (AED-C)** and **Functional-Speculation via Inferential-Circuits (InfCs)** for efficient *Adaptive Voltage Over-Scaling* (AVOS).

2. Timing-Speculation via AED-C

The main intuition behind this technique [1][2][3][4] is that the availability of an AED-C mechanism represents a smart option to control the QoR-energy tradeoff.



(a) Aggressive AED-C (b) Slow AED-C

Figure 1. AED-C Working Principle

The proposed AED-C scheme leverages the resolution of in-situ timing error detection mechanisms as a knob to dynamically accelerate (Fig. 1a) or slow-down (Fig. 1b) AVOS, thus to achieve lower energy consumption or higher QoR.

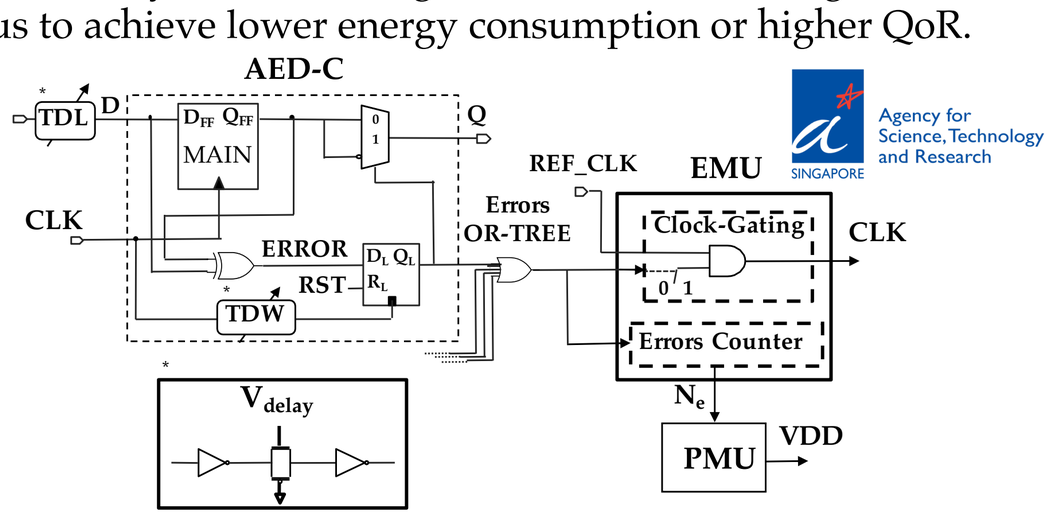
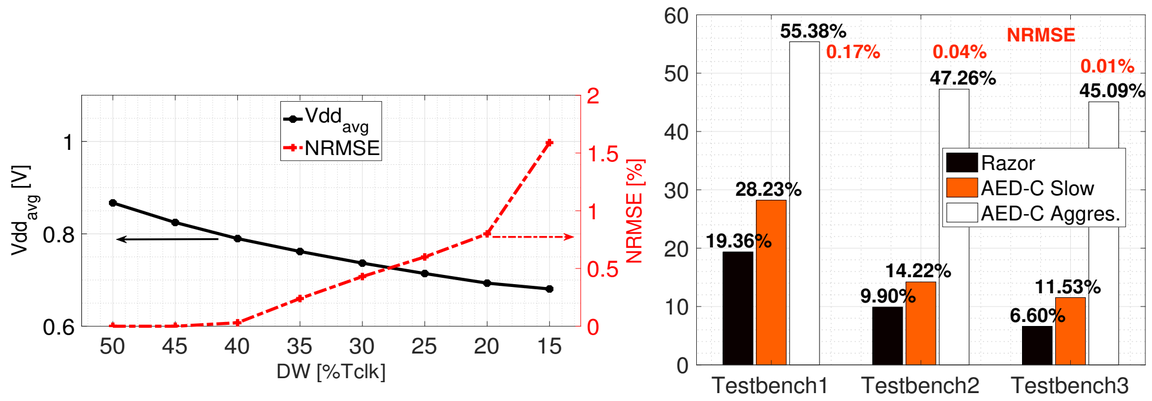


Figure 2. Error detection and logic masking circuitry in AED-C

An *Error Management Unit* (EMU) implements an error-driven clock-gating enabling the error logic masking, while *a Power Management Unit* (PMU) uses the Error Rate (N_e) to handle voltage scaling (refer to Fig. 2). A parametric analysis on a set of representative benchmarks discloses AED-C efficiency (Fig. 3), providing an assessment of the QoR-energy trade-off. Also, when applied to a real-life application, i.e., FDCT into a JPEG compressor, aggressive AED-C shows 51.9% energy savings (vs. baseline FDCT) and a PSNR of 48.5dB (vs. baseline JPEG). Worst and best quality images are shown in Fig. 4a and 4b.



(a) MAC: Vdd_{avg} / $Error_{avg}$ vs. TDW (b) FIR Filter: Energy savings

Figure 3. Parametric Analysis on AED-C applications

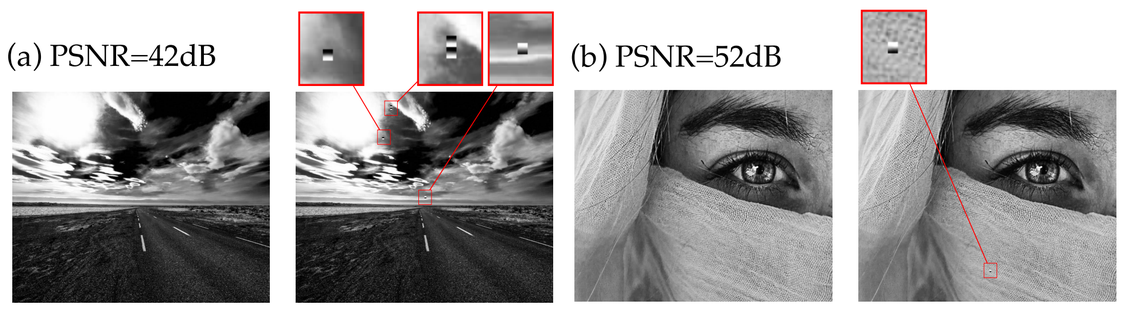


Figure 4. AED-C for FDCT in JPEG compression

3. Functional-Speculation via InfCs

InfCs, as design-level speculative architectures, are able to skip arithmetic operations inferring output values by evaluating the key features of a function learned during a training stage, i.e., *Machine Learning* (ML). Since InfCs leverage statistical inference engine as *Classification Trees* (CT), an additional stage for CT building is integrated into the standard synthesis flow (Fig. 5).

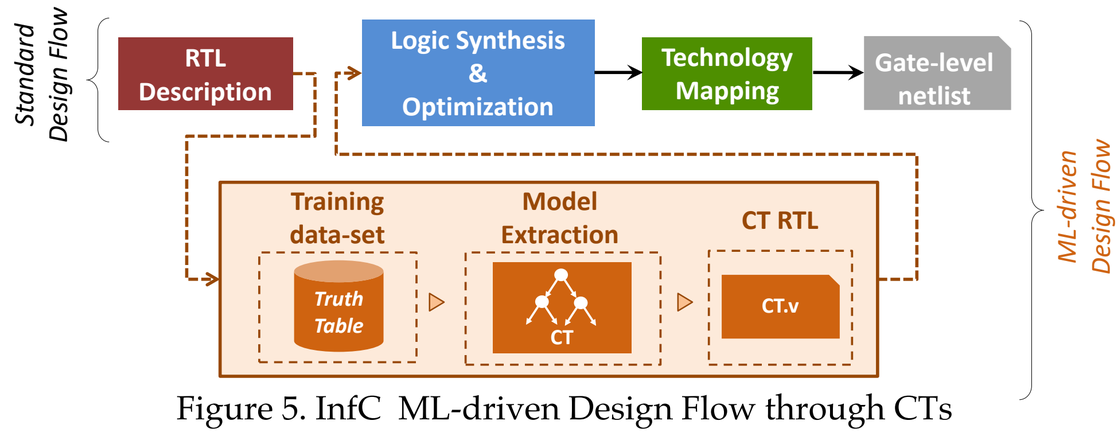


Figure 5. InfC ML-driven Design Flow through CTs

Being inference faster than arithmetic computation, InfCs show larger margins to push VDD under the "safety" point, trading QoR for energy. As a case study, we implemented an **Inferential Multiplier** (I-MULT): it presents a 22% area savings being 2.2x faster than a Booth Mult. [5]. We tested the efficiency of I-MULT on *Image Blending* application, over a set of 56 blended images. I-MULT enables an aggressive VOS pushing the VDD from 1.10V to 0.64V, with 80% power savings (w.r.t. Booth Mult.) at the cost of 5.4% average error (NRMSE).

4. References

1. R.G. Rizzo et al. "Early Bird Sampling: a Short-Paths Free Error Detection-Correction Strategy for Data-Driven VOS", Proc. VLSI-SoC, 2017
2. R.G. Rizzo et al. "Tunable Error Detection-Correction for Efficient Adaptive Voltage Over-Scaling", Proc. NGCAS, 2017
3. R.G. Rizzo et al. "On the Efficiency of Early Bird Sampling (EBS) An Error Detection-Correction Scheme for Data-Driven Voltage Overs-Scaling", VLSI-SoC Book, Springer, 2018 (In Press)
4. R.G. Rizzo et al. "Approximate Error Detection-Correction for Efficient Adaptive Voltage Over-Scaling", Integration, the VLSI Journal, 2018
5. R.G. Rizzo et al. "Multiplication by Inference using Classification Trees: A Case-Study Analysis", Proc. ISCAS, 2018