Poster Abstract: Energy vs. Quality of Approximate Non-volatile Writes in Intermittent Computing

Rei Barjami
Politecnico di Milano
Italy

Antonio Miele
Politecnico di Milano
Italy

Luca Mottola
Politecnico di Milano and RI.SE
Italy, Sweden

ABSTRACT

We explore how hardware approximation techniques can be used to reduce the overhead introduced by state persistence operations in intermittent computing. We do so by exploring the trade-off between energy consumption and quality of the results. We specifically adjust the energy/quality ratio of write operations in Spin Transfer Torque Magnetic Random Access Memory (STT-MRAM) by modifying the current applied during these operations. Our evaluation on a heterogeneus set of benchmarks demonstrates up to ≈50% reduction in the state persistence overhead while maintaining an acceptable output quality.

KEYWORDS

Non volatile memory, Approximate computing, Energy-harvesting

1 OVERVIEW

Energy harvesting allows embedded devices to mitigate their dependency on traditional batteries. They do so by relying on sources such as solar radiations and thermal gradients. However, ambient energy sources are unpredictable, leading to frequent energy failures [2]. To maintain program forward progress across these failures, energy harvesting devices store their state in Non Volatile Memory (NVM). This operation largely represents an overhead, which does not directly contribute to executing application logic and is thus detrimental to the system's energy efficiency.

We explore how approximate computing techniques can be used to reduce the overhead of NVM writes in intermittent computing while accepting a degradation in the accuracy of the obtained results. Unlike prior work [1, 3], we concentrate on techniques that work close to the hardware.

We specifically exploit the stochastic switching property of Spin Transfer Torque Magnetic Random-Access Memory (STT-MRAM) [4]. *By piloting the current used for writing* in STT-MRAM, we modify the energy it consumes while altering the Write Error Rate (WER). We thus explore the trade-off between saving energy by reducing

Benchmark	Domain	Pipeline	Quality Metric
FFT	Signal Processing	No	Avg. Rel. Error
PicoJpeg	Image Processing	No	RMSE
Susan	Image Processing	Yes	Precision and Recall
Only Writes	Micro Benchmark	No	-

Table 1: Benchmarks from MiBench2.

the amount of current during write operations and the errors we consequently introduce in the written data, which ultimately affect the program's Quality Of Results (QoR) [8].

The aim of our research is therefore: *i)* quantifying the energy savings achieved through approximated NVM writes, *ii)* analyzing the impact of approximation on data processing and QoR, and *iii)* assessing the feasibility of substantial reductions in NVM write energy consumption while maintaining acceptable QoR. We apply this technique to the widely-used MSP430 MCU, hooked to an STT-MRAM NVM. Using a diverse set of benchmarks, we reveal a ≈50% reduction in NVM write overhead while preserving an acceptable QoR.

2 EVALUATION

We describe the setting we use for the evaluation and the results we obtain.

2.1 Setting

Platform and simulation. Our hardware comprises an MSP430 MCU utilizing STT-MRAM as NVM. We consider three different versions of MSP430 MCU that expose different performance vs. energy trade-off: the MSP430Lxxxx and MSP430Gxxx, along with a custom version designed by Singhal et al. [9]. Since the bit-switching probability of STT-MRAM during write operations varies with write current, we define five quality levels, ranging from Q0 with almost 100% bit-switching probability, to Q4 with higher energy efficiency but a WER of 10^{-3}.

We build a simulation framework by integrating NVSim [5] to estimate STT-MRAM energy consumptions with MSPSim [6] to emulate the MSP430 MCU and measure the performance.

Benchmarks and metrics. In Table 1 we show the set of benchmarks we select from the commonly used MiBench2 [7] suite for evaluation. We choose a heterogeneous set of benchmarks to cover various aspects relevant to the evaluation, while keeping into account resource constraints of the target platform that may prevent certain benchmarks from running in the first place, for example, due to memory limitations.

We concentrate on the following aspects: *i)* whether the benchmark is amenable to approximation or it requires 100% accurate

Quality	WER	Set Current (μA)	Write Energy x bit (pJ)
Q0	10^{-8}	1153	167
Q1	10^{-6}	865	94
Q2	10^{-5}	769	74
Q3	10^{-4}	673	57
Q4	10^{-3}	577	43

Table 2: Energy consumption for various quality levels in STT-MRAM writes.

computations, and *ii)* whether the benchmark may be considered a pipeline of subtasks or a single task. For each benchmark, we select a quality metric to assess the QoR of the approximate execution in comparison to the correct execution. This metric is necessarily application-specific.

2.2 Results

Memory characterization. Table 2 summarizes energy consumption for various quality levels in STT-MRAM writes. Energy savings are significant, with Q4 consuming only 25% of the energy in the correct case. As expected, energy reduction is more pronounced at lower quality levels, becoming less significant at higher levels.

Energy consumption. Figure 1 reports the total energy consumption for each benchmark using the MSP430Singhal, normalized to the benchmark executed with Q0. Results vary widely between benchmarks, reflecting the extent of NVM utilization. Benchmarks that rely heavily on NVM show substantial energy savings, while those primarily using volatile memory enjoy lower energy savings.

For instance, the `Only Writes` microbenchmark that only includes write operations and may be considered as a baseline of sorts, does experience a significant energy reduction with Q4 consuming only 30% of the energy of Q0. In contrast, FFT computation exhibits limited energy savings, particularly with higher approximation levels. `Susan Edge Detection` demonstrates notable energy savings at Q3, reducing energy consumption by approximately 34%.

Quality vs. energy trade-off. The energy savings come at the cost of a reduction in the quality of the output. The discussion of such trade-off is necessarily application-specific.

Consider `PicoJpeg`, where we measure the QoR with the RMSE of the resulting picture. The trends are shown in Figure 2. At higher
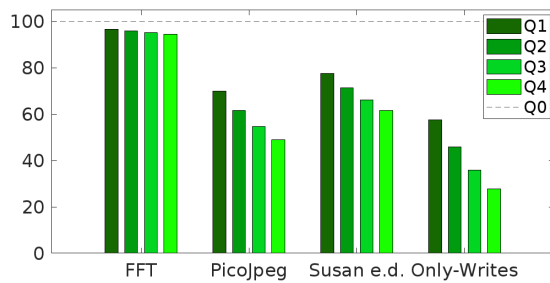

Figure 1: Energy consumption with various quality levels, normalized to energy consumed with Q0.
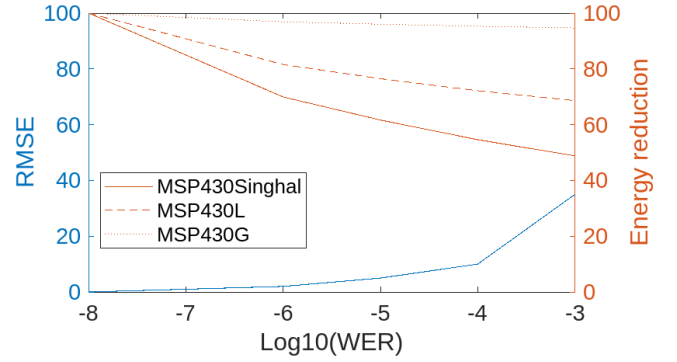

Figure 2: Quality vs. energy trade-off for `PicoJpeg`.

approximation levels, the QoR decreases linearly, with the RMSE increasing correspondingly. Energy reduction is most pronounced at lower approximation levels. The best trade-off is arguably around Q3, which corresponds to a significant reduction in energy consumption for two platforms out of three and a limited loss of quality in the output. The system using the MSP430G is the worst-performing in this case, as the MCU dominates energy consumption anyway. In contrast, the system equipped with the MSP430Singhal gains the most benefits from more efficient NVM writes as the MCU is also the most energy efficient, and thus NVM operations dominate the energy figure.

Similar considerations apply to `Susan Edge Detection`. Errors in the pipeline stages impact the QoR, resulting in false positives and false negatives. Precision and recall decrease almost linearly with increasing approximation levels. Q3 provides again a balance between energy savings and QoR degradation, reducing energy consumption by about 34% while limiting the errors.

FFT behaves differently. Energy consumption decreases, but the percentage of unacceptable results increases drastically with higher approximation levels. Errors are too large beyond Q1, where, however, the energy savings are limited.

REFERENCES

[1] Bambusi et al. 2022. The case for approximate intermittent computing. In *ACM/IEEE IPSN*.
[2] Bhatti et al. 2016. Energy harvesting and wireless transfer in sensor network applications: Concepts and experiences. *ACM Transactions on Sensor Networks* 12, 3 (2016).
[3] Adriano Branco et al. 2019. Intermittent asynchronous peripheral operations. In *ACM SENSYS*.
[4] Devolder et al. 2008. Single-shot time-resolved measurements of nanosecond-scale spin-transfer induced switching: Stochastic versus deterministic aspects. *Physical Review Letters* 100, 5 (2008), 057206.
[5] Dong et al. 2012. Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 31, 7 (2012).
[6] Eriksson et al. 2007. MSPSim–an extensible simulator for MSP430-equipped sensor boards. In *EWSN*.
[7] MiBench2. 2019. MiBench2: A benchmark suite for micro-architectural research. https://github.com/impedimentToProgress/MiBench2.
[8] Monazzah et al. 2020. CAST: content-aware STT-MRAM cache write management for different levels of approximation. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 39, 12 (2020).
[9] Singhal et al. 2015. 8.3 A 10.5 μA/MHz at 16MHz single-cycle non-volatile memory access microcontroller with full state retention at 108nA in a 90nm process. In *2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers*. IEEE.