Xor-Masking: A Novel Statistical Method for Instruction Read Energy Reduction in Contemporary SRAM Technologies

Citation

Year
2016

Version
Peer reviewed version (post-print)

Link to publication
TUTCRIS Portal (http://www.tut.fi/tutcris)

Published in
2016 IEEE International Workshop on Signal Processing Systems (SiPS)

DOI
10.1109/SiPS.2016.19

Copyright
© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Take down policy
If you believe that this document breaches copyright, please contact tutcris@tut.fi, and we will remove access to the work immediately and investigate your claim.

Download date: 15.04.2018
Xor-Masking: A Novel Statistical Method for Instruction Read Energy Reduction in Contemporary SRAM Technologies

Joonas Multanen, Timo Viitanen, Pekka Jääskeläinen, Jarmo Takala
Tampere University of Technology, Finland
Email: joonas.multanen@tut.fi, timo.2.viitanen@tut.fi, pekka.jaaskelainen@tut.fi, jarmo.takala@tut.fi

Abstract—Pervasive computing calls for ultra-low-power devices to extend the battery life enough to enable usability in everyday life. Especially in devices involving programmable processors, the energy consumption of integrated memories often plays a critical role. Consequently, contemporary memory technologies focus more on the energy-efficiency aspects with new custom CMOS SRAM cells with tailored energy consumption profiles constantly being proposed.

This paper proposes a method that exploits such contemporary low power SRAM memories that are energy optimized for storing a certain logic value to improve the energy-efficiency of instruction fetching, a major energy overhead in programmable designs. The method utilizes a low overhead xor-masking approach combined with statistical program analysis to produce optimal masks to reduce the occurrence of the more energy consuming bit values in the fetched instructions.

In comparison to the “bus invert” technique typically used with similar SRAMs, the proposed method incurs minimal area overhead while still reducing the total energy consumption of an example LatticeMico32 core up to 5%. The improvement to instruction memory energy consumption alone is up to 13% with a set of benchmarks.

I. INTRODUCTION

With the emerging Internet of Things (IoT), more and more devices are becoming wearable, intelligent and wireless. Having small, usable devices that can process data, communicate wirelessly and operate for extended periods of time without external power source creates challenges for battery and processor technologies. Improved intelligence in devices often results in increased battery consumption due to the additional required digital logic. Since limitations of the current battery technologies inflict a hard upper bound on device usage time without having an external power source available, low-power and energy-efficient operation is an important requirement for embedded devices.

While processing devices are getting smaller, the amount of data being processed is growing, which makes power consumption of memories a critical aspect in ultra-low-power compute devices. It is not rare for on-chip memories to be responsible for half of the power consumption in CPU-based digital designs [1], [2]. In order to keep memories up to speed with other components, new technologies such as Spin-Transfer Torque RAM (STT-RAM) and Embedded DRAM (eDRAM), are studied as possible replacements to SRAM. However, these technologies are not yet mature, and process challenges such as cost, durability, and variability control must be overcome before they can be widely adopted. Unlike SRAM, they are dedicated-process [3] technologies, where significant modifications to the fabrication process need to be performed.

According to ISSCC, SRAM is still the technology of choice for fast on-chip memories and caches. However, also within the space of SRAM designs there is room for optimization. Traditional cell designs do not consider that instruction and program data are often biased towards containing either high or low logic values. If statistical information about the data to be stored exists, memory cells can be designed in an asymmetric fashion, so that they consume less power when holding, reading or writing the more often occurring logic value in that data, allowing reduction of the total energy consumed. These asymmetric cells have been designed and fabricated for memories and caches [4]–[7].

This paper takes under closer inspection the asymmetric SRAM designed and fabricated by Mori et al. [6], where reading the logical value one results in lower energy consumption compared to the value zero. In order to reduce the total energy consumption, the authors added a majority voting logic to increase the amount of logical ones stored, and thus read from their proposed memory.

In this paper we show that this type of asymmetric SRAMs can be beneficially exploited also for program instruction memories in statically scheduled processor cores by adding a very small area hardware logic, and utilizing offline program binary analysis. The paper is organized as follows. Section II reviews previous methods for low-power instruction and data encoding. Section III introduces the proposed method. It is evaluated and compared to an existing low-power encoding in Section IV. Section V concludes the paper.

II. RELATED WORK

Previous work on low-power encoding has mostly concentrated on instruction address bus power and program data bus power reduction. The methods can be divided into static and dynamic methods, depending of the time the encoding is performed. Static encoding happens at program compile time, and dynamic at runtime using additional hardware logic.
Bus-invert encoding [8] was introduced by Stan and Burleson in 1995. They presented the encoding as a method to lower off-chip data bus power consumption. Later work has applied the encoding to segments of divided buses [9], [10]. In the bus-invert method, data is dynamically re-encoded based on the number of toggling bits between two consecutive data words (also known as the Hamming distance). If the number of toggling bits is more than half of the total bits, the data word is inverted with a logical NOT operation. This is done using a logic connected to the memory block, at the end of the bus. At the other end of the bus, where the data is consumed (typically a load-store unit or an instruction fetch), NOT is performed again to restore the original word, in case indicated by an extra control bit. The most apparent drawback of this technique is the additional control bit that needs to be stored per each word in the memory, and transferred to the consumers.

Petrov and Orailoglu [11] proposed a static method where the instruction data bus was encoded with 16 possible data transformation operations. The optimal set of transformations were found by using an exhaustive search considering two consecutive instruction words at a time. The encoded words were stored in memory when loading the program in, and the transformations to perform were communicated to the processor decode unit either in the beginning of the program or before application hot-spots, such as loops. This method needs frequently executed loops to perform well, also the reduction of toggled bits is better with smaller block sizes in comparison to large ones. In terms of energy consumption, the benefits from reduced successive word toggle activity can be significant when applied to an off-chip DRAM, or when the words are transferred via a long on-chip bus. However, the effect of bit toggles in the case of mostly reading data from a SRAM residing close to the consumer, a typical scenario with on-chip instruction memories, is small. In addition, the energy overhead of the additional decoding logic to implement the 16 transformations was omitted from the analysis.

Su et al. [12] used Gray coding and Cold scheduling to reduce bit switching on the instruction address bus. Gray coding exploits the fact that instructions are often fetched from consecutive memory addresses. By reordering the instructions in memory so that sequential addresses are located in Gray coded addresses, the instruction address bus toggling can be reduced. The authors combined Gray coding with Cold scheduling, where the program compiler schedules instructions using their relative energy costs to each other to minimize the total energy.

Musoll et al. [13] introduced an encoding based on the observation that usually programs operate on sets of same addresses, or working zones, repeatedly at a time. Like in Gray coding, this can reduce the address bus toggling. Working zones are given identifiers and these along with address offsets are sent to the instruction control logic to signal which working zone to use. This method can be expensive logic-wise, since the encoding and decoding algorithms are complex.

Benini et al. [14] analyzed instruction address traces to create custom encoder and decoder logic for a given processor. According to the authors, the method is generic and can be applied to various processors. This method showed little savings in the total energy when taking into account the energy required by the additional encoding and decoding logic.

Regarding instruction memory power consumption optimization, previous work mostly concentrates on the address bus power reduction, and, in particular, minimizing its bit switching activity. Many of the original methods focus on off-chip data bus energy reduction and disregard the energy overhead of the extra logic needed for encoding and decoding. In contemporary compute devices, the instruction data bus energy consumption might not be relevant due to instruction memories or caches integrated close to the consumer, the instruction fetch unit. On the other hand, in ultra-low-power designs, the energy and area overhead of the additional encoding and decoding logic required might be an important factor, since a complex implementation of the method can reverse the achieved gains.

Unlike previous solutions, our proposed method is able to utilize statistical analysis of the individual bit positions in instruction execution to produce an optimal, low-power encoding during compile-time with minimal required additional dynamic decoding logic.

### III. Proposed Method

Most of the existing bus encoding algorithms concentrate on instruction address bus and program data bus encoding. They aim to reduce the bit switching activity on the buses, whereas the proposed method maximizes the occurrence of one logical value over the other. For the instruction address bus, the existing approaches exploit the fact that instruction addresses are often accessed in a sequential order, where techniques such as Gray coding the addresses can help. The fetched instruction data, on the other hand, is usually less sequential – the individual bits do not typically correlate with the previously fetched instruction word.

The proposed method we call xor-masking, uses statistical information about instruction memory accesses for individual programs to statically determine an optimal xor-mask to encode the instructions in each particular program to maximize the appearance of the desired logical value. The method is intended for designs in microcontroller or signal processing, where no operating system is used. This allows simpler handling of the xor-mask, since no dynamically linked code, system calls or context switches are used.

This encoding logic was used dynamically by Mori et al. [6] for general data. It resembles the original bus-invert logic, where the inverting decision is based on the Hamming distance between words.

The Hamming distance for two words of data can be calculated by taking the logical XOR between the words, and counting the ’1’ bits of the result. This is presented in Fig. 1. Here the previous (top) word has been inverted and we calculate the Hamming distance to the next word, including the added bit. In this case, we would invert the next (bottom) word, since the Hamming distance is greater than half of the word length. Counting the bits can be done with majority voters. Their implementation cost depends on the width of the data, and the type of the voter [15]. The decoding is done by XORing a given word with its added bit. For their SRAM, Mori et al., modified the original bus-invert by not minimizing the amount of toggles, but maximizing the amount of ‘1’s. This corresponds to calculating the Hamming distance with the other word constantly being all ‘1’s. This simplified the logic,
since the previous instruction or data value was not compared to the current one, thus saving a register and a logical XOR between the words. In this paper, this method is referred to as Majority Voter Encoding (MVE).

Encoding words for low power like this is not ideal, when the SRAM is used as an instruction memory. Statistical instruction analysis, such as in the proposed method, can lead to better maximizing of the low read-energy logic values in the memory. The analysis (see Fig. 3) starts by taking the instruction memory bit image of the optimized program. An instruction address trace with an unmodified instruction memory image is produced by simulating the targeted program using typical input data. Then, for each bit index, the total amount of logical ‘1’s and ‘0’s is weighted according to execution count is calculated. This process is illustrated in Fig. 2. If the amount of ‘0’s is greater than half of total bits in the word, a ‘1’ is assigned to that index of a xor-mask that is applied to the instruction word. That is, all the bits in that index will be inverted in the memory. The xor-mask is created for each application separately. For the memory examined in this paper, logic ‘1’ reads are preferred for low energy. However, the method can equally efficiently be applied to cases, where logic ‘0’ is the less energy consuming alternative.

Like the static instruction transformation of Petrov and Orailoglu [11], the proposed method conveys application-specific information to the processor. In our case, this information is reduced to a mere xor-mask, which is used to decode the encoded words during the instruction fetch stage. Fig. 4 depicts the logic required. Decoding the words is done by performing a logical XOR between the mask word and each encoded word. In addition to incurring extremely little additional logic, this allows easy implementation of the method to existing architectures.

The update of the XOR mask can be integrated with the program image, so that after reset the first instruction address holds the xor-mask. This is fetched and treated as a mask word which is stored to the mask register. After this, the succeeding words are treated normally as instructions. For even easier integration with program compilers, the mask word and a jump to reset address can be placed to the last addresses in instruction memory to avoid relocating the program instructions. For real world applications, program code for the static analysis should be weighted realistically and simulated with a typical input. If the simulated execution does not match a realistic scenario, the resulting suboptimal mask could cause an increase in energy consumption.

The decoding logic overhead in the proposed method is very low. Moreover, encoding the words incurs no additional hardware logic, since the encoding is performed offline as a final pass over the instruction bit image.

As a side effect of an encoding that aims to maximize one logical value over another, the proposed method often reduces bit toggling between successive instruction words. In case of a multi-level instruction memory hierarchy, the reduced toggling can save additional energy since off-chip buses can consume considerable amounts of energy.

IV. EVALUATION

Xor-masking was implemented on LatticeMico32 processor [16]. LatticeMico32 is a RISC architecture with an open licensing agreement and freely available RTL source code. Detailed specifications of the evaluated processor are listed in Table III. The evaluation platform was a minimal setup, where instruction and data memories were scaled to be large enough to accommodate all of the individual benchmark program instructions and data at a time. The benchmark programs were chosen to represent typical applications in microcontrollers in low-power scenarios. The benchmarks are listed in Table IV.
The energy consumption was evaluated by calculating energy costs for reading a single ‘1’ and ‘0’ based on the previously published measurements for the original SRAM [6]. In the original 64kB SRAM, for a 16-bit word, energy consumption for reading all ‘0’s and reading all ‘1’s was 1440 fJ/cycle and 148.5 fJ/cycle, respectively. Dividing these by 16 yielded 90.00 fJ/cycle and 9.28 fJ/cycle for the cost of reading a single ‘1’ and a ‘0’ bit, respectively. Next, instruction address traces for the benchmark programs were produced from Modelsim simulations. Using the calculated read energies, instruction traces and instruction memory bit images encoded according to each of the two methods, the total SRAM energy consumption for each benchmark program was calculated. These are presented in Table I.

Considering the write energy consumption is relevant for processors with an instruction cache. Cache misses translate to writing cache lines. The difference in energy reduction for writing all ‘0’s and all ‘1’s to the referred SRAM, compared to a regular SRAM, was reported as negligible, 1%. This is illustrated in Fig. 5. The proposed method does not affect the time of replacing cache lines compared to the referred method. In this sense, write energy evaluation is not interesting, since the difference between the proposed method and the referred one is not significant.

If both the instruction memory and the instruction cache are implemented with the referred SRAM technology, the total energy reduction depends on the combined energy consumption of the two. In this case, the decoding logic would be implemented after the instruction cache.

In the case of the proposed method, the numbers also include the energy consumed by the majority logic proposed with the referred SRAM technology. In our case, this majority logic is not used and instead, the decoding logic would be implemented in the instruction fetch unit. This overhead is present in our energy numbers for the proposed method. The improvement in energy consumption of our proposed method compared to the referred method would, therefore, be better than the results presented in this paper. However, this overhead is difficult to estimate, since the authors did not report the majority logic energy consumption individually, but rather the overall SRAM’s. Moreover, the majority logic relies on a pull-down network and a sense amplifier. However, regardless of this overhead, our proposed method still achieves lower energy consumption in all 11 benchmark programs.

Energy comparison normalized to MVE for the benchmark programs is presented in Fig. 8. As is expected, the energy consumption depends on the dynamic instruction mix in each of the benchmarks. The more there are instructions resembling each other on the bit level, the more the proposed method can save energy. The worst case is when the occurrence of ‘1’ and ‘0’ at each bit position is exactly the same. In this case, inverting the bit index results in no savings in energy.

The energy overhead of the added logic is small compared to the overall energy, on average 1.0% of the SRAM energy. The best energy reduction, 13.3%, was achieved in matrix benchmark and the lowest reduction, 1.5%, in coremark. On average, the reduction was 6.2%. To estimate the CPU total energy consumption, the described LatticeMico32 was synthesized on a 28nm ASIC standard cell technology. The instruction memory consumed 37.7% of the total energy after synthesis for this particular implementation. Using this number, the effect on the total CPU energy was calculated. This is presented in Fig. 6. The largest total energy reduction, 24.8%, was achieved with fir. In matrix, total CPU energy consumption was 5.0% less compared to MVE. On average, this reduction was 2.4%.
The reduction in bit switching activity is compared to the original instruction words and the majority-voter-encoded words. This is presented in Table II. Both of the encoding methods add a toggle bit to the unencoded words, increasing the total amount of bits read. The proposed method reduces the bit switching activity in all but two benchmark programs compared to the MVE. The reductions are small and in the best case 3.3%. In a realistic implementation in an ultra-low-power IoT device, off-chip memory would be unlikely to be used for storing instructions and the effect in energy reduction for an on-chip bus would be negligible.

To evaluate the possible energy reduction from using multiple masks, basic blocks for the benchmark programs were formed from the instruction traces. A xor-mask was computed for each basic block. Then, energy consumption was again calculated using the numbers for the referred SRAM. The results are presented in Fig. 7. The numbers do not include the energy overhead from mask updating. The results seem encouraging, and suggest that further investigation could be beneficial, since theoretically, up to 74% reduction could be achieved (in coremark benchmark).

V. CONCLUSIONS

Energy consumption of on-chip and off-chip memories offers optimization opportunities in pervasive compute devices. In this paper, a novel statistical method, xor-masking was proposed to reduce the instruction fetch energy consumption in asymmetric SRAM technologies. The proposed method was evaluated on LatticeMico32 RISC soft-core with 11 benchmark programs.

Including the energy overhead of the decoding logic, the proposed method consumes up to 13% less energy compared to the state-of-the-art majority voter encoding on the same SRAM. The total CPU energy reduction is up to 5% compared to majority voter encoding. The energy consumption with the proposed method was smaller in all benchmark programs compared to the majority-voter-encoding. In addition, the proposed method reduces instruction data bus toggling up to 3.3% compared to the referred method.

Initial evaluation of using multiple xor-masks per program suggest further possibilities for energy reduction. Future work involves researching the use of multiple masks for a given program. This involves at least investigating the granularity of the masking (basic block/instruction level), number of masks and mask updating strategies for maximal energy reduction. Careful consideration of the mask updating overhead is required.

VI. ACKNOWLEDGMENTS

The authors would like to thank the Finnish Funding Agency for Technology and Innovation (project "Parallel Acceleration 3", funding decision 1134/31/2015) and ARTEMIS JU under grant agreement no 621439 (ALMARVI).

REFERENCES


