Energy Efficient and Reliable Inference in Nonvolatile Memory under Extreme Operating Conditions

SALONIK RESCH, University of Minnesota, Twin Cities, USA
S. KAREN KHATAMIFARD, University of Minnesota, Twin Cities, USA
ZAMSHED I. CHOWDHURY, University of Minnesota, Twin Cities, USA
MASOUD ZABIHI, University of Minnesota, Twin Cities, USA
ZHENGYANG ZHAO, University of Minnesota, Twin Cities, USA
HUSREV CILASUN, University of Minnesota, Twin Cities, USA
JIAN-PING WANG, University of Minnesota, Twin Cities, USA
SACHIN S. SAPATNEKAR, University of Minnesota, Twin Cities, USA
ULYA R. KARPUZCU, University of Minnesota, Twin Cities, USA

Beyond edge devices can operate outside the reach of the power grid and without batteries. Such devices can be deployed in large numbers to regions which are difficult to access. Using machine learning, these devices can solve complex problems and relay valuable information back to a host. Such devices have been proposed for use as nano-satellites, where many devices are deployed into low earth orbit. Due to the harsh and unpredictable environment, devices must be highly energy efficient, capable of operating intermittently, and be tolerant of radiation. Here, we propose non-volatile processing-in-memory (PIM) architecture which is capable of extreme energy efficiency, has low overhead checkpointing mechanisms, can operate at wide range of temperatures, and has a natural resilience to radiation.

Additional Key Words and Phrases: processing in memory, beyond edge computing

ACM Reference Format:
Salonik Resch, S. Karen Khatamifard, Zamshed I. Chowdhury, Masoud Zabihi, Zhengyang Zhao, Husrev Cilasun, Jian-Ping Wang, Sachin S. Sapatnekar, and Ulya R. Karpuzcu. 2021. Energy Efficient and Reliable Inference in Nonvolatile Memory under Extreme Operating Conditions. *ACM Trans. Embedd. Comput. Syst.* 37, 4, Article 111 (August 2021), 32 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Beyond edge devices collect energy from the environment, allowing them to operate off the grid and without a battery [11, 55]. This enables them to function in environments which would otherwise be impossible, such as in the remote wilderness [82], within the human body [39], and out in

Authors' addresses: Salonik Resch, resc0059@umn.edu, University of Minnesota, Twin Cities, 200 Union St SE, Minneapolis, Minnesota, USA, 55455; S. Karen Khatamifard, khatami@umn.edu, University of Minnesota, Twin Cities, 200 Union St SE, Minneapolis, Minnesota, USA, 55455; Zamshed I. Chowdhury, University of Minnesota, Twin Cities, 200 Union St SE, Minneapolis, Minnesota, USA, 55455; Masoud Zabihi, University of Minnesota, Twin Cities, 200 Union St SE, Minneapolis, Minnesota, USA, 55455; Zhengyang Zhao, University of Minnesota, Twin Cities, 200 Union St SE, Minneapolis, Minnesota, USA, 55455; Husrev Cilasun, University of Minnesota, Twin Cities, 200 Union St SE, Minneapolis, Minnesota, USA, 55455; Jian-Ping Wang, University of Minnesota, Twin Cities, 200 Union St SE, Minneapolis, Minnesota, USA, 55455; Sachin S. Sapatnekar, University of Minnesota, Twin Cities, 200 Union St SE, Minneapolis, Minnesota, USA, 55455; Ulya R. Karpuzcu, University of Minnesota, Twin Cities, 200 Union St SE, Minneapolis, Minnesota, USA, 55455.

1539-9087/2021/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

space [73]. This capability opens up many new possible applications. Running machine-learning algorithms on beyond edge devices is particularly attractive to due its versatility [37]. Utilizing neural networks (NN) or support vector machines (SVM), such devices will be able to solve a wide variety of problems.

However, engineering devices to operate beyond the edge is difficult. As they must collect energy from their environment, the power source will be unreliable. The devices must frequently power off, turning back on when energy is available. This is referred to as *intermittent* computing, which comes with a performance and energy efficiency overhead. In order to prevent total loss of information during intermittent operation, beyond edge devices must do additional work in 3 categories [32]: 1) Backup: Save data and the current architectural state (writes to non-volatile memory), 2) Dead: Re-perform work which was not able to be saved on the last shutdown, and 3) Restore: Re-start the device after a shutdown. Beyond the additional latency and energy overheads, intermittency also makes it a challenge to guarantee correctness of a program. Power interuptions can introduce memory inconsistencies which lead to incorrect operation [16, 73]. For conventional embedded systems, sophisticated software strategies are required to ensure that an interruption at any point does not induce corruption [37, 95].

Previous work has shown that non-volatile processing-in-memory (PIM) architectures are promising for use in beyond edge devices. MOUSE [93] is a PIM architecture which demonstrated high performance, extreme energy efficiency, and low complexity checkpointing mechanisms. It has 3 advantages over traditional architectures.

(1) Inherently intermittent safe logic operations
(2) Automatic and instantaneous data backup
(3) Highly energy efficient and highly parallel operations

Advantage 1 enables MOUSE to simplify its checkpointing strategies. Operations performed in the memory can be interrupted or performed multiple times without introducing corruption. Hence, data remains consistent as long as operations are performed sequentially. Advantage 2 reduces overhead for checkpointing. Normally, a device will have to write volatile data back to memory to save progress before shutdown. Since MOUSE does all its computation in non-volatile memory, progress is saved automatically after every operation. Finally, Advantage 3 enables high performance within low power budgets.

Recently, there has been much excitement about the use of beyond edge devices as nano-satellites [74] deployed in low earth orbit (LEO). Such devices could provide valuable services such as agricultural monitoring [114], security, and environmental and structural monitoring [74]. Nano-satellites can be much more cost effective than traditional monolithic satellites. However, orbital deployment for use as a satellite further increases the challenge of engineering such devices. One difference is that the cost of communication is much greater than computation (even more so than for terrestrial deployment) [37, 74]. This shifts emphasis towards performing more computation and holding more data on the device, and away from frequent communication [26]. MOUSE is well-suited for this challenge as it has a large memory capacity (due to consisting nearly entirely of high density non-volatile memory), enabling it to potentially go long periods of time and store more results before attempting transmitting.

An additional challenge for beyond edge devices deployed in LEO is that they must operate at a wide range of temperatures. Satellites can get both very cold (-170C) and very hot (123C) [65]. Large scale satellites can be engineered to perform temperature modulation [8]. However, small, cheap beyond edge devices may not be able to use such strategies. Fortunately, CMOS can perform well across this wide temperature range, and the performance of CMOS circuit actually tends to

increase with decreasing temperature [105, 122]. However, cold operation can have adverse impacts on non-volatile memory, where the energy efficiency degrades [48, 62, 92, 121, 125].

Another complication of orbital deployment is radiation. Even in terrestrial deployment, radiation can induce soft errors in CMOS circuits [9], potentially corrupting the architectural state. Without the earth's atmosphere to shield radiation from space, satellites will be exposed to increased levels of radiation. Circuit level strategies should be used to mitigate this impact [102], however these will come with power, latency, and area overheads.

In this work, we extend the design of MOUSE [93] to be suitable for orbital deployment. We demonstrate that MOUSE can operate over a wide temperature range (despite non-ideal impacts on non-volatile memory) and evaluate its performance at the extremes. MOUSE has an inherent resilience to radiation due to ideal properties of the non-volatile memory it uses [35, 57], but still requires CMOS circuitry to drive operations. We show that even after adding overhead to account for the hardening of CMOS circuitry to radiation [130], MOUSE remains highly energy efficient and performant. We also extend the MOUSE PIM instruction set [93] and add architectural support for branch instructions, which increases the programmability of the device. Finally, we introduce more hardware efficient column activation mechanisms for enabling logic in the memory. The result is a programmable, high performance, and extremely energy efficient beyond edge device which is suitable for deployment in space.

2 NON-VOLATILE PIM

Any non-volatile memory technology can be used as a PIM substrate, including RRAM [118] and PC-RAM [64]. In this work, we focus on Magnetoresistive RAM (MRAM) [85, 116] which has both high density and high endurance. Due to its highly ideal properties, it is even being considered as a universal memory replacement [28] and a few commercial products are already available [1, 2]. Spin-Torque Transfer (STT) MRAM uses the magnetic tunnel junction (MTJ) as its memory element and each memory cell contains one MTJ and one access transistor.

STT-MRAM arrays were modified to enable PIM, creating Computational RAM (CRAM) [14]. It has a unique advantage in that it does not require the use of sense amplifiers or any digital circuitry in the array periphery. The structure of CRAM allows boolean logic operations to be performed within the memory, simply by applying a voltage along the bitlines. The computation remains *entirely* inside the array at all times. In this work, use two variants of CRAM, an optimized version of the standard STT [94] and an extension which increases energy efficiency with a Spin-Hall Effect (SHE) channel [128].

2.1 Magnetic Tunnel Junctions

Magnetic Tunnel Junctions (MTJs) are the resistive memory elements used by STT-MRAM. MTJs consist of two magnetic layers (a fixed layer and a free layer) separated by a thin insulator. The polarity of the free layer can change, but the fixed cannot. If the two magnetic layers are aligned (referred to as the parallel (P) state), the MTJ has a low resistance - which is assigned the logic value 0. When the layers are not aligned (the anti-parallel (AP) state), the MTJ has a high resistance - which is assigned logic value 1. The MTJ will change states if a (relatively) large amount of current is passed through it. The state it transitions to is determined by the direction of the current. If electrons flow from the free (fixed) layer to the fixed (free) layer, the MTJ will switch to the AP (P) state. The current required to change the state is referred to as I_{switch}. A current greater than or equal to I_{switch} will induce switching and a current below I_{switch} will leave the state as is. The voltage required to induce switching can be referred to as V_{switch}, and is determined by $V_{switch} \geq I_{switch} \times R_{MTJ}$, where R_{MTJ} is the resistance of the MTJ. Hence, V_{switch} will also depend

on the state (P or AP) of the MTJ. As will be discussed in Section 5.1, the operating temperature will also significantly impact the voltage required.

To read the value of the MTJ, a voltage below V_{switch} (to avoid switching) is applied across it. The amount of current which passes through it is a function of its resistance (state), allowing it to be detected by a sense amplifier. To write the state of the MTJ, a voltage higher than V_{switch} is applied across it. The voltage will induce a sufficient amount of current through the MTJ to change its state.

2.2 Logic Operations with MTJs

MTJs can be used to perform logic operations if they are connected together in a circuit. An example of a circuit which can perform a 2-input logic operation is shown in Figure 1. Two input MTJs are connected in parallel, which are in series with an output MTJ. Before the logic gate, the two inputs can be in any state, but the output MTJ must be preset to a known value. After the logic gate is performed, the state of the output MTJ should be a function of the two input MTJs.

For example, a NAND gate requires the output to be preset to 0 (low resistance). A voltage is then applied across the terminals V_1 and V_2, such that electrons flow from the input MTJs to the output MTJ. If both inputs are 1 (high resistance), the current will be held sufficiently low to prevent switching of the output MTJ - it will remain at 0. If either of the inputs is 0 (low resistance), there will be enough current to induce switching on the output MTJ. As electrons are flowing from the free layer to the fixed layer of the output MTJ, it will switch to 1.[1] Hence, the output MTJ is the logical NAND of the two inputs; 0 if both are 1, and 1 if either is 0.

Different gates, such as NOT, AND, and N(OR) can be performed by changing the number of inputs, direction of the current, and the preset of the output. Sequences of these gates can be used to perform more complex operations. A full-add can be done with 9 NAND gates. Additions and multiplications can be performed with sequences of full-adds. As the gate set is universal, any computation can implemented.

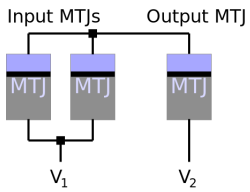


Fig. 1. MTJs connected to implement a 2-input logic gate. The preset value of the output MTJ and the polarity and magnitude of the voltage applied between V_1 and V_2 determines the type of logic gate. The fixed layer is colored in grey and the free layer in light blue.

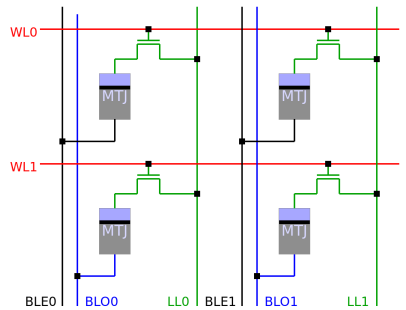


Fig. 2. 4 cells in 2 columns and 2 rows in 1T1M (one access transistor, one MTJ) STT configuration.

2.3 STT Array Architecture

The optimized STT variant of CRAM is nearly identical to a standard STT-MRAM array [94]. Four cells located in two adjacent rows and columns are shown in Figure 2. It has one MTJ and one access transistor in each cell. A standard STT-MRAM array contains two bitlines (typically referred to as bitline and bitline bar). CRAM has three bitlines per column, bit line even (BLE), bit line odd (BLO),

[1]Due to the direction of the current, the output MTJ can only be switched to 1, it cannot be switched back to 0.

and the logic line (LL). BLE connects to even rows and BLO connects to odd rows. LL connects to all rows through the access transistors. The existence of three bitlines is essential for enabling computation, which is explained below. As with standard STT-MRAM, there is a wordline (WL) in each row that controls the access transistor. We now describe how memory and logic operations are performed in the array.

Read: To read from row n, activate WLn. Apply a voltage differential, V_{read}, across LL and the BLE/BLO. Current can be sensed on the bitlines. V_{read} should be lower than V_{switch}.
Write: To write to rown n, activate WLn. Apply a voltage differential, V_{write}, across LL and the BLE/BLO. To write 0 (1), the voltage on BLE/BLO should be higher (lower) than on LL. V_{write} should be larger than V_{switch}.

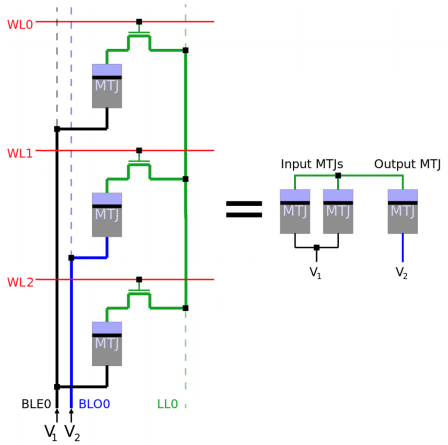


Fig. 3. Demonstration of how a (2-input) NAND gate is performed within the array.

Logic Operation: To perform a logic gate with two inputs in rows n_1, n_2 and with the output in row m, preset row m by performing a write operation. [2] n_1 and n_2 must have the same parity (i.e., both even or both odd) and m the opposite. Activate WLn_1, WLn_2 and WLm. Apply a voltage differential, V_{logic}, across BLE and BLO. Due to the parity requirement, in Figure 1, if V_1 is connected to BLE, V_2 must be connected to BLO, and vice versa. LL connects the free layers (in light blue) of the input and the output MTJs. Current travels from one bit line (either BLO or BLE, depending on the parity of the input cells), through the MTJs in rows n_1 and n_2, through the LL, through the MTJ in row m, and back to the other bitline. Depending on the states of the MTJs in rows n_1 and n_2, the state of the MTJ in row m will either change or not. Figure 3 shows how a NAND gate can be performed inside the array. V_{logic} must be within a specified range for each type of logic operation [94, 127].

Only one operation (read, write, or logic) can be performed in each column at a time. However, operations can proceed in many columns simultaneously. The restriction is that (within a single array) the same operation with the same inputs and outputs must be performed. For example, a NAND gate can be performed in all columns with the inputs in rows n_1 and n_2 and the output in row m.

It may be desirable to perform computation in all columns, or in just a subset of columns. The peripheral circuitry determines which columns participate in every operation. The mechanism for activating columns is covered in Section 3.3 and the instructions which control it in Section 3.4.

[2]Write operation need only be performed if it is not already known to be the correct preset value.

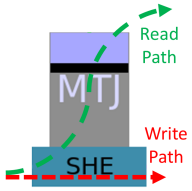


Fig. 4. A magnetic tunnel junction (MTJ) integrated with a Spin-Hall Effect (SHE) channel has seperate read and write paths.

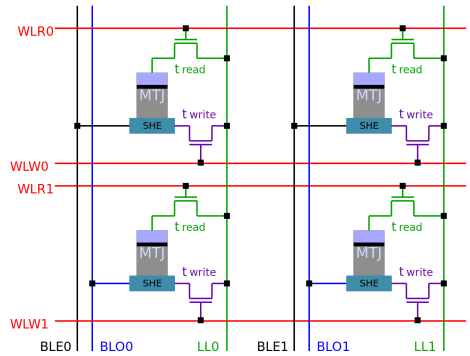


Fig. 5. 4 cells in 2 columns and 2 rows in 2T1M SHE configuration.

Effectively, each column acts as an independent thread which has access to the memory cells within the column. This is highly analogous to the the SIMD lanes of a GPU architecture, where each lane (column) performs the same operation on different data. The active columns act like the bitmask in a GPU, where only active columns perform the operation. However, each column is limited to boolean logic gates (whereas a GPU would have an ALU). Hence, computations in each column are relatively slow, but performance is achieved with a high degree of parallelism.

2.4 SHE Array Architecture

The energy efficiency of MTJ write and logic can be significantly improved by utilizing a Spin Hall Effect (SHE) channel [128]. SHE channels are compatible with both CMOS and MTJs and prototypes have been successfully demonstrated [34]. Proposed memory technology based on this same technology is called Spin-Orbit Torque (SOT) MRAM [34, 85].

The SHE channel provides a more efficient mechanism for switching the state of the MTJ. The combined MTJ and SHE channel is shown in Figure 4. There are two current paths through the device. For the write path, current passes only through the SHE channely. Despite not travelling throught the MTJ body, this current can change the MTJ state. A lower current density is required and the voltage V_{write} can be lowered. This benefit extends to logic operations and V_{logic} as well. The read path passes through the MTJ body, just as with STT-MRAM, and read operations remain the same.

Four augmented cells in two rows and two columns are shown in Figure 5. For the SHE design, there are two word lines per row, word line for read (WLR) and word line for write (WLW). WLR connects the cell to the read path, via t_{read}. WLW connects the cell to the write path, via t_{write}. t_{write} connects the SHE channel directly to LL, allowing current to bypass the MTJ body. t_{write} is activated when the memory cell is being written, or when it is the output of a logic operation. t_{read} connects the end of the MTJ to LL, causing current to travel through the MTJ body. t_{read} is activated when the memory cell is being read, or when it is an input to a logic operation.

The energy efficiency provided by the SHE channel is highly beneficial for beyond edge devices. Reducing the amount of energy per operation can reduce to total execution time (if the device is limited by the available energy), as will be shown in the evaluation. Beyond energy efficiency, the SHE channel also enables more robust logic operations. When performing logic with STT, the resistance of the output MTJ is in series with the input MTJs (Figure 1). With SHE, only the SHE channel is in series with the input MTJs. This makes it easier to distinguish the resistances of the input MTJs, reducing susceptibility to voltage fluctuations [128].

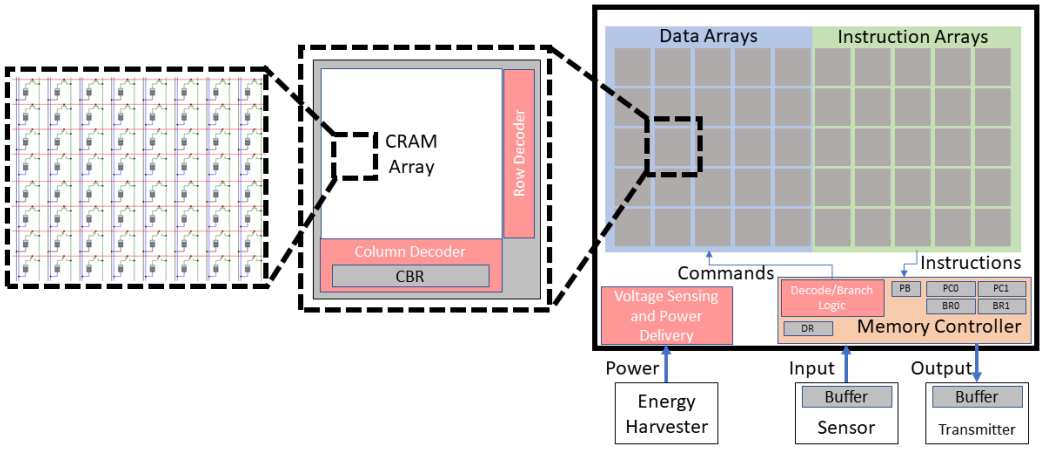


Fig. 6. Overview of MOUSE. Each memory array contains an array of MTJs, a row and column decoder, and a non-volatile register storing the column bitmask. Sense amplifiers are required for reads but are not used in computation. The memory controller contains non-volatile registers to maintain the architectural state.

3 MOUSE DESIGN

In this section, we describe the architecture of MOUSE and show how it is uniquely well suited for beyond edge deployment. MOUSE utilizes CRAM arrays and minimal support circuitry. The computations performed in MOUSE are energy efficient, highly parallel, and have an inherent robustness to intermittent operation. The basic architecture and operation semantics are the same as our previous work [93]. However, we expand the instruction set, improve the method of activating memory to perform computation, and add hardware support for branch instructions.

3.1 Hardware Organization

Figure 6 shows the architecture of MOUSE. It consists predominantly of CRAM arrays.[3] MOUSE can afford to have a large number of arrays (and hence more memory than is typical for a beyond edge device) due to ideal properties of MRAM. The memory arrays are not energy costly as MRAM has a near zero standby power. Hence, MOUSE does not have to worry about static power with large memory capacity (as would be the case for SRAM/DRAM). Additionally, the area overhead of MRAM is very small.[4] For example, NVSIM [29] reports the size of 64MB STT-MRAM array (nearly twice the capacity required by our largest benchmark) as 15.12mm^2. Commercial products of 256MB and 1GB STT-MRAM memory manufactured by Everspin come in a packages that are 130mm^2 [1, 2]. For reference, MSP430FR5994 micro-controller, commonly used as a sub-component of beyond edge devices [18, 37, 42–44, 99], consumes over 100mm^2. Additionally, as computation is performed within the memory arrays, there is no need for an external processor or area costly volatile memory (such as SRAM). Nearly the entire area budget of MOUSE is available for memory arrays. We use a size of 1024x1024 for each CRAM array. In addition to the CRAM arrays, MOUSE requires the following minimal hardware to drive operations and maintain the architectural state

1) A memory controller that reads, decodes, and issues instructions;
2) A non-volatile register for the Program Counter (PC);
3) A 128B memory data register (DR) that facilitates reads and writes;
4) Two non-volatile registers, BR1 and BR2, for branch evaluation;

[3]Each array also contains a row decoder and column decoder.
[4]The SHE configuration consumes more area than STT due to the presence of a 2nd transistor, which we detail in Section 6. However, the area budget still remains modest.

5) Voltage sensing circuitry for monitoring the power source.

Minimal hardware is required for the memory controller.[5] With the exception of resolving branches (covered in Section 3.4.4) and updating architectural variables, its sole responsibility is repeatedly reading instructions, decoding them, and broadcasting them to the CRAM arrays. We use a highly simplified instruction set, covered in Section 3.4, hence decoding requires very little computation. The DR is the same size as one row of the MOUSE arrays and is used for intermediate storage when transferring data to and from different arrays. BR1 and BR2 hold data near the memory controller, enabling quick comparison tests for branch resolution. Finally, the voltage sensing circuitry is standard in beyond edge devices, and is as described in [71].

3.2 Row Activation

In standard memory technology a row decoder activates wordlines for read and write operations. As described in Section 2.3 and depicted in Figure 3, logic operations require the activation of multiple rows (up to 3) simultaneously. To avoid increasing complexity of the row decoder, we use a latching mechanism which holds wordlines high after a row activation [67]. In this manner, the row decoder can activate rows sequentially with normal operation. The hardware cost is two transistors per row. Additionally, each logic operation must wait for the three sequential activations, which increases latency.

3.3 One-Hot Column Decoder

Typically it is desirable to drive logic operations in every column simultaneously. However, it is frequently preferable to perform computation only in a subset of the columns, leaving data in other columns unperturbed. Hence, in addition to a row decoder we also require a column decoder, which will select which columns participate in each operation.

Column activation patterns are different than for rows. With rows, 1-3 rows are activated for every instruction, and the rows which are activated are typically different on each consecutive instruction. For columns, typically many are activated simultaneously (commonly all columns or a large subset). Additionally, columns tend to remain active for long periods of time - (de)activating columns is a rare event.

Previously, MOUSE relied on a decoder which allowed bulk addressing [93, 104]. Activation for the column decoder was the same as for the row decoder, except there were reserved addresses which corresponded to groups of columns and up to 5 column addresses could be specified simultaneously [93]. In this work, we propose a One-Hot bitmask decoding to reduce the significant complexity of the decoder.

Rather than an address, a bitmask is supplied to the column decoder. In a 1:1 bit to column scheme, each bit indicates whether each column should be activated or not. The bitmask is stored a 1024-bit non-volatile register (corresponding to the 1024 columns in each array) within each CRAM array. We call this register the *column bitmask register* (CBR). The CBR can be written with a standard write operation. The advantage of One-Hot bitmask activation is that the column decoder complexity is low: no addresses need to be resolved. The activation signal of each bit can be supplied directly to the columns. The disadvantage is that each activation of the columns (if the column addresses are changing) must be preceded by a write operation to CBR.

It it possible to use different bit to column ratios. Fore example, a 1:32 scheme could be used, where each bit activates a consecutive set of 32 columns. This would allow a 32 bit mask to activate all 1024 columns. This comes with the drawback of being less precise - it cannot activate less than

[5]Our memory controller is a standard memory controller which has been augmented with the capability to read, decode, and issue PIM instructions. We simply refer to it as the memory controller for the remainder of the paper.

32 columns at a time. If computation requires less than 32 columns, additional (and unnecessary) operations will be performed in all 32 columns. This wastes energy. As a 1024 bitmask is easily handled by the CBR we maintain a 1:1 ratio.

3.4 Instructions

Instructions for MOUSE are 64-bit and have the formats shown in Figure 7. There are four categories of instructions, which we explain here.

3.4.1 Memory Instructions. Reads and writes are the standard memory operations, but have additional overhead to account for intermittent operation. The data register (DR) is a non-volatile register the same size as the rows of the CRAM arrays (128B) that holds data between read and write instructions. A read instruction will read from a CRAM array (at the specified address) and write the data into the DR. A write instruction will read data from the DR and write it into a CRAM array (at the specified address). Hence, if there is a power interruption between consecutive reads and writes, the DR will maintain the data being transferred - circumventing the need to re-perform the prior read operation. In addition to memory operations which use the DR, there is also a write immediate instruction, which allows instructions to write data directly into memory.

3.4.2 Logic Instructions. Logic instructions correspond 1:1 with logic gates (as covered in Section 2.2). For example, NOT, (N)AND, and (N)OR are all individual instructions. The instruction specifies the CRAM array address the operation is to be performed in and the row addresses of the logic gate (which rows the inputs and outputs reside in). NOT requires two row addresses (1 input, 1 output) and all others require three row addresses (2 inputs, 1 output). For example, a NAND instruction may specify it is to be performed in CRAM array 15, with inputs in row 7 and 9, and the output in row 12. We restrict logic operations to at most two inputs, which are known to be reliable [94, 127], hence there are only 5 unique logic instructions.

The CRAM array address can specify a single array, or multiple arrays with bulk addressing [104]. There are reserved memory addresses which correspond to groups of memory arrays. For example, it may be desirable to trigger an operation in all CRAM arrays. Array address 111111111 can be a reserved address which will send the instruction to all arrays.

3.4.3 Activate Columns Instruction. It is necessary to specify which columns should participate in each operation. As noted in Section 3.3, consecutive operations typically use the same columns. Hence, which columns to activate changes infrequently. To take advantage of this, we use a strategy where columns are activated and then held active. All following logic operations will be performed in the columns which are held active. To (de)activate columns we use a dedicated instruction, the activate columns (AC) instruction. As described in Section 3.3, the column decoder simply activates the columns depending on the values in CBR. Hence, a column activation consists of two components

(1) A write to the CBR
(2) Triggering of the column decoder to activate columns

Typically an AC instruction will do both components. However, when restarting the device it is only necessary to do the 2nd. Hence there are two variants of the AC instruction, one which does both parts (set and activate) and one which only does the second (activate).

The write to CBR acts like a standard write. As noted in Section 3.4.1, a write can use the value in the DR or an immediate field in the instruction. Hence, there are a total of 3 unique versions of the AC instruction:

(1) Re-activate: Activate using pre-existing value in CBR
(2) Set and Activate: Use data in DR to set CBR and then activate

(3) Set and Activate (Immediate): Use data in immediate field of instruction to set CBR and then activate

3.4.4 Branch Instructions. Branch instructions involve an update to the program counter (PC) in the event a logical condition holds true. As the logic required to evaluate the condition (e.g., checking equality of two numbers) is not complex, it can be implemented efficiently within the memory controller.

Non-volatile registers, BR1 and BR2, reside in the memory controller and are used for condition evaluation. We support simple standard branches based on BR1 and BR2.

(1) beq BR1 BR2: branch if BR1 and BR2 contents are equal
(2) bge BR1 BR2: branch if BR1 is equal or greater than BR2
(3) beqz BR1: if BR1 equal to 0

Hence, the memory controller evaluates a simple condition based on BR1 (and BR2) and updates the PC accordingly. Additional instructions are required to write values to BR1 and BR2. This follows the same semantics as writing the CBR. A dedicated instruction writes to BR1 or BR2, and the value can come either from the DR or an immediate field in the instruction.

Branch instructions increase programmability by enabling function calls, repetitions of computational blocks, and handling I/O events. However, as the computation for branch instructions happens in the memory controller, it cannot capitalize on the extreme energy efficiency and large degrees of parallelism provided by the CRAM arrays. Hence, to remain efficient, branch instructions should remain a low percentage of instruction count.[6]

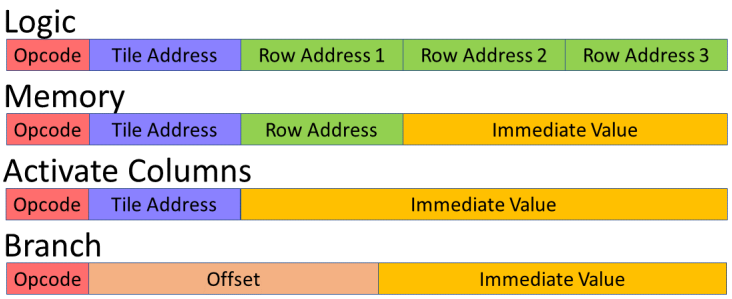


Fig. 7. MOUSE instruction formats. There are three types of instructions, logic, memory, and an additional activate columns instruction for configuration. Opcodes are 5 bits; tile addresses, 9 bits; and row addresses 10 bits each. Branch offset is 20 bits. For instructions which allow for an optional immediate value, it consumes the remaining bits.

3.4.5 Compilation. Compiling high-level code to MOUSE instructions (or any PIM substrate) requires knowledge of the PIM hardware in order to make efficient use of available parallelism. This situation is similar to compiling Open-CL or CUDA code for GPU architectures. Unfortunately, there does not yet exist an equivalent software compiler for PIM.

There is a rich design space, where a multi-dimensional trade-off exists between efficiency, area, power, and performance. Higher degrees of parallelism are possible by spreading computation out over more columns. However, this increases power, consumes more area, and adds communication overhead which reduces energy efficiency. Our strategy was to minimize area by using as few columns as possible, to maximize energy efficiency. Our data layout is similar to a number of other

[6]Avoiding branch instructions is easy for machine learning applications. For our benchmarks, we do not need any branch instructions during a single inference pass. Branches are used only to repeat inference or to handle I/O.

works which have mapped applications to PIM substrates [67, 104], including machine learning algorithms [94].

3.4.6 Issuing Instructions. While operations can occur in multiple arrays simultaneously, arrays do not operate autonomously. All operations are triggered by the memory controller (discussed in more detail in Section 4). Effectively, there is a single controlling "thread", and hence there are no concurrency concerns between individual arrays.

The CRAM arrays in MOUSE hold both data and the instructions. For clarity, we categorize arrays into *instruction arrays* and *data arrays* based on the contents they store. However, all arrays have identical hardware - arrays can be re-categorized to fit the programmer's needs.

Instructions and required data are written into the arrays prior to deployment. During operation, the memory controller repeatedly fetches each instruction from the instruction arrays, decodes it, and then broadcasts it to the data arrays. Instructions are performed entirely sequentially. The next instruction does not start until the previous has finished. This is to guarantee correctness, which will be explained further in Section 3.6 and Section 4.

Different instructions can take different amounts of time to complete. This is because instructions can activate different numbers of rows, each of which has an associated latency. To guarantee that all instructions complete, the memory controller waits longer than the longest instruction before committing and starting the next instruction. This time lapse forms a *cycle*. This conservative approach to issuing instructions comes with a performance cost. For example, a more complex event-driven strategy could issue instructions faster. However, we opt for the conservative approach for three main reasons. The first is that MOUSE is already capable of high performance relative to other devices in the beyond edge domain (as will be shown in Section 6), hence aggressive optimization is unnecessary. Second, complex hardware is less energy efficient and makes it more difficult to guarantee correctness during intermittency. Simplicity is a strength for beyond edge devices. And finally, energy efficiency (rather than high performance hardware) is the limiting factor for performance beyond the edge [38]. Devices which use less energy will be able to complete their programs faster because they will spend less time waiting for sufficient power.

3.5 Power Draw

As power sources for beyond edge devices are highly variable, it is undesirable to connect them directly to the compute circuitry. A solution is to utilize an energy buffer (capacitor) which is charged by the power source. The device can consume energy from this buffer, without having to match the power supplied from the source in real time [73]. Power delivery systems, such as Capybara [19], have been specifically designed to harvest energy and reliably power beyond edge devices in such a manner. Because MOUSE uses such an energy buffer, it accumulates energy over time and consumes it in bursts. This allows MOUSE to consume more power during its power-on time than the power source provides. However, it is possible to program MOUSE to tune the amount of power it consumes. The amount of parallelism that is exploited is determined by the instructions. High degrees of parallelism (many active columns) will enable high performance but also draw more power. This enables a trade-off between latency and power consumption, which can be controlled by the programmer.

3.6 Intermittent Processing

Beyond edge devices are powered by unreliable sources and must frequently shut off. Hence, the devices must be able to maintain correctness of the program during these frequent power cycles. This is referred to as intermittent processing. In addition to providing a correctness guarantee, the hardware support for intermittent processing must also be highly energy efficient. Energy is

Table 1. Four possible cases for re-performing an interrupted AND gate. The output MTJ either should or should not switch for correct operation, and it either did or did not prior to the power being cut.

	Output did not switch before interrupt	**Output did switch before interrupt**
Output should not switch	Input MTJs prevent the switching of the output MTJ, both before and after interruption.	Not possible. Input MTJs prevent switching of output MTJ. By construction, repetition cannot induce switching.
Output should switch	Inputs and output did not change prior to interrupt. Second attempt has same inputs and will produce correct output.	The output has already switched to 0 (correct output). Second attempt has the wrong output preset value. However, due to the *direction* of the current, the output MTJ will remain at 0.

a precious resource for beyond edge devices, and any spent on guaranteeing correctness will be unavailable for normal program execution.

Satisfying both of these constraints is a challenging task. Previously, sophisticated software and hardware strategies have enabled intermittent processing on more traditional architectures [17, 36, 81, 96]. MOUSE is in significant contrast to these strategies. MOUSE is able to checkpoint after every instruction and maintain correctness with extremely limited additional hardware. In fact, the memory controller of MOUSE need only maintain a valid copy of the program counter (PC) and an additional non-volatile status bit in order to guarantee correctness. This strategy is extremely simple and would be crude for more traditional architectures. However, the novel PIM architecture enables this strategy to work effectively and efficiently. More sophisticated and complex strategies are unsuitable and unnecessary. As MOUSE performs the computation in non-volatile memory, progress is automatically saved after *every* operation. Hence, there is no additional backup operations required - a task that has high overhead and complexity in traditional architectures. Hence, MOUSE can checkpoint after every operation with very low overhead. When MOUSE restarts, only two pieces of information are required.

(1) What was the last instruction that was completed (valid value of PC)
(2) Which columns were active

Item 2 is held in the CBR register within all CRAM arrays. Hence, the memory controller need only send a *Re-Activate* instruction to all arrays, causing the column decoders to re-activate all columns. Item 1, the PC, needs to be maintained by the memory controller. In order to keep a copy of the PC up to the last instruction, the memory controller writes (checkpoints) the PC into a non-volatile register after the completion of every instruction. The correctness of the PC is covered more in Section 4.2.

In the worst case, MOUSE looses power after an instruction has been completed, but before the PC can be updated and saved. When power is restored, MOUSE will re-issue the same instruction, performing it for a second time. However, this does not break correctness as the same result is obtained if a single instruction is repeated multiple times, i.e., each such repetition is *idempotent*[46, 113] as will be shown in Section 4.1. The only requirement is that the PC checkpoint happens strictly after each instruction is performed. It must be known that an instruction has been fully completed before the PC is updated.

Checkpointing after every instruction not only minimizes the amount of work potentially lost on shutdown, but it also simplifies the restart process. The simple correctness guarantee, an operation being *idempotent*, does not apply to sequences of operations (over multiple instructions). This is because over the course of multiple instructions, the inputs to logic operations can be overwritten. If we were to re-perform multiple instructions, these input values may be incorrect. To guarantee correctness when repeating multiple instructions, software-level policies and additional presetting operations would be required. While possible, such strategies introduce additional (and unnecessary) complexity.

The second requirement on restart is to restore the previously active columns. As the active columns are stored in the CBR of each memory array, all that is required is for the memory controller to issue a re-active AC instruction. The column decoder in each memory array will re-activate its columns and the memory controller can resume issuing instructions. Correctness during intermittent operation is covered further in Section 4.

3.7 System Integration

When performing the computation for inference, MOUSE is a self-contained system. Memory arrays hold all the instructions and data and the memory controller drives operations. To act as a full beyond edge device, MOUSE will need to be integrated with an energy harvesting power source, a sensor to provide input data, and a transmitter. We assume that input data is stored in a non-volatile buffer within the sensor. The sensor is given a memory address and is considered to be a single memory array, where MOUSE can use read instructions to retrieve data from it. Additionally, the sensor has a non-volatile valid bit, which indicates if new input data is ready. When MOUSE is ready to receive new input data, it can check the valid bit and begin reading from the sensor and writing the data into the MOUSE *data arrays*. These reads and writes are controlled by instructions in the *instruction arrays*, hence data transfer is a software controlled (programmable) process.

When MOUSE finishes inference, the memory controller reads out the data from the arrays, and writes it into a non-volatile buffer for the transmitter. This buffer is also considered to be another memory array, where standard write instructions can be used. In this work, we focus only on the accelerator and do not consider any overhead for the sensor or transmitter.

The programmability of the data transfer process is important. For example, it is possible that MOUSE could loose power during the process of transferring data in to be processed. If power is not available for an extended period of time, when MOUSE restarts there may be a new set of data in the sensor. MOUSE can handle this with branch instructions. If the data in the sensor is timestamped, the first instruction in the transfer process can be used to copy this timestamp into *BR1*. The last instruction of the transfer process can be used to copy the timestamp into *BR2*. MOUSE can check the equivalence of *BR1* and *BR2*, and branch back to the beginning of the transfer to overwrite old data.

4 INTERMITTENT CORRECTNESS GUARANTEE

Beyond edge devices need to protect program correctness in spite of power outages. If not unaccounted for, interruption due to power outage can corrupt the architectural state. In this section we show how MOUSE remains correct, even with unexpected power outages. There are two components that need to be considered, the correctness of individual in-memory operations when interrupted or re-performed (Section 4.1) and the correctness of architectural state variables in transitions between states (Section 4.2). As MOUSE checkpoints after every instruction, we need only show that each instruction and the transitions between instructions remain correct when interrupted. We will show that all instructions and transitions are *idempotent* [46, 113], which means they produce the same results, even if repeated multiple times. The key to remaining idempotent is not over-writing data that will be required on restart (or if it is overwritten, in a manner that does not change the outcome). The architectural state variables and their protection mechanisms are listed in Table 2. Note that the correctness guarantee covered in this section applies only to interruptions and power outages. It does not cover errors in the computation itself or perturbations due to soft errors from radiation.

Table 2. Architectural state variables and how they are protected under power interruptions.

Variable	Volatility	Protection Mechanism
Program Counter	Non-Volatile	Duplicated, valid copy is read only
Parity Bit	Non-Volatile	Only flipped after instruction has finished. Flip is an atomic operation
BR1 and BR2	Non-Volatile	Write operation guarantee (Section 4.1.2)
CBR	Non-Volatile	Write operation guarantee (Section 4.1.2)
DR	Non-Volatile	Read and Write operation guarantee (Section 4.1.2)
Active Columns	Volatile	Bitmask stored in CBR. Re-actived on restart with AC instruction (Section 3.4.3)
Active Rows	Volatile	Activated by every instruction
Data	Non-Volatile	Idempotent logic operations (Section 4.1), Read and Write operation guarantee (Section 4.1.2)

4.1 Operation Level Correctness

In this section we cover the correctness of individual operations performed in the memory when interrupted and re-performed. We are considering the most general case, where the power can be cut at any moment (unexpectedly). Hence, we need to consider what happens when an operation is interrupted in all its possible stages.

4.1.1 Logic Operations. All logic operations are threshold operations (the output MTJ either switches or it doesn't). Hence, there are only two stages for each logic operation, pre- and post-switching. An additional complication to consider is whether the output MTJ *should* switch or not. To be explicit, we use an AND operation as an example. However, the observations here apply to all gates.

To perform an AND gate, the output MTJ must be in the logical 1 (high resistance) state. Voltage is applied across the two inputs and the output (as in Figure 1), such that electrons flow from the fixed layer to the free layer of the output MTJ. This current can potentially change the state of the output MTJ to 0. If either of the two inputs is 0 (low resistance), the current will be sufficiently high to switch it to 0. If both inputs are 1, the current will be too low and it will remain at 1. We must now consider what happens when this operation is interrupted due to power outage, and we need to re-perform it a second time once power is restored.

Consider first the case where the output MTJ *should not* switch. This means the input MTJs are preventing the output MTJ from switching. Hence, the output MTJ could not have switched prior to the interruption. When we re-perform the operation, the initial values of all MTJs are the same. Hence this is identical to performing it the first time, and again the output MTJ will not switch.

Now consider the case where the output MTJ *should* switch. In this case, there are two possibilities: 1) The output MTJ did not switch before interrupt and 2) the output MTJ did switch prior to interrupt. For possibility 1, when re-performing the operation, the initial states of all MTJs are the same. Hence, performing it the second time is identical to performing it the first time (minus the interruption). Hence, this time the operation will be able to finish, and the output MTJ will switch as desired. In possibility 2, the MTJs are not in the same state: the output MTJ has already switched to 0, whereas it should be preset to 1. However, the operation remains correct when performing it a second time. The current applied can only cause the output MTJ to switch to 0, it cannot revert it back to 1. Hence, after performing it a second time, the output MTJ will remain in the 0 state, as desired.

All four cases are listed in Table 1. The catch here is that repeating a logic gate is effectively the same as performing the gate for a longer duration. Doing so results in an identical outcome, regardless of whether the output MTJ switched before interruption (i.e., power outage) or not.

4.1.2 Memory Operations. Re-performing a read operation has no effect on the read data, reading it a second time will produce the same results. Re-performing a write will over-write whatever was written the first time. If the data was unable to be written successfully the first time (due to

interruption) it will be written the second time. If it was written successfully the first time, the same value will be written twice. As noted in Section 3.4.1, read (write) instructions can involve a write to (read from) the *DR*. These are protected by the idempotency of both read and write operations - a memory operation does not write to any address/register that it also reads.

4.1.3 Column Activation. Column activation involves a write to the CBR in a data array and then a triggering of the column decoder. The write to the CBR is kept correct by the same semantics as memory operations (a write can be performed multiple times). The column activation by the column decoder does not change any non-volatile data, and hence cannot introduce corruption. The volatile state is entirely lost on shutdown and will be overwritten on restart.

4.1.4 Summary. Power interruptions can waste energy (due to re-performing instructions) but cannot corrupt the data in memory. Idempotency of all instructions guarantees they produce the same results, even if performed multiple times. Idempotency is not required beyond a single instruction as only one instruction is performed between each checkpoint.

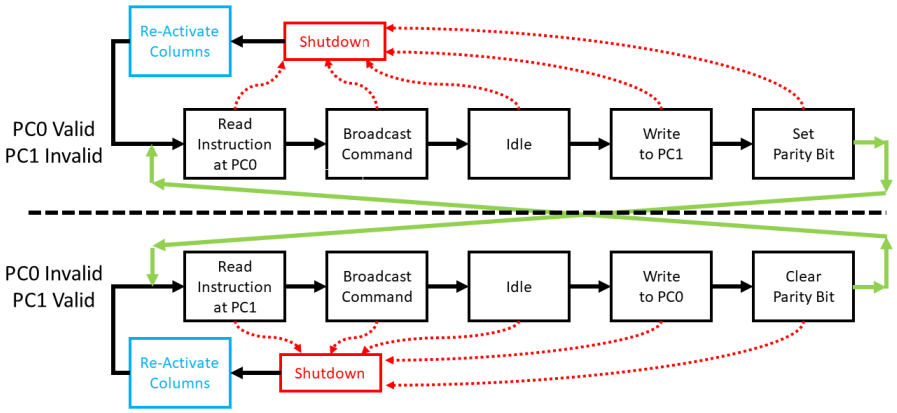


Fig. 8. Memory controller's state transitions to ensure correctness of the program counter as MOUSE transitions from one instruction to the next. Effect of interrupts are dashed and highlighted in red, corrective measures in blue, and forward progress (guaranteed completion of an instruction) in green.

4.2 Maintaining Correct State

The previous section showed that the individual operations performed in the memory are idempotent. We must also be sure that the memory controller can drive the operations and maintain the architectural state in an intermittent safe fashion.

4.2.1 Memory Controller. The memory controller repeatedly reads instructions from the *instruction arrays*, decodes them, and broadcasts them to the *data arrays*. The memory controller waits a sufficient amount of time for the instruction to complete, then updates the program counter (PC). The PC must be stored in a non-volatile register to prevent loss on shutdown. However, a concern remains if the update to the PC gets interrupted. If power is lost during a write to the PC register, it can be corrupted - resulting in incorrect behavior on startup.

We circumvent the issue by maintaining two PC registers, PC0 and PC1, and an additional non-volatile parity bit. If the parity bit is 0, PC0 is valid, and if the parity bit is 1, PC1 is valid. When the memory controller updates the PC, it takes the value in the valid PC register, updates it accordingly, and stores it into the invalid PC register. After which, it flips the value of the parity bit, indicating the advancement to the next instruction. Hence, the memory controller never writes to the valid PC register, and there is no risk of corruption.

The setting of the parity bit is analogous to the committing of an instruction in traditional architectures. As it is a single bit, the operation is *atomic*, it cannot be interrupted mid-way through. It either is set or it is not. If an interruption occurs before the setting of the parity bit, the memory controller will re-issue the same instruction on restart, which is safe to do (covered in Section 4.2.2). If the interruption occurs after the setting of the parity bit, the instruction has completed and the memory controller will issue the next instruction on restart. This process is depicted in Figure 8.

There are other non-volatile registers that hold the architectural state. This includes the data register (DR), branch registers (BR1 and BR2), and the column bitmask registers (CBR) in each array. These registers are protected by the same semantics as in Section 4.1.2. If values are overwritten, they are guaranteed to be completed before the memory controller commits the instruction. No register is both read and written by the same instruction, so no required data can be corrupted.

The currently active columns is part of the architectural state. When MOUSE restarts, the columns will need to be re-activated. The non-volatile CBR in each memory array will maintain the currently valid bitmask. Hence, all that is required is for the columns decoders to re-activate columns. The memory controller can accomplish this by issuing a re-activate columns instruction to all arrays on restart (Section 3.4.3).

4.2.2 Data in Arrays. The previous section showed that the memory controller itself remains correct during intermittent operation. We must also ensure that the memory controller does not generate any signals which corrupt the data residing in the memory arrays.

The memory controller broadcasts instructions to the data arrays. This broadcast is not atomic, and thus can be interrupted at any stage. However, all the operations that it can trigger are idempotent (Section 4.1), meaning they can safely be interrupted at any point in their progression. As a direct result, the broadcast cannot cause corruption as it's only effect is the initiation of the operation. Power can be cut before the broadcast reaches a memory array, while the operation is being performed, or after the operation has finished –none of these cases can introduce error.

5 IMPACT OF ORBITAL DEPLOYMENT

An exciting domain for beyond edge devices is low earth orbit (LEO) where they can act as nano-satellites [74]. One impact of LEO deployment is that the cost of communication is much greater than computation (even more so than for terrestrial deployment) [37, 74]. For this, MOUSE is well suited as it has a large memory capacity relative to other beyond edge devices (due to consisting nearly entirely of high density non-volatile memory). The high memory capacity would enable MOUSE to go long periods of time and store many results of interest before attempting communication. However, orbital deployment will also introduce challenges related to operating temperature and subjection to radiation. We discuss here how MOUSE can tolerate such conditions.

5.1 Temperature

Satellites in LEO can experience a wide range of temperatures, from -170C to 123C [65]. Maintaining proper temperature on large scale satellites is an important engineering challenge [8]. However, nano-satellites might not have sufficient resources to do any temperature modulation. Hence, it is desirable that they be able to operate properly across a wide range of temperatures. A potential challenge for MOUSE is that it consists mostly of non-volatile memory, which has characteristics which are significantly impacted by temperature [92].

The resistance of MTJ devices is dependent on temperature. It increases with decreasing temperature. The resistance at -170C can be as much as 30% higher than at room temperature [62, 122]. This increases the voltage required to write MTJs, and consequently, increases energy consumption. The SHE architecture will be less affected than STT as the SHE channel is metallic and will not increase

in resistance. Hence, write operations with SHE will remain largely unaffected. However, SHE still requires current to travel through the body of input MTJs for read and logic operations. Hence, the overall energy efficiency of SHE will still decrease. Because of this, PIM which uses MTJs (or other non-volatile technologies), is less energy efficient at cold temperatures [92]. However, the change is modest, overall remaining within approximately 10% additional energy consumption (relative to room temperature), even at cryogenic (77K) temperatures [92]. Given that MTJs are extremely energy efficient [126, 128], this increase in energy consumption remains tolerable. Additionally, there is a benefit of cold temperature. The ratio between the high and low resistance state increases [62, 125]. This leads to more robust logic gates which are less susceptible to voltage fluctuations [94, 127]. The overhead we consider for cold operation is described further in Section 6.

The inverse is true at high temperatures. The overall MTJ resistance and the ratio between the high and low resistance states are both lower. MTJ resistance at 123C is roughly 86% of its resistance at room temperature. Hence, the MTJ logic gates themselves will be more energy efficient. However, they will also be more susceptible to voltage fluctuations. This increases the critical nature of the power delivery system. Power systems, such as Capybara [19], will be necessary to ensure the proper voltage is applied across a variety of temperatures. Using switched-capacitor voltage converters [53, 89, 89], the necessary voltages can be delivered to facilitate operation. The overhead of voltage conversion is discussed more in Section 6.

In contrast to the resistive memory devices, the peripheral circuitry in MOUSE should benefit from cold temperatures. At colder temperatures CMOS transistors have higher ON current [122], switch faster [86], have a higher trans-conductance, and have a steeper subthreshold slope [105] leading to lower leakage current. However, MOUSE will not benefit significantly from these characteristics. The non-volatile memory already has extremely low static power and the latency will be limited by the switching time of the MTJ devices. CMOS performance can degrade with increasing temperature due to increasing leakage current. However, typical CMOS devices can operate well up to 175C [54]. Radiation hardened bulk CMOS technology can increase this further to 250C [54, 72]. Hence, CMOS technology is well suited to operate within the expected temperature range of LEO satellite. We discuss overhead of our CMOS components further in Section 6.

5.2 Radiation

Exposure to radiation can cause bit-flips (soft errors) on CMOS circuitry. When deployed in orbit, beyond edge devices will be exposed to significantly higher levels of radiation. The CMOS components of MOUSE, such as the memory controller, will be equally susceptible to soft errors as other beyond edge devices. A variety of circuit level strategies to mitigate soft errors [102] exists (such as increasing node capacitance and transistor drive current [130]), which will introduce additional area, latency, or power overhead. However, the vast majority of MOUSE's computation and all of its memory exist in non-volatile MTJ devices. Fortunately, MTJs are considerably more robust to soft errors than other forms of memory [57, 58]. In fact, MTJs have been shown to be highly resilient to radiation from heavy ions [20, 58], neutrons [91], protons [49], and gamma rays [49, 91]. For these reasons, MTJs have been considered a leading candidate for usage in space applications [35, 57]. Since MOUSE consists mostly of MTJs, it will be less susceptible to radiation than traditional architectures. As only minimal circuitry is required external to the memory arrays, using circuit level strategies to increase CMOS resilience to radiation [130] will have a lower total overhead. To account for the overhead, we increase the latency and energy of MOUSE's peripheral circuitry, discussed further in Section 6.

Table 3. Parameters for MTJ devices.

Parameter	Modern	Projected
P State Resistance	3.15 kΩ	7.34 kΩ
AP State Resistance	7.34 kΩ	76.39 kΩ
Switching Time	3 ns [84, 97]	1 ns [50, 126]
Switching Current	40 μA [97]	3 μA [126]

6 EVALUATION SETUP

Benchmarks: The exact use case of beyond edge devices can vary significantly, applications include agricultural monitoring [74, 114], security, and structural and environmental monitoring [26]. However, general sensor processing algorithms can be used to solve a wide variety of problems. We use benchmarks which are representative for many possible use cases - machine learning inference on data sets which are tenable for beyond edge devices. The specific input problem will vary depending on the user, however the computation involved should remain highly similar.

We implement two machine learning algorithms, Support Vector Machines (SVM) and Binary Neural Networks (BNN). Both are widely used and light weight, which makes them highly suitable for the beyond edge domain. We built customized SVM implementations and trained and tested them in R [88]. We use only operations that are efficient in MOUSE, all bit-wise and integer arithmetic. We were able to achieve a similar accuracy as standard SVM implementations from libSVM [12]. For inference, the main computation is effectively performing the dot product between an input vector and each of the support vectors. The results of these dot products are then squared, multiplied by a set of coefficients, and finally summed together. By construction, SVMs have two class outputs, where the sign of the output value is the classification. We extend to multi-class classification by training a separate SVM for each possible output class, where each has the task of identifying a single class. BNNs are neural networks use neurons and weights represented by a single bit each [22]. This enables multiplications to be replaced by XNOR operations and addition is simplified to a popcount operation. This gives BNNs extreme energy efficiency. Previous work has efficiently mapped BNNs onto FPGAs, including FINN [111] and FP-BNN [68]. We copy their network configurations exactly. We modify the algorithms only in transforming them to run on our PIM substrate. Hence, our accuracy is identical.

Data Sets: For small scale image recognition we use MNIST [63]. The task is digit recognition, where a 28×28 pixel image with 8-bit precision is to be classified into one of ten digits (0-9). We use both BNNs and SVMs on this benchmark. With SVM, the pixels are a 784 element vector. We also create a binarized version, where pixels that have a value below $255/4 \approx 63$ are assigned 0 and those above are assigned 1. This allows us to replace multiplications with AND gates, significantly reducing the time, energy, and area overhead. For BNNs, we use the network configurations of FPGA-based FINN [111] and FP-BNN [68]. FINN [111] uses binarized input. It has three hidden layers of 1024 neurons (bits) each and the output layer has 10 neurons with 10-bit precision. FP-BNN [68] 8-bit inputs and has three hidden layers of 2048 neurons each. The output layer has 10 neurons with 16-bit precision.

Human Activity Recognition (HAR) [3] is a data set which has accelerometer and gyroscope measurements from a smartphone, which is carried by participants performing a variety of activities. The problem is to classify the physical activity the individual is performing. Each input is a vector of 561 elements. We convert the input to fixed point representation with 8-bit precision.

ADULT [59] contains census information. The problem is to classify whether an individual makes greater than $50K per year or not. We use a reformatted version of the data set from libSVM [12]. Each input is a 15 element vector where each element is an 8-bit integer.

Performance and Energy Model: We use an in-house simulator to evaluate MOUSE. We set each array in MOUSE to 1024x1024, which is a recommended subarray size for non-volatile memories from NVSIM [29].

We simulate with both modern MTJ parameters [98] and with projections of MTJ parameters expected to be possible within a few years [126, 128]. MTJs are expected to significantly more energy efficient as the technology matures. Two techniques will enable a reduction in the switching current, 1) decreasing the damping constant of ferromagnetic materials [30, 83, 101] and 2) using a dual-reference layer structure [27, 47]. It is possible switching currents will be as low as 1 μA, however we assume 3 μA to be conservative. The parameters we use are shown in Table 3. For Modern MTJs we use only the STT architecture, for projected MTJs we use both the STT and SHE architectures. The benefit of SHE is providing a more efficient write mechanism. We model the SHE channel as a 1 kΩ resistance. This provides a conservative estimate of SHE energy efficiency.

Due to the different switching times of modern and projected MTJs, we clock MOUSE at different speeds for each. With Modern MTJs MOUSE operates at 30.3 MHz clock rate (33ns per cycle) and for projected MTJs MOUSE operates at 90.9 MHz clock rate (11ns per cycle). This enables sufficient time for instruction read, decode, and the peripheral circuitry latency and MTJ switching time.

For modelling peripheral circuitry, we take data from NVSIM [29] which reports the relative overhead of peripheral circuitry for modern MRAM memory. We set the latency and energy overhead of MOUSE so that is consumes the same relative share of total latency and energy as reported by NVSIM. We also account for the energy required to read instructions, update program counter and valid bit, and the specification of row addresses.

We first evaluate MOUSE with continuous power (using a power source which can supply as much power as MOUSE desires). Then, we evaluate with an energy harvesting power source where MOUSE will have to operate intermittently. We model the energy harvester as a (small) constant power source which is filling an energy buffer (capacitor). When MOUSE is off, the power source charges the capacitor and the voltage will rise. When MOUSE is on, it will consume the energy and the voltage will drop. MOUSE will shut off when the voltage hits a pre-defined minimum value, hence the voltage on the capacitor will fluctuate within a specified range. When the voltage hits the lower end of the range, power is instantaneously cut - MOUSE does not do any preparation for the shutdown. We start all benchmarks with a capacitor that has voltage just below the cutoff, hence all benchmarks begin with an initial charging time. Modern MTJs and Projected MTJs have different operating voltages [126], so we use a different voltage range for each technology. We let the voltage fluctuate between between 400 mV and 420 mV when using Modern MTJs and between 100 mV and 120 mV when using Projected MTJs. Switched-capacitor converters are used for upconversion and downconversion [41] to supply the required voltages for all operations. All required voltages can be acquired by using conversion ratios of 0.75, 1, 1.5, and 1.75 [53, 89]. We evaluate MOUSE on the power supplied by the converter, the evaluation does not include regulator efficiency overhead. The converter may have an efficiency anywhere between 35-80%, hence the energy harvester may need to provide roughly 1.25-2.85× the energy that MOUSE consumes. As noted in Section 3.4.6, a single instruction is performed in every cycle. A portion of the cycle must be dedicated to changing the output voltage of the converter (if consecutive operations require different voltage levels). The time overhead can be overlapped with the row activations.

It is desirable to match the capacitor size to the expected energy consumption. Hence, we also use different capacitor sizes for modern and project MTJs. We use a 100 μF capacitor (energy buffer) with Modern MTJs and a 10 μF capacitor for Projected MTJs. The optimal capacitor size depends on the technology and the program being executed. When deployed, a system such as Capybara [19] could be used to tune the parameters of the energy buffer.

Given that energy harvesting power sources can vary significantly in how much power they can provide, we sweep the power source over a wide range. At the low end, we test from 60 µW which is approximately what can be harvested from a 1cm^2 thermal energy harvester running on body heat [56, 66]. This is well below the operating power of MOUSE. At the high end we use 5 mW, which is the same power harvested by the beyond edge device SONIC [37]. This can nearly power MOUSE continuously. Beyond edge devices deployed as satellites will likely use solar cells as power sources [74]. The amount of power which can be harvested will depend on the size of the cells (typically very small) and their orientation which is likely to change over time.

Area Overhead: The CRAM arrays used in MOUSE have a similar area overhead as MRAM arrays. The extra overhead of STT CRAM is an extra bit line per column, which is a minor impact. For SHE CRAM, a second transistor and SHE channel is required in each cell, which has a significant impact.

We base our cell area estimates from Zabihi et. al. [127]. We use configurations where the access transistors have a resistance less than 1 kΩ and give an extra 10% to account for spacing and layout issues. The access transistors and MTJs can be placed on separate layers. As the transistors are much larger, they dominate the area overhead. As the SHE architecture has twice as many transistors, it is approximately twice as large. We use NVSIM [29] to estimate the area overhead of peripheral circuitry. NVSIM reports the percentage of chip area which much be dedicated to the peripheral circuitry for different memory sizes. We increase the area overhead for each of benchmarks accordingly. Our conservative area estimates are shown in Table 4.

Impact of Temperature: MTJs have been demonstrated to function over a wide range of temperatures [62, 125]. However, the MTJ resistance increases at colder temperatures, which will increase energy consumption. We test MOUSE both at -170C (cold) and 123C (hot). To model the impact on MTJs we take data from Yuan et. al. [125]. For cold temperatures, we conservatively estimate the MTJ resistance increases by 30%. For the STT architecture, this increases the write, read, and logic energy consumption by 30%. For SHE, the write energy remains unaffected as the SHE channel (which is metallic) is use for write operations. However, energy consumption still increases for read and logic operations. The CMOS circuitry will generally perform better at cold temperatures, having a lower latency and potentially lower energy [86, 105, 122]. However, to be conservative, we assume no additional efficiency of the peripheral circuitry. The latency improvement of CMOS does not benefit MOUSE as we choose maintain the same clock rate across temperature ranges. Hence, the latency of each instruction remains the same. At hot temperatures, the MTJ resistance drops by approximately 13% [125]. We model this in identical fashion to cold temperatures, where we change the energy efficiency of each operation.

Impact of Radiation: As noted in Section 5.2, MTJs have an inherent resilience to radiation. However, the CMOS components of MOUSE remain vulnerable. Circuit level strategies, such as increasing node capacitance and transistor drive currents [102, 130], can make CMOS circuits resistance. These strategies come with a power and delay cost. We choose to be conservative, and assume a large overhead of a 60% increase in CMOS energy and a 10% increase in CMOS latency [130].

Baseline for Comparison: We compare MOUSE with SONIC [37], a beyond edge device which performs machine learning inference on the same benchmarks we use. As SONIC was evaluated at room temperature, we must estimate its performance at different temperature ranges. To be conservative, we assume SONIC can fully exploit the benefit of CMOS operation at cold temperatures, increasing performance by 30% [86]. We also assume it suffers no negative consequences of varying temperature (hot or cold) and it pays no overhead resilience to radiation. We also compare against estimates of the vector architecture MANIC [38]. We also give MANIC overly optimistic assumptions, a 30% boost in performance and no overhead for temperature or radiation. MANIC was not evaluated on end-to-end inference, rather on computational kernals required for inference

Table 4. Area required for MOUSE for different benchmarks and configurations. Units are in mm^2.

Benchmark	Total Memory	Modern STT[127]	Projected STT [127]	SHE
SVM MNIST	64MB	28.04	21.27	42.54
Binarized	8MB	2.99	2.27	4.53
SVM HAR	16MB	5.97	4.53	9.06
SVM ADULT	1MB	0.39	0.29	0.58
BNN FINN MNIST	8MB	2.99	2.27	4.53
BNN FPBNN MNIST	16MB	5.97	4.53	9.06

Table 5. Continuously powered MOUSE at room temperature (using STT design and modern MTJ devices) and related work under continuous power. The CPU does not benefit from MNIST binarization as it still performs 64-bit integer multiplication.

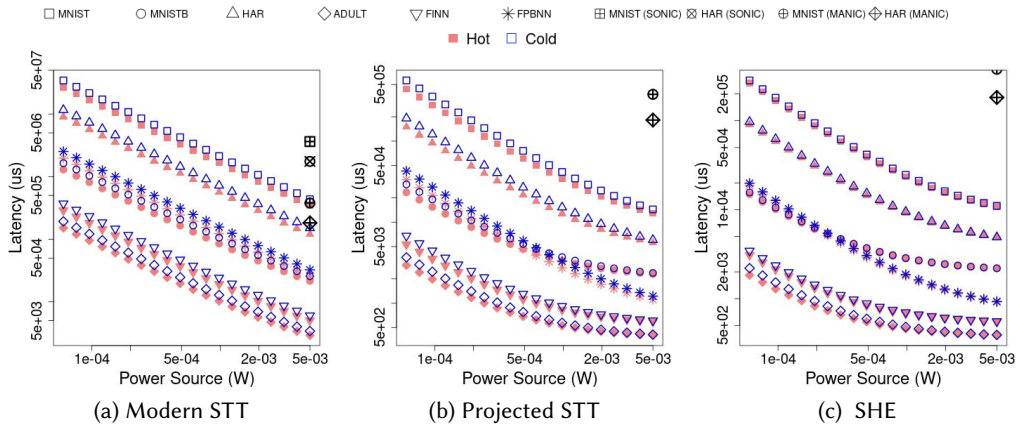
Benchmark	Latency (µs)	Energy (µJ)	#SV	I/D Mem (MB)	Area (mm^2)	Accuracy
SVM (CPU)						
MNIST	169,824	5,094,702	11,813	-	-	97.55
MNIST (Binarized)	192,370	5,771,085	12,214	-	-	97.37
HAR (integer) [3, 110]	127,494	3,824,822	2,809	-		95.96
ADULT	4,368	131,052	1,909	-	-	76.12
MOUSE SVM (Modern STT)						
MNIST	23,116	1,700	11,813	4.5 / 30.0	28.04	97.55
MNIST (Binarized)	6,071	81.43	12,214	1.25 / 6.0	2.99	97.37
HAR (integer) [3, 110]	11,312	575.8	2,809	2.25 / 10.0	5.97	94.57
ADULT	1,104	9.06	1,909	0.25 / 0.5	0.39	76.12
MOUSE BNN (Modern STT)						
MNIST (Binarized) FINN	1,605	18.04	NA	3.15/1.71	2.99	98.4
MNIST FP-BNN	2,150	125.4	NA	4.20 / 8.00	5.97	98.24
libSVM [12]						
MNIST	7,830	234,900	8,652	-	-	98.05
MNIST (Binarized)	19,037	571,116	23,672	-	-	92.49
HAR (integer)	1,701	51,042	2,632	-	-	93.69
ADULT	379	11,370	15,792	-	-	78.62
SONIC [37]						
MNIST	2,740,000	27,000	NA	0.256	> 100	99
HAR	1,100,000	12,500	NA	0.256	> 100	88

(i.e. convolution). Hence, we rely on rough estimates of its performance on the same benchmarks. We follow the authors' statement, that MANIC 9.6× more energy efficient than SONIC [38].

7 EVALUATION

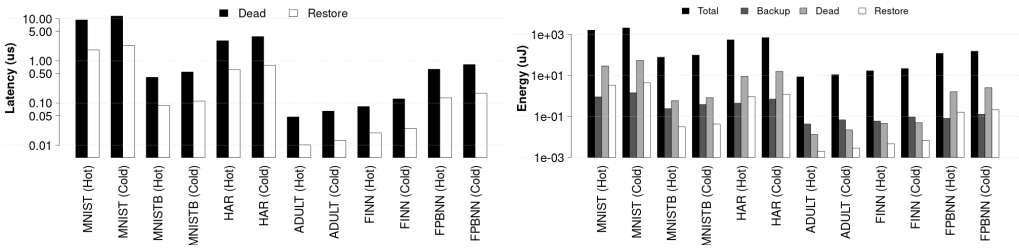
Continuous Power: Continuously powered MOUSE at room temperature and related work is reported in Table 5. MOUSE implements both BNNs and SVMs. SONIC [37] is beyond edge device which uses TI-MSP430FR5994 microcontroller to run neural networks on the same benchmarks. For reference, our custom SVM implementation and optimized SVMs from libSVM [12] are run on a Intel Haswell 5-2680v3 processor. To be conservative, we account only for the processor power consumption and assume it operates at its idle power. Overall, MOUSE has a significant energy efficiency advantage and a competitive latency. Notably, MOUSE consumes more memory than SONIC. However, this is reasonable as MOUSE consists nearly entirely of non-volatile memory, which has high density. MOUSE does not require external processing logic or area costly volatile memory.

Intermittent Operation: We now evaluate MOUSE with intermittent computation, where a small power source is charging a capacitor that MOUSE can draw energy from. The latency (including time powered off) of all benchmarks with each MTJ device (and different operating temperatures) over the range of power sources (60 µW - 5 mW) is plotted in Figure 9, along with a comparison to SONIC



(a) Modern STT (b) Projected STT (c) SHE

Fig. 9. Latency (µs) vs. Power Source (W) for each MOUSE configuration and SONIC [37]. MOUSE at hot temperature is shown in Red/Filled shapes and at cold temperature is shown in Blue/empty shapes.



(a) Latency (b) Energy

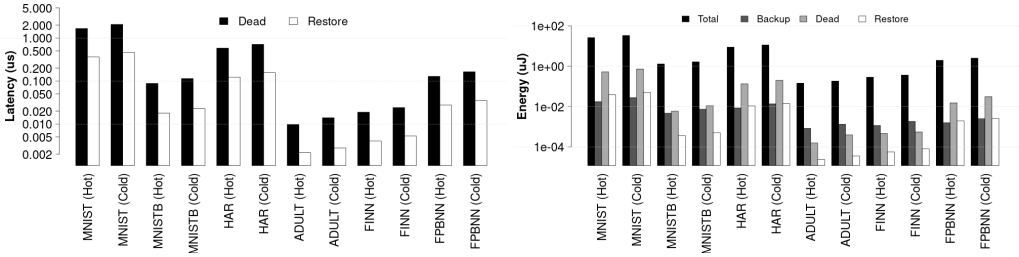
Fig. 10. Latency/Energy Breakdown: Modern STT.

[37] and MANIC [38]. All MOUSE configurations are able to significantly outperform SONIC for the same power budget. Despite conservative estimates of MTJ performance, conservative estimates of peripheral circuitry, and very optimistic estimation of MANIC across the temperature range (30% boost in performance and no overhead for temperature or radiation) MOUSE has a similar performance with MANIC. On the MNIST data set, if MOUSE uses 8-bit inputs, its latency is 0.91× (1.15×) that of MANIC at hot (cold) temperatures. On the HAR data set, MOUSE has a latency that is 0.66× (0.83×) that of MANIC at hot (cold) temperatures. Hence, MOUSE has better performance on average, with better results at warmer temperatures.

At cold temperatures MOUSE has a higher latency than when hot. At 60 µW, overall cold is 23.4% slower on average. While MOUSE has the same clock rate and issues instructions at the same rate, the instructions consume more energy when cold. Hence, MOUSE will run out of energy and have to power off more frequently. Temperature has a varying level of impact on each MTJ technology. Modern STT has a 33.3% higher latency and Projected STT has 28.5% higher latency at cold temperature, across all benchmarks. SHE is less effected by temperature because write and logic operations use the SHE channel, which is not only more energy efficient but less affected by temperature. SHE has an 8.6% higher latency across all benchmarks at cold temperature.

Independent of temperature, SHE is the most energy efficient. Because of this it drains the capacitor less often, and hence has fewer power outages leading to the overall lowest latency.

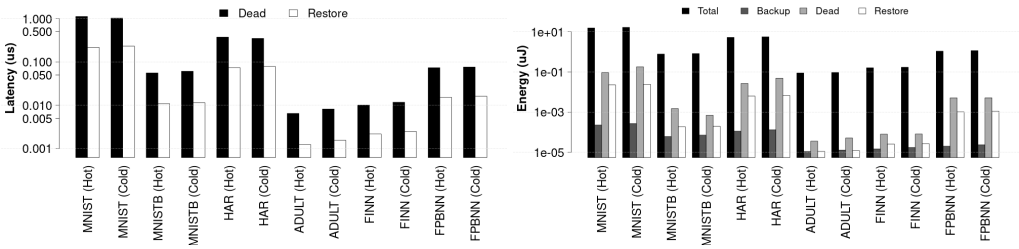
Projected STT has a lower latency than Modern STT, as it can operate at higher frequency (11ns per instruction vs. 33ns) and it is more energy efficient.



(a) Latency (b) Energy

Fig. 11. Latency/Energy Breakdown: Projected STT.

MOUSE spends negligible amounts of energy while powered off. Hence, the energy consumption is nearly independent of the power supply. The vast majority of the energy is dedicated to normal program execution. A small portion is dedicated to overhead for intermittent execution, which will vary depending on the number of interruptions (which is determined by energy efficiency and the capacitor size). The total energy is plotted in Figure 10(b) for Modern STT; in Figure 11(b) for Projected STT; and in Figure 12(b) for SHE; assuming a 60 µW power source.



(a) Latency (b) Energy

Fig. 12. Latency/Energy Breakdown: SHE.

There are metrics specific to beyond edge devices which indicate how efficient the checkpointing strategy is [100]. In addition to the total energy, we report the Backup energy, Dead energy, and the Restore energy. Backup refers to any actions required prior to shutdown to save the state. For more traditional architectures, this involves writing data back to non-volatile memory. For MOUSE, the only backup operations are saving the PC, flipping the parity bit, and writing values into the CBR (to indicate which columns are active). MOUSE does the first two on every instruction, and the second only AC instructions. Dead refers to any computation that must be re-performed after restart (which was lost due to the shutdown). For MOUSE, this is at most re-performing the very last instruction. Restart is any actions required to put the device back into operating condition after a shutdown. For MOUSE, this is the re-activation of columns with an AC instruction.

Backup has no associated latency, as MOUSE's backup operations are overlapped with normal program execution. However, we do report Dead latency, which is the time it takes to re-perform the last instruction, and the Restore latency, which is the time it takes to re-activate columns. To

remain efficient, a beyond edge device should have low Backup, Restore, and Dead latency and energy.

Overhead for Backup, Restore, and Dead are reported in Figure 10 for Modern STT; in Figure 11, for Projected STT; and in Figure 12, for SHE. Note that the y-axis is log scale. The total energy encapsulates all energy used for computation, as well as Backup, Restore, and Dead energy. Also note the total latency is provided for all architectures in Figure 9 – where the breakdown figures capture the data for the 60 µW power source.

The overheads for Backup, Dead, and Restore increase with cold temperature. This is for two reasons. The first is that the actions required for each will cost more energy due to the MTJ characteristics. For exampling, writing the PC value or re-performing the last instruction will involve MTJ operations, which will take more energy at cold temperatures. The second reason is that the overall lower energy efficiency at cold temperatures leads to more power outages. At cold temperature, across all benchmarks and technologies, MOUSE restarts 24.4% more often than at hot temperatures. This increases the number of instructions that need to be re-performed and the number of times architectural state variables will be saved.

Modern STT is the least energy efficient, which means it must restart the most. Because of this it has the largest relative Dead energy. At the extremely low power of 60 µW, on average, across all benchmarks, Dead energy is 0.98% (1.09%) of the total energy at hot (cold) temperature. The projected MTJs have lower overhead, where Dead energy (on average) becomes 0.796% (0.804%) of the energy for Projected STT and 0.194% (0.323) of the total for SHE at hot (cold) temperatures. Dead latency, on the other hand, is 0.068% (0.084%) of the total for Modern STT, 0.040% (0.050%) of the total for Projected STT, and 0.020% (0.020%) of the total for SHE with hot (cold) temperatures. Restore is only 0.013% (0.016%) of the latency and 0.066% (0.069%) of the energy for Modern STT; 0.008% (0.010%) of the latency and 0.048% (0.049%) of the energy for Projected STT; and 0.0037% (0.0040%)of the latency and 0.0436% (0.0436%) of the energy for SHE with hot (cold) temperatures. As Restore latency and energy is due to peripheral circuitry, SHE has no advantage over STT for an individual restart. However, SHE still requires fewer restart operations due to its overall increased energy efficiency. Backup energy is, on average across all benchmarks, 0.304% (0.337%) for Modern STT; 0.350% (0.340%) for Projected STT; and 0.009% (0.009%) for SHE. Backup has no associated latency as it is performed at the same time as each instruction on every cycle. Overall the Backup, Dead, and Restore overheads increase only modestly at cold temperatures. Hence, the checkpointing mechanisms remain efficient across the wide temperature range and MOUSE is suitable for use as an intermittent accelerator in the harsh environments of LEO.

Restore and Dead latency and energy are all zero for the case of a continuously powered system. This is because there are no power outages and, hence, never a need to restart the system or re-perform any potentially unfinished instructions.

8 RELATED WORK

Orbital Edge Computing (OEC) was proposed by Denby and Lucia [25, 26] as a new model for satellite computation. The authors describe architectures for computational nano-satellites. Additionally, they proposed a strategy called the computational nano-satellite pipeline, which parallelizes computation across collections of satellites in order to reduce latency. MOUSE could be used as a sub-component within such computational satellites.

MANIC [38] is a vector architecture for low power systems. The authors suggest that energy efficiency is the most important metric for beyond edge devices, more so than efficient checkpointing mechanisms.

Traditional architectures have been significantly modified to be intermittent safe. A strategy has been to tightly integrate non-volatile memory with volatile registers to enable a fast and more

efficient backup process just prior to shutdown. These architectures are known as non-volatile processors (NVP) [71, 79]. A system utilizing a THU1010N non-volatile processor was analyzed, where trade-offs in checkpointing strategies are evaluated [71]. Follow up work has increased the resilience of NVPs to power interruptions [77, 78]. The NVP in [77] can complete the FFT benchmark from MiBench [40] in 4.2 ms. Cilasun et. al. [23] evaluated FFT implementations on CRAM, the same PIM substrate which MOUSE uses. Performing a similarly sized problem, the best latency they were able to achieve is 1.63 ms. However, adapting this implementation to be intermittent safe in the same manner in MOUSE would add a latency overhead. PIM has been incorporated into beyond edge devices previously, using RRAM arrays for acceleration [106]. However, this design still requires a CPU to perform logic and orchestrate control. PIM is only a sub-component of the system, hence the efficient checkpointing strategy of MOUSE cannot be applied to this architecture.

ResiRCA [87] uses an adaptable RRAM crossbar accelerator for MAC (multiply+accumulate) operations for CNNs. The architecture is able to adapt to varying levels of input power to efficiently utilize the PIM components. However, a battery is required in order to maintain an external controller. Additionally, a significant amount of computation occurs outside the memory array (only MACs are processed by the memory). Hence, the MOUSE's checkpointing mechanism is also not applicable to this architecture. Many RRAM accelerators have been developed [107, 108, 119, 123]. However, these architectures only use the RRAM array as an accelerator for specific operations. The full system contains much additional circuitry and logic in addition to the memory arrays. This significantly increases the difficulty to adapt to intermittent processing. Additionally, they require ADCs (analog to digital converters) for every PIM operation, which has a significant area and energy overhead.

Capybara uses a re-configurable hardware energy storage mechanism and a software interface that allows the specification of energy needs for different tasks. This gives the system more flexibility in satisfying the requirements of different kinds of tasks. In this work we assumed a constant capacitor size, however Capybara could enable variable size energy buffers to more closely match the requirements of each application.

Hibernus [7] is a system that reactively hibernates and wakes up. This is a similar shutdown policy to MOUSE. However, Hibernus performs an additional back-up operation before shutting down, whereas MOUSE does not need to.

Many strategies have proposed to enable more traditional systems to operate intermittently. CleanCut [17] works with LLVM to compile programs with checkpoints, and uses a statistical energy model to find potential non-terminating paths. Chinchilla [81] uses adaptive checkpointing, where the frequency of checkpoints is a function of the number of interrupts. Coati [96] developed methods to ensure correctness of concurrent threads in the presence of interrupts for intermittent systems. The What's Next Intermittent Architecture [33] uses approximation to improve performance. Rather than following an all-or-nothing approach, What's Next computes approximate results and continually improves the output. If an acceptable output is achieved it will skip to processing the next input. This enables the device to process more inputs as it does not waste time and energy achieving unnecessary accuracy.

The EH model [100] is a design space exploration tool for energy harvesting architectures. As noted by the authors, energy harvesting systems can generally be divided into two types, 1) multi-backup, which perform many backups between power outages, and 2) single back-up, which only save state once before a power outage. Multi-backup systems include Mementos, [90], DINO [75], Chain [15], Alpaca [80], Mayfly [45], Ratchet [113], and Clank [46]. Single-backup systems include Hibernus [6], QuickRecall [51], and many others [4, 5, 10, 70, 76]. MOUSE is a multi-backup system as it is constantly saving the architectural state.

Many PIM architectures exist, such as Pinatubo [67], for DRAM with Ambit [104], and for SRAM with Neural Cache [31]. These technologies target traditional memory hierarchies and have not considered intermittent operation. Ambit and Neural Cache are not suitable for energy harvesting as they are volatile technologies. Pinatubo has the potential to be adapted and used similarly as CRAM in MOUSE. However, Pinatubo uses logic external to the memory array for some operations. This adds complexity which is difficult to manage during intermittent execution. Additionally, Pinatubo requires sense amplifiers for every operation, which comes with a high energy cost.

Neural networks [13, 120] and BNNs [109, 124] have been previously mapped to PIM substrates for acceleration, including on CRAM [94]. However, such designs have not considered intermittent computing and would be unsuitable for the beyond edge domain.

A number of high performance and low power accelerator exist, but which have not been adapted for intermittent executio. The Phoenix processor [103] is an extremely low power processor with a sophisticated sleep strategy. A number of accelerators have demonstrated high performance and energy efficiency on inference. PuDianNao [69] is an ASIC accelerator which also targets SVM. The XNOR Neural engine is microcontroller based system for BNN acceleration [21]. An in/near memory SRAM substrate is proposed in [115], which performs bit-serial arithmetic, and which was shown to have high performance and efficiency on the AlexNet [61] network. A number of PIM accelerators also exist, including a BNN accelerator for Cifar-10 image classification [52], an analog SRAM accelerator for MNIST classification [129], and another which does both MNIST and Cifar-10 classification [112]. Adapting such accelerators to support safe intermittent computing is not straight-forward and would likely come –if at all possible– at significant performance and efficiency cost.

Orthogonal to our work, recent papers have made progress on problems relevant in the energy harvesting domain. Low power and accurate time keeping was developed in [24]. SRAM was used as a an efficient check-pointing memory, being able to maintain state for short periods of power off time [117]. A new platform for intermittent computing is proposed in [60] which simplifies the task of adapting pre-existing embedded applications to work in intermittent environments.

9 CONCLUSION

We extended the work of MOUSE [93], a non-volatile processing-in-memory (PIM) accelerator, creating a more hardware efficient and programmable design and enabling proper functionality for orbital deployment. We extended the PIM instruction set and added architectural support for branch instructions, increasing the programmability of the device. We developed more efficient mechanisms for column activation, reducing the complexity of the peripheral circuitry. It was shown that MTJ devices and supporting CMOS circuitry can operate correctly across a wide temperature range. Even when accounting for overhead to maintain resilience against radiation, the architecture has high performance and extreme energy efficiency. Combined with the intermittent safe operation and inherent low-cost checkpointing mechanisms, the final result is a device well suited for use as a satellite in low earth orbit.

REFERENCES

[1] 2019. https://www.everspin.com/supportdocs/EMD3D256M08G1-150CBS1. Accessed: 2019-08-10.
[2] 2019. https://www.everspin.com/family/emd4e001g?npath=3557. Accessed: 2019-11-25.
[3] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. 2013. A public domain dataset for human activity recognition using smartphones.. In *Esann*.
[4] Faycal Ait Aouda, Kevin Marquet, and Guillaume Salagnac. 2014. Incremental checkpointing of program state to NVRAM for transiently-powered systems. In *2014 9th International Symposium on Reconfigurable and Communication-Centric Systems-on-Chip (ReCoSoC)*. IEEE, 1–4.
[5] Domenico Balsamo, Anup Das, Alex S Weddell, Davide Brunelli, Bashir M Al-Hashimi, Geoff V Merrett, and Luca Benini. 2016. Graceful performance modulation for power-neutral transient computing systems. *IEEE Transactions on*

Computer-Aided Design of Integrated Circuits and Systems 35, 5 (2016), 738–749.

[6] Domenico Balsamo, Alex S Weddell, Anup Das, Alberto Rodriguez Arreola, Davide Brunelli, Bashir M Al-Hashimi, Geoff V Merrett, and Luca Benini. 2016. Hibernus++: a self-calibrating and adaptive system for transiently-powered embedded devices. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 35, 12 (2016), 1968–1980.

[7] Domenico Balsamo, Alex S Weddell, Geoff V Merrett, Bashir M Al-Hashimi, Davide Brunelli, and Luca Benini. 2014. Hibernus: Sustaining computation during intermittent supply for energy-harvesting systems. *IEEE Embedded Systems Letters* 7, 1 (2014), 15–18.

[8] Mark Barton and Jon Miller. 2005. Modular Thermal Design Concepts: Thermal Design of a Spacecraft on a Module Level for LEO Missions. (2005).

[9] Robert Baumann. 2005. Soft errors in advanced computer systems. *IEEE Design & Test of Computers* 22, 3 (2005), 258–266.

[10] Gautier Berthou, Tristan Delizy, Kevin Marquet, Tanguy Risset, and Guillaume Salagnac. 2017. Peripheral state persistence for transiently-powered systems. In *2017 Global Internet of Things Summit (GIoTS)*. IEEE, 1–6.

[11] Anantha P Chandrakasan, Denis C Daly, Joyce Kwong, and Yogesh K Ramadass. 2008. Next generation micro-power systems. In *2008 IEEE Symposium on VLSI Circuits*. IEEE, 2–5.

[12] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 27.

[13] Lerong Chen, Jiawen Li, Yiran Chen, Qiuping Deng, Jiyuan Shen, Xiaoyao Liang, and Li Jiang. 2017. Accelerator-friendly neural-network training: Learning variations and defects in RRAM crossbar. In *Proceedings of the Conference on Design, Automation & Test in Europe*. European Design and Automation Association, 19–24.

[14] Zamshed Chowdhury, Jonathan D Harms, S Karen Khatamifard, Masoud Zabihi, Yang Lv, Andrew P Lyle, Sachin S Sapatnekar, Ulya R Karpuzcu, and Jian-Ping Wang. 2017. Efficient in-memory processing using spintronics. *IEEE Computer Architecture Letters* 17, 1 (2017), 42–46.

[15] Alexei Colin and Brandon Lucia. 2016. Chain: tasks and channels for reliable intermittent programs. In *ACM SIGPLAN Notices*, Vol. 51. ACM, 514–530.

[16] Alexei Colin and Brandon Lucia. 2018. Termination checking and task decomposition for task-based intermittent programs. In *Proceedings of the 27th International Conference on Compiler Construction*. 116–127.

[17] Alexei Colin and Brandon Lucia. 2018. Termination checking and task decomposition for task-based intermittent programs. In *Proceedings of the 27th International Conference on Compiler Construction*. ACM, 116–127.

[18] Alexei Colin, Emily Ruppel, and Brandon Lucia. 2018. A reconfigurable energy storage architecture for energy-harvesting devices. In *ACM SIGPLAN Notices*, Vol. 53. ACM, 767–781.

[19] Alexei Colin, Emily Ruppel, and Brandon Lucia. 2018. A reconfigurable energy storage architecture for energy-harvesting devices. In *ACM SIGPLAN Notices*, Vol. 53. ACM, 767–781.

[20] Y Conraux, JP Nozieres, V Da Costa, M Toulemonde, and K Ounadjela. 2003. Effects of swift heavy ion bombardment on magnetic tunnel junction functional properties. *Journal of applied physics* 93, 10 (2003), 7301–7303.

[21] Francesco Conti, Pasquale Davide Schiavone, and Luca Benini. 2018. XNOR Neural Engine: A Hardware Accelerator IP for 21.6-fJ/op Binary Neural Network Inference. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 37, 11 (2018), 2940–2951.

[22] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. 2016. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830* (2016).

[23] Husrev Cılasun, Salonik Resch, Zamshed Iqbal Chowdhury, Erin Olson, Masoud Zabihi, Zhengyang Zhao, Thomas Peterson, Jian-Ping Wang, Sachin S. Sapatnekar, and Ulya Karpuzcu. 2020. CRAFFT: High Resolution FFT Accelerator In Spintronic Computational RAM. In *Proceedings of the 57th Annual ACM/IEEE Design Automation Conference*.

[24] Jasper de Winkel, Carlo Delle Donne, Kasim Sinan Yildirim, Przemysław Pawełczak, and Josiah Hester. 2020. Reliable Timekeeping for Intermittent Computing. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*. 53–67.

[25] Bradley Denby and Brandon Lucia. 2019. Orbital edge computing: Machine inference in space. *IEEE Computer Architecture Letters* 18, 1 (2019), 59–62.

[26] Bradley Denby and Brandon Lucia. 2020. Orbital edge computing: Nanosatellite constellations as a new class of computer system. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*. 939–954.

[27] Zhitao Diao, Alex Panchula, Yunfei Ding, Mahendra Pakala, Shengyuan Wang, Zhanjie Li, Dmytro Apalkov, Hideyasu Nagai, Alexander Driskill-Smith, Lien-Chang Wang, et al. 2007. Spin transfer switching in dual MgO magnetic tunnel junctions. *Applied Physics Letters* 90, 13 (2007), 132508.

[28] Xiangyu Dong, Xiaoxia Wu, Guangyu Sun, Yuan Xie, Helen Li, and Yiran Chen. 2008. Circuit and microarchitecture evaluation of 3D stacking magnetic RAM (MRAM) as a universal memory replacement. In *2008 45th ACM/IEEE Design Automation Conference*. IEEE, 554–559.

[29] Xiangyu Dong, Cong Xu, Yuan Xie, and Norman P Jouppi. 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), 994–1007.

[30] Philipp Dürrenfeld, Felicitas Gerhard, Jonathan Chico, Randy K Dumas, Mojtaba Ranjbar, Anders Bergman, Lars Bergqvist, Anna Delin, Charles Gould, Laurens W Molenkamp, et al. 2015. Tunable damping, saturation magnetization, and exchange stiffness of half-Heusler NiMnSb thin films. *Physical Review B* 92, 21 (2015), 214424.

[31] Charles Eckert, Xiaowei Wang, Jingcheng Wang, Arun Subramaniyan, Ravi Iyer, Dennis Sylvester, David Blaauw, and Reetuparna Das. 2018. Neural cache: Bit-serial in-cache acceleration of deep neural networks. In *Proceedings of the 45th Annual International Symposium on Computer Architecture*. IEEE Press, 383–396.

[32] Karthik Ganesan, Joshua San Miguel, and Natalie Enright Jerger. 2019. The What's Next Intermittent Computing Architecture. In *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 211–223.

[33] Karthik Ganesan, Joshua San Miguel, and Natalie Enright Jerger. 2019. The What's Next Intermittent Computing Architecture. In *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 211–223.

[34] Kevin Garello, Farrukh Yasin, S Couet, Laurent Souriau, J Swerts, S Rao, Simon Van Beek, Wonsub Kim, Enlong Liu, S Kundu, et al. 2018. SOT-MRAM 300mm integration for low power and ultrafast embedded memories. In *2018 IEEE Symposium on VLSI Circuits*. IEEE, 81–82.

[35] Simone Gerardin and Alessandro Paccagnella. 2010. Present and future non-volatile memories for space. *IEEE Transactions on Nuclear Science* 57, 6 (2010), 3016–3039.

[36] Graham Gobieski, Nathan Beckmann, and Brandon Lucia. 2018. Intermittent Deep Neural Network Inference.

[37] Graham Gobieski, Brandon Lucia, and Nathan Beckmann. 2019. Intelligence beyond the edge: Inference on intermittent embedded systems. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*. 199–213.

[38] Graham Gobieski, Amolak Nagi, Nathan Serafin, Mehmet Meric Isgenc, Nathan Beckmann, and Brandon Lucia. 2019. Manic: A vector-dataflow architecture for ultra-low-power embedded systems. In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*. 670–684.

[39] Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers. 2016. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging* 35, 5 (2016), 1153–1159.

[40] Matthew R Guthaus, Jeffrey S Ringenberg, Dan Ernst, Todd M Austin, Trevor Mudge, and Richard B Brown. 2001. MiBench: A free, commercially representative embedded benchmark suite. In *Proceedings of the fourth annual IEEE international workshop on workload characterization. WWC-4 (Cat. No. 01EX538)*. IEEE, 3–14.

[41] Ramesh Harjani and Saurabh Chaubey. 2014. A unified framework for capacitive series-parallel DC-DC converter design. In *Proceedings of the IEEE 2014 Custom Integrated Circuits Conference*. IEEE, 1–8.

[42] Josiah Hester, Travis Peters, Tianlong Yun, Ronald Peterson, Joseph Skinner, Bhargav Golla, Kevin Storer, Steven Hearndon, Kevin Freeman, Sarah Lord, et al. 2016. Amulet: An energy-efficient, multi-application wearable platform. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*. ACM, 216–229.

[43] Josiah Hester, Lanny Sitanayah, and Jacob Sorber. 2015. Tragedy of the coulombs: Federating energy storage for tiny, intermittently-powered sensors. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*. ACM, 5–16.

[44] Josiah Hester and Jacob Sorber. 2017. Flicker: Rapid prototyping for the batteryless internet-of-things. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. ACM, 19.

[45] Josiah Hester, Kevin Storer, and Jacob Sorber. 2017. Timely execution on intermittently powered batteryless sensors. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. ACM, 17.

[46] Matthew Hicks. 2017. Clank: Architectural support for intermittent computation. In *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 228–240.

[47] G Hu, JH Lee, JJ Nowak, JZ Sun, J Harms, A Annunziata, S Brown, W Chen, YH Kim, G Lauer, et al. 2015. STT-MRAM with double magnetic tunnel junctions. In *2015 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 26–3.

[48] Xiao-Di Huang and et.al. 2020. Forming-free, fast, uniform, and high endurance resistive switching from cryogenic to high temperatures in W/AlO x/Al 2 O 3/Pt bilayer memristor. *IEEE Electron Device Letters* 41, 4 (2020), 549–552.

[49] Harold Hughes, Konrad Bussmann, Patrick J McMarr, Shu-Fan Cheng, Robert Shull, Andrew P Chen, Simon Schafer, Tim Mewes, Adrian Ong, Eugene Chen, et al. 2012. Radiation studies of spin-transfer torque materials and devices. *IEEE Transactions on Nuclear Science* 59, 6 (2012), 3027–3033.

[50] Guenole Jan, Luc Thomas, Son Le, Yuan-Jen Lee, Huanlong Liu, Jian Zhu, Ru-Ying Tong, Keyu Pi, Yu-Jen Wang, Dongna Shen, et al. 2014. Demonstration of fully functional 8Mb perpendicular STT-MRAM chips with sub-5ns

writing for non-volatile embedded memories. In *2014 Symposium on VLSI Technology (VLSI-Technology): Digest of Technical Papers*. IEEE, 1–2.

[51] Hrishikesh Jayakumar, Arnab Raha, and Vijay Raghunathan. 2014. QuickRecall: A low overhead HW/SW approach for enabling computations across power cycles in transiently powered computers. In *2014 27th International Conference on VLSI Design and 2014 13th International Conference on Embedded Systems*. IEEE, 330–335.

[52] Hongyang Jia, H Valavi, Y Tang, J Zhang, and N Verma. 2019. A Programmable Embedded Microprocessor for Bit-scalable In-memory Computing. In *2019 IEEE Hot Chips 31 Symposium (HCS)*. IEEE, 1–29.

[53] Wanyeong Jung, Sechang Oh, Suyoung Bang, Yoonmyung Lee, Dennis Sylvester, and David Blaauw. 2014. 23.3 A 3nW fully integrated energy harvester based on self-oscillating switched-capacitor DC-DC converter. In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*. IEEE, 398–399.

[54] Holger Kappert, Norbert Kordas, Stefan Dreiner, Uwe Paschen, and Rainer Kokozinski. 2015. High temperature SOI CMOS technology and circuit realization for applications up to 300° C. In *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 1162–1165.

[55] Sangkil Kim, Rushi Vyas, Jo Bito, Kyriaki Niotaki, Ana Collado, Apostolos Georgiadis, and Manos M Tentzeris. 2014. Ambient RF energy-harvesting technologies for self-sustainable standalone wireless sensor platforms. *Proc. IEEE* 102, 11 (2014), 1649–1666.

[56] Sangkil Kim, Rushi Vyas, Jo Bito, Kyriaki Niotaki, Ana Collado, Apostolos Georgiadis, and Manos M Tentzeris. 2014. Ambient RF energy-harvesting technologies for self-sustainable standalone wireless sensor platforms. *Proc. IEEE* 102, 11 (2014), 1649–1666.

[57] Daisuke Kobayashi, Kazuyuki Hirose, Takahiro Makino, Shinobu Onoda, Takeshi Ohshima, Shoji Ikeda, Hideo Sato, Eli Christopher Inocencio Enobio, Tetsuo Endoh, and Hideo Ohno. 2017. Soft errors in 10-nm-scale magnetic tunnel junctions exposed to high-energy heavy-ion radiation. *Japanese Journal of Applied Physics* 56, 8 (2017), 0802B4.

[58] Daisuke Kobayashi, Yuya Kakehashi, Kazuyuki Hirose, Shinobu Onoda, Takahiro Makino, Takeshi Ohshima, Shoji Ikeda, Michihiko Yamanouchi, Hideo Sato, Eli Christopher Enobio, et al. 2014. Influence of heavy ion irradiation on perpendicular-anisotropy CoFeB-MgO magnetic tunnel junctions. *IEEE Transactions on Nuclear Science* 61, 4 (2014), 1710–1716.

[59] Ron Kohavi. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.. In *Kdd*, Vol. 96. Citeseer, 202–207.

[60] Vito Kortbeek, Kasim Sinan Yildirim, Abu Bakar, Jacob Sorber, Josiah Hester, and Przemysław Pawełczak. 2020. Time-sensitive Intermittent Computing Meets Legacy Software. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*. 85–99.

[61] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[62] Lili Lang and et.al. 2020. A low temperature functioning CoFeB/MgO-based perpendicular magnetic tunnel junction for cryogenic nonvolatile random access memory. *Applied Physics Letters* 116, 2 (2020).

[63] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[64] SH Lee, HC Park, MS Kim, HW Kim, MR Choi, HG Lee, JW Seo, SC Kim, SG Kim, SB Hong, et al. 2011. Highly productive PCRAM technology platform and full chip operation: Based on 4F 2 (84nm pitch) cell scheme for 1 Gb and beyond. In *2011 International Electron Devices Meeting*. IEEE, 3–3.

[65] LEO Temperatures 2021. https://www.oreilly.com/library/view/diy-satellite-platforms/9781449312756/ch01s05.html. Accessed: 2021-07-14.

[66] Vladimir Leonov. 2013. Thermoelectric energy harvesting of human body heat for wearable sensors. *IEEE Sensors Journal* 13, 6 (2013), 2284–2291.

[67] Shuangchen Li, Cong Xu, Qiaosha Zou, Jishen Zhao, Yu Lu, and Yuan Xie. 2016. Pinatubo: A processing-in-memory architecture for bulk bitwise operations in emerging non-volatile memories. In *Proceedings of the 53rd Annual Design Automation Conference*. ACM, 173.

[68] Shuang Liang, Shouyi Yin, Leibo Liu, Wayne Luk, and Shaojun Wei. 2018. FP-BNN: Binarized neural network on FPGA. *Neurocomputing* 275 (2018), 1072–1086.

[69] Daofu Liu, Tianshi Chen, Shaoli Liu, Jinhong Zhou, Shengyuan Zhou, Olivier Teman, Xiaobing Feng, Xuehai Zhou, and Yunji Chen. 2015. Pudiannao: A polyvalent machine learning accelerator. In *ACM SIGARCH Computer Architecture News*, Vol. 43. ACM, 369–381.

[70] Qingrui Liu and Changhee Jung. 2016. Lightweight hardware support for transparent consistency-aware checkpointing in intermittent energy-harvesting systems. In *2016 5th Non-Volatile Memory Systems and Applications Symposium (NVMSA)*. IEEE, 1–6.

[71] Yongpan Liu, Zewei Li, Hehe Li, Yiqun Wang, Xueqing Li, Kaisheng Ma, Shuangchen Li, Meng-Fan Chang, Sampson John, Yuan Xie, et al. 2015. Ambient energy harvesting nonvolatile processors: from circuit to system. In *Proceedings*

of the 52nd Annual Design Automation Conference. ACM, 150.

[72] R Lowther, W Morris, D Gifford, D Duff, and R Fuller. 2011. Latchup Immunity in High Temperature Bulk CMOS Devices. *Additional Papers and Presentations* 2011, HITEN (2011), 000215–000220.

[73] Brandon Lucia, Vignesh Balaji, Alexei Colin, Kiwan Maeng, and Emily Ruppel. 2017. Intermittent computing: Challenges and opportunities. In *2nd Summit on Advances in Programming Languages (SNAPL 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[74] Brandon Lucia, Brad Denby, Zachary Manchester, Harsh Desai, Emily Ruppel, and Alexei Colin. 2021. Computational Nanosatellite Constellations: Opportunities and Challenges. *GetMobile: Mobile Computing and Communications* 25, 1 (2021), 16–23.

[75] Brandon Lucia and Benjamin Ransford. 2015. A simpler, safer programming and execution model for intermittent systems. In *ACM SIGPLAN Notices*, Vol. 50. ACM, 575–585.

[76] Giedrius Lukosevicius, Alberto Rodriguez Arreola, and Alex S Weddell. 2017. Using sleep states to maximize the active time of transient computing systems. In *Proceedings of the Fifth ACM International Workshop on Energy Harvesting and Energy-Neutral Sensing Systems*. ACM, 31–36.

[77] Kaisheng Ma, Xueqing Li, Jinyang Li, Yongpan Liu, Yuan Xie, Jack Sampson, Mahmut Taylan Kandemir, and Vijaykrishnan Narayanan. 2017. Incidental computing on IoT nonvolatile processors. In *2017 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 204–218.

[78] Kaisheng Ma, Xueqing Li, Huichu Liu, Xiao Sheng, Yiqun Wang, Karthik Swaminathan, Yongpan Liu, Yuan Xie, John Sampson, and Vijaykrishnan Narayanan. 2017. Dynamic power and energy management for energy harvesting nonvolatile processor systems. *ACM Transactions on Embedded Computing Systems (TECS)* 16, 4 (2017), 1–23.

[79] Kaisheng Ma, Yang Zheng, Shuangchen Li, Karthik Swaminathan, Xueqing Li, Yongpan Liu, Jack Sampson, Yuan Xie, and Vijaykrishnan Narayanan. 2015. Architecture exploration for ambient energy harvesting nonvolatile processors. In *2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 526–537.

[80] Kiwan Maeng, Alexei Colin, and Brandon Lucia. 2017. Alpaca: intermittent execution without checkpoints. *Proceedings of the ACM on Programming Languages* 1, OOPSLA (2017), 96.

[81] Kiwan Maeng and Brandon Lucia. 2018. Adaptive dynamic checkpointing for safe efficient intermittent computing. In *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*. 129–144.

[82] Milos Manic, Kasun Amarasinghe, Juan J Rodriguez-Andina, and Craig Rieger. 2016. Intelligent buildings of the future: Cyberaware, deep learning powered, and human interacting. *IEEE Industrial Electronics Magazine* 10, 4 (2016), 32–49.

[83] S Mizukami, D Watanabe, M Oogane, Y Ando, Y Miura, M Shirai, and T Miyazaki. 2009. Low damping constant for Co 2 FeAl Heusler alloy films and its correlation with density of states. *Journal of Applied Physics* 105, 7 (2009), 07D306.

[84] Hiroki Noguchi, Kazutaka Ikegami, Keiichi Kushida, Keiko Abe, Shogo Itai, Satoshi Takaya, Naoharu Shimomura, Junichi Ito, Atsushi Kawasumi, Hiroyuki Hara, et al. 2015. 7.5 A 3.3 ns-access-time 71.2 μW/MHz 1Mb embedded STT-MRAM using physically eliminated read-disturb scheme and normally-off memory architecture. In *2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers*. IEEE, 1–3.

[85] Fabian Oboril, Rajendra Bishnoi, Mojtaba Ebrahimi, and Mehdi B Tahoori. 2015. Evaluation of hybrid memory technologies using SOT-MRAM for on-chip cache hierarchy. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 34, 3 (2015), 367–380.

[86] Bishnu Patra and et.al. 2017. Cryo-CMOS circuits and systems for quantum computing applications. *IEEE Journal of Solid-State Circuits* 53, 1 (2017).

[87] Keni Qiu, Nicholas Jao, Mengying Zhao, Cyan Subhra Mishra, Gulsum Gudukbay, Sethu Jose, Jack Sampson, Mahmut Taylan Kandemir, and Vijaykrishnan Narayanan. 2020. ResiRCA: A resilient energy harvesting ReRAM crossbar-based accelerator for intelligent embedded processors. In *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 315–327.

[88] R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[89] Yogesh K Ramadass and Anantha P Chandrakasan. 2007. Voltage scalable switched capacitor DC-DC converter for ultra-low-power on-chip applications. In *2007 IEEE Power Electronics Specialists Conference*. IEEE, 2353–2359.

[90] Benjamin Ransford, Jacob Sorber, and Kevin Fu. 2011. Mementos: system support for long-running computation on RFID-scale devices. In *ACM SIGARCH Computer Architecture News*, Vol. 39. ACM, 159–170.

[91] Fanghui Ren, Albrecht Jander, Pallavi Dhagat, and Cathy Nordman. 2012. Radiation tolerance of magnetic tunnel junctions with MgO tunnel barriers. *IEEE Transactions on Nuclear Science* 59, 6 (2012), 3034–3038.

[92] Salonik Resch, Husrev Cilasun, and Ulya Karpuzcu. 2021. Cryogenic PIM: Challenges & Opportunities. *IEEE Computer Architecture Letters* (2021).

[93] Salonik Resch, S Karen Khatamifard, Zamshed I Chowdhury, Masoud Zabihi, Zhengyang Zhao, Husrev Cilasun, Jian-Ping Wang, Sachin S Sapatnekar, and Ulya R Karpuzcu. 2020. MOUSE: Inference In Non-volatile Memory for Energy Harvesting Applications. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture*

(MICRO). IEEE, 400–414.

[94] Salonik Resch, S Karen Khatamifard, Zamshed Iqbal Chowdhury, Masoud Zabihi, Zhengyang Zhao, Jian-Ping Wang, Sachin S Sapatnekar, and Ulya R Karpuzcu. 2019. PIMBALL: Binary Neural Networks in Spintronic Memory. *ACM Transactions on Architecture and Code Optimization (TACO)* 16, 4 (2019), 41.

[95] Emily Ruppel and Brandon Lucia. 2019. Transactional concurrency control for intermittent, energy-harvesting computing systems. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*. 1085–1100.

[96] Emily Ruppel and Brandon Lucia. 2019. Transactional concurrency control for intermittent, energy-harvesting computing systems. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*. ACM, 1085–1100.

[97] Daisuke Saida, Saori Kashiwada, Megumi Yakabe, Tadaomi Daibou, Naoki Hase, Miyoshi Fukumoto, Shinji Miwa, Yoshishige Suzuki, Hiroki Noguchi, Shinobu Fujita, et al. 2016. Sub-3 ns pulse with sub-100 μA switching of 1x–2x nm perpendicular MTJ for high-performance embedded STT-MRAM towards sub-20 nm CMOS. In *2016 IEEE Symposium on VLSI Technology*. IEEE, 1–2.

[98] Daisuke Saida, Saori Kashiwada, Megumi Yakabe, Tadaomi Daibou, Naoki Hase, Miyoshi Fukumoto, Shinji Miwa, Yoshishige Suzuki, Hiroki Noguchi, Shinobu Fujita, et al. 2016. Sub-3 ns pulse with sub-100 μA switching of 1x–2x nm perpendicular MTJ for high-performance embedded STT-MRAM towards sub-20 nm CMOS. In *2016 IEEE Symposium on VLSI Technology*. IEEE, 1–2.

[99] Alanson P Sample, Daniel J Yeager, Pauline S Powledge, Alexander V Mamishev, and Joshua R Smith. 2008. Design of an RFID-based battery-free programmable sensing platform. *IEEE transactions on instrumentation and measurement* 57, 11 (2008), 2608–2615.

[100] Joshua San Miguel, Karthik Ganesan, Mario Badr, Chunqiu Xia, Rose Li, Hsuan Hsiao, and Natalie Enright Jerger. 2018. The EH Model: Early Design Space Exploration of Intermittent Processor Architectures. In *2018 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 600–612.

[101] H Sato, ECI Enobio, M Yamanouchi, S Ikeda, S Fukami, S Kanai, F Matsukura, and H Ohno. 2014. Properties of magnetic tunnel junctions with a MgO/CoFeB/Ta/CoFeB/MgO recording structure down to junction diameter of 11 nm. *Applied Physics Letters* 105, 6 (2014), 062403.

[102] Selahattin Sayil. 2019. A survey of circuit-level soft error mitigation methodologies. *Analog Integrated Circuits and Signal Processing* 99, 1 (2019), 63–70.

[103] Mingoo Seok, Scott Hanson, Yu-Shiang Lin, Zhiyoong Foo, Daeyeon Kim, Yoonmyung Lee, Nurrachman Liu, Dennis Sylvester, and David Blaauw. 2008. The Phoenix Processor: A 30pW platform for sensor applications. In *2008 IEEE Symposium on VLSI Circuits*. IEEE, 188–189.

[104] Vivek Seshadri, Donghyuk Lee, Thomas Mullins, Hasan Hassan, Amirali Boroumand, Jeremie Kim, Michael A Kozuch, Onur Mutlu, Phillip B Gibbons, and Todd C Mowry. 2017. Ambit: In-memory accelerator for bulk bitwise operations using commodity DRAM technology. In *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture*. ACM, 273–287.

[105] M Shin and et.al. 2014. Low temperature characterization of 14nm FDSOI CMOS devices. In *WOLTE 2014*.

[106] Fang Su, Wei-Hao Chen, Lixue Xia, Chieh-Pu Lo, Tianqi Tang, Zhibo Wang, Kuo-Hsiang Hsu, Ming Cheng, Jun-Yi Li, Yuan Xie, et al. 2017. A 462GOPs/J RRAM-based nonvolatile intelligent processor for energy harvesting IoE system featuring nonvolatile logics and processing-in-memory. In *2017 Symposium on VLSI Technology*. IEEE, T260–T261.

[107] Xiaoyu Sun, Xiaochen Peng, Pai-Yu Chen, Rui Liu, Jae-sun Seo, and Shimeng Yu. 2018. Fully parallel RRAM synaptic array for implementing binary neural network with (+ 1,- 1) weights and (+ 1, 0) neurons. In *Proceedings of the 23rd Asia and South Pacific Design Automation Conference*. IEEE Press, 574–579.

[108] Tianqi Tang, Lixue Xia, Boxun Li, Yu Wang, and Huazhong Yang. 2017. Binary convolutional neural network on rram. In *Design Automation Conference (ASP-DAC), 2017 22nd Asia and South Pacific*. IEEE, 782–787.

[109] Tianqi Tang, Lixue Xia, Boxun Li, Yu Wang, and Huazhong Yang. 2017. Binary convolutional neural network on RRAM. In *2017 22nd Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE, 782–787.

[110] UCI Machine Learning Repository 2019. https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+\using+smartphones. Accessed: 2019-06-02.

[111] Yaman Umuroglu, Nicholas J Fraser, Giulio Gambardella, Michaela Blott, Philip Leong, Magnus Jahre, and Kees Vissers. 2017. Finn: A framework for fast, scalable binarized neural network inference. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 65–74.

[112] Hossein Valavi, Peter J Ramadge, Eric Nestler, and Naveen Verma. 2019. A 64-tile 2.4-Mb in-memory-computing CNN accelerator employing charge-domain compute. *IEEE Journal of Solid-State Circuits* 54, 6 (2019), 1789–1799.

[113] Joel Van Der Woude and Matthew Hicks. 2016. Intermittent computation without hardware support or programmer intervention. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 17–32.

[114] Deepak Vasisht, Zerina Kapetanovic, Jongho Won, Xinxin Jin, Ranveer Chandra, Sudipta Sinha, Ashish Kapoor, Madhusudhan Sudarshan, and Sean Stratman. 2017. Farmbeats: An iot platform for data-driven agriculture. In *14th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 17)*. 515–529.

[115] Jingcheng Wang, Xiaowei Wang, Charles Eckert, Arun Subramaniyan, Reetuparna Das, David Blaauw, and Dennis Sylvester. 2019. A 28-nm Compute SRAM With Bit-Serial Logic/Arithmetic Operations for Programmable In-Memory Vector Computing. *IEEE Journal of Solid-State Circuits* 55, 1 (2019), 76–86.

[116] Jian-Ping Wang, Mahdi Jamaliz, Angeline Klemm Smith, and Zhengyang Zhao. 2016. Magnetic tunnel junction based integrated logics and computational circuits. *Nanomagnetic and Spintronic Devices for Energy-Efficient Memory and Computing* (2016), 133.

[117] Harrison Williams, Xun Jian, and Matthew Hicks. 2020. Forget Failure: Exploiting SRAM Data Remanence for Low-overhead Intermittent Computation. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*. 69–84.

[118] H-S Philip Wong, Heng-Yuan Lee, Shimeng Yu, Yu-Sheng Chen, Yi Wu, Pang-Shiu Chen, Byoungil Lee, Frederick T Chen, and Ming-Jinn Tsai. 2012. Metal–oxide RRAM. *Proc. IEEE* 100, 6 (2012), 1951–1970.

[119] Lixue Xia, Tianqi Tang, Wenqin Huangfu, Ming Cheng, Xiling Yin, Boxun Li, Yu Wang, and Huazhong Yang. 2016. Switched by input: power efficient structure for RRAM-based convolutional neural network. In *Proceedings of the 53rd Annual Design Automation Conference*. ACM, 125.

[120] Lixue Xia, Tianqi Tang, Wenqin Huangfu, Ming Cheng, Xiling Yin, Boxun Li, Yu Wang, and Huazhong Yang. 2016. Switched by input: Power efficient structure for RRAM-based convolutional neural network. In *2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*. IEEE, 1–6.

[121] Jeng-Bang Yau and et.al. 2017. Hybrid Cryogenic Memory Cells for Superconducting Computing Applications. In *ICRC 2017*.

[122] MUSTAFA BERKE Yelten. [n.d.]. Cryogenic DC Characteristics of Low Threshold Voltage (VTH) n-channel MOSFETs. *Balkan Journal of Electrical and Computer Engineering* 7, 3 ([n. d.]).

[123] Shimeng Yu, Zhiwei Li, Pai-Yu Chen, Huaqiang Wu, Bin Gao, Deli Wang, Wei Wu, and He Qian. 2016. Binary neural network with 16 Mb RRAM macro chip for classification and online training. In *Electron Devices Meeting (IEDM), 2016 IEEE International*. IEEE, 16–2.

[124] Shimeng Yu, Zhiwei Li, Pai-Yu Chen, Huaqiang Wu, Bin Gao, Deli Wang, Wei Wu, and He Qian. 2016. Binary neural network with 16 Mb RRAM macro chip for classification and online training. In *2016 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 16–2.

[125] L Yuan and et.al. 2006. Temperature dependence of magnetoresistance in magnetic tunnel junctions with different free layer structures. *Physical Review B* 73, 13 (2006).

[126] Masoud Zabihi, Zamshed Iqbal Chowdhury, Zhengyang Zhao, Ulya R Karpuzcu, Jian-Ping Wang, and Sachin S Sapatnekar. 2018. In-memory processing on the spintronic CRAM: From hardware design to application mapping. *IEEE Trans. Comput.* 68, 8 (2018), 1159–1173.

[127] Masoud Zabihi, Arvind K Sharma, Meghna G Mankalale, Zamshed Iqbal Chowdhury, Zhengyang Zhao, Salonik Resch, Ulya R Karpuzcu, Jian-Ping Wang, and Sachin S Sapatnekar. 2020. Analyzing the Effects of Interconnect Parasitics in the STT CRAM In-memory Computational Platform. *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits* 6, 1 (2020), 71–79.

[128] Masoud Zabihi, Zhengyang Zhao, DC Mahendra, Zamshed I Chowdhury, Salonik Resch, Thomas Peterson, Ulya R Karpuzcu, Jian-Ping Wang, and Sachin S Sapatnekar. 2019. Using spin-Hall MTJs to build an energy-efficient in-memory computation platform. In *20th International Symposium on Quality Electronic Design (ISQED)*. IEEE, 52–57.

[129] Jintao Zhang and Naveen Verma. 2019. An In-memory-Computing DNN Achieving 700 TOPS/W and 6 TOPS/mm 2 in 130-nm CMOS. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9, 2 (2019), 358–366.

[130] Quming Zhou and Kartik Mohanram. 2005. Gate sizing to radiation harden combinational logic. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 25, 1 (2005), 155–166.