MINERGY: A 19µW @ 4MHz, 256 MOPS/mW, RISC-V microcontroller with embedded MRAM main memory and vector-dataflow co-processor in 22nm bulk finFET CMOS

***Abstract*—Whether powered by a battery or energy harvested from the environment, low-power (LP) sensor devices are becoming pervasive and require extreme energy efficiency. We present MINERGY, an energy-efficient microcontroller (MCU) augmented with a vector-dataflow (VDF) co-processor. MINERGY is 60% more energy efficient than a low-power scalar MCU, achieving peak efficiency of 256 MOPS/mW (2.6× prior work) while consuming only 19.1µW (@4MHz). To make the system viable for intermittently powered applications that require non-volatile storage, MINERGY includes a 256KB embedded MRAM.**

I. INTRODUCTION

Emerging sensing applications demand extreme energy efficiency to enable sophisticated computation in remote environments, e.g., for on-device machine learning. Existing programmable devices waste significant energy on supplying instructions and data. MINERGY improves energy efficiency by *changing the execution model* to avoid wasting energy.

Fig. 1 illustrates the execution models of a baseline low-power, scalar MCU, a simple vector design, and MINERGY. Blue arrows denote data flow, and orange arrows denote control flow. The baseline MCU requires one instruction per operation, burning energy in i-fetch and register-file (RF) access. By contrast, each vector instruction applies an operation to multiple data items, amortizing i-fetch. But vector execution still communicates through vector registers (VRF). To minimize VRF energy, MINERGY implements *vector-dataflow execution* [2]. With VDF execution, MINERGY operates on a window of SIMD instructions, identifying dependencies between operations and forwarding intermediate values directly from producers to consumers. Forwarded values are not read from or written to the VRF.

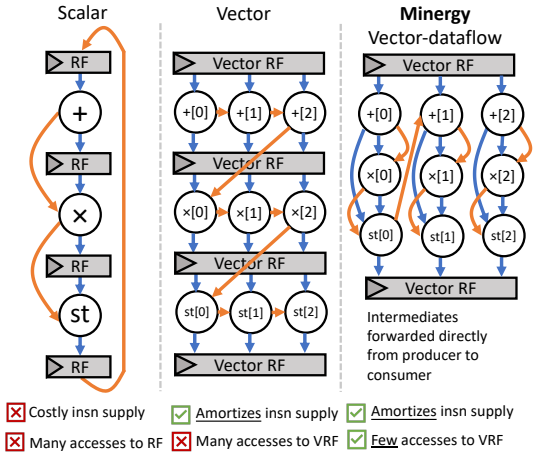


Figure 1: Comparison of execution models. In MINERGY's *vector-dataflow execution*, vectors reduce instruction-supply energy and dataflow forwarding reduces data-supply energy.

II. EXECUTION MODEL AND MICROARCHITECTURE

Fig. 2 shows MINERGY's microarchitecture, split along the two phases of execution: Decode & Rename and Execute.

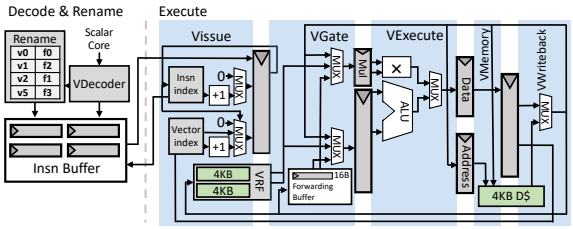


Figure 2: Pipeline diagram of MINERGY.

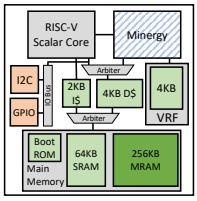


Figure 3: Block diagram.

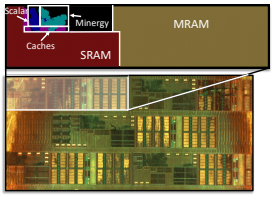


Figure 4: Die shot.

During the Decode & Rename Phase, MINERGY buffers a window of decoded instructions ("Insn Buffer") and identifies dataflow between them. Then, where dataflow allows, MINERGY renames operands to point to the "Forwarding Buffer" (a 16B buffer) instead of the 4KB VRF.

The Execute Phase begins once the instruction buffer is full, a `vfence` instruction is reached, or the forwarding buffer is fully allocated. MINERGY has a five-stage execution pipeline. Unlike conventional vector execution, MINERGY's VDF executes a *series of operations per element* before going to the next element (Fig. 1). This execution order minimizes the number of in-flight values so that they fit in the 16B Forwarding Buffer. VIssue tracks execution progress and maintains a pointer into the instruction buffer, decodes instructions, and (only if necessary) initiates VRF reads; VGate determines the source for each operand (the VRF, Forwarding Buffer, or bypass paths) and steers operands to the multiplier or ALU; VExecute computes the ALU and multiplier results; VMemory issues loads and stores; and VWriteback writes results to the Forwarding Buffer or VRF, as appropriate. VGate reduces switching activity in VExecute by steering operands to dedicated input registers for the ALU or multiplier to, e.g., prevent a VADD from toggling the multiplier. This is important because, unlike conventional vector execution, the active instruction changes every cycle in MINERGY, increasing activity on control and data signals.

III. CHIP DESIGN

Fig. 3 shows a block diagram of the MINERGY system-on-chip. MINERGY comprises a low-power-optimized RISC-V (RV32IMEC) MCU, 2KB instruction cache, 4KB data cache (shared with the coprocessor), IO-bus supporting I^2C and GPIO, main-memory (1KB boot ROM, 64KB SRAM, and 256KB MRAM), and VDF co-processor with 4KB 1r1w VRF.

MINERGY was fabricated in a 22nm bulk finFET process with an area of $0.57mm^2$. Fig. 4 shows the die photo of the 4mm × 8mm testchip. MINERGY has separate power domains for SRAM, MRAM, and logic that can be controlled and

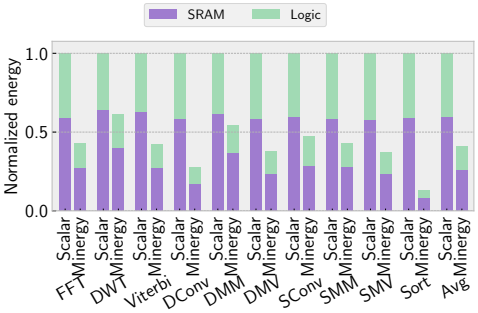


Figure 5: Energy (normalized to scalar).

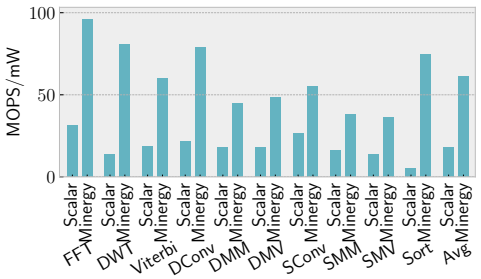


Figure 6: Energy efficiency.

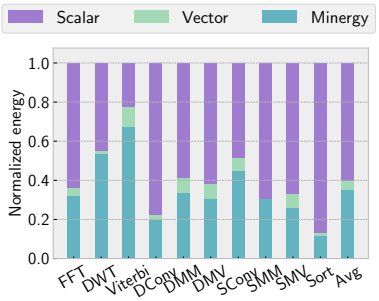


Figure 7: Scalar vs. vector vs. MINERGY.

	2017 [4]	2018 [3]	2019 [1]	2020 [5]	This work
Architecture	Scalar & Vector	Scalar	Scalar	Scalar w/ SIMD ext.	Scalar & Vector-dataflow
ISA	RISC-V	Thumb-2	Thumb-2	Thumb-2	RISC-V
Process (nm)	28 FD-SOI	14 Tri-gate	28 FD-SOI	65 LP	22 bulk FF
Core Area (mm2)	1.07	6.25	0.675	6	0.57
Voltage (V)	0.48-1.0	0.4-1.0	0.4	0.4-0.75	0.4-1.05 Core 1.1 MRAM
Frequency (MHz)	20-797	0.2-950	40-80	0.8-38	4-48.9
Memory (KB)	56KB SRAM	64+64+384 SRAM	32+32 SRAM	128 ROM, 16+4 SRAM	64 SRAM, 256 MRAM
Power Budget (mW)	1-200	1-20	1	1-4	0.019-2 w/o MRAM 1-2 w/ MRAM
Average Power (µW)	50000	80	144	47	19.1 w/o MRAM @ 4MHz[1] 1.7mW w/ MRAM @ 49MHz[1]
Peak Efficiency (MOPS/mW)[2]	41.8 MFlops/mW	Not reported	97	Not reported	256 w/o MRAM 33.2 w/ MRAM
Best Active Energy (pJ/Cycle)	Not reported	6.2	3	10.9	3.7 w/o MRAM 29 w/ MRAM

[1] Over all benchmarks [2] 32b operations

Table I: MINERGY vs. prior work. MINERGY is 2.6× more efficient than prior work, achieving 256 MOPS/mW (@19μW & 4MHz).

measured independently. MINERGY is optimized to run with a 4MHz to 50MHz clock from an on-die clock generator at 0.4V to 1.0V logic, 0.4V to 1.0V SRAM, and 1.10V MRAM.

IV. EVALUATION

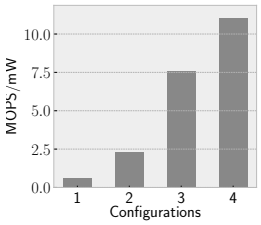
We evaluate MINERGY across ten benchmarks with random 32b inputs. Fig. 5 shows energy normalized to the baseline low-power scalar MCU, and Fig. 6 shows energy efficiency (MOPs/mW). All results were collected with MRAM disabled and core voltage at 0.4V. On average, MINERGY reduces energy by 60% vs. the baseline and achieves 60MOPS/mW at 19µW.

MINERGY is energy-efficient: Table I compares MINERGY with prior work [1, 3, 5]. MINERGY was designed for energy-minimal, low-power operation: MINERGY consumes 19µW at 4MHz, significantly lower than prior work. MINERGY is more energy-efficient than prior work (by 2.6×), with a peak efficiency of 256 MOPS/mW (vector increment, 32b ops) and 3.7pJ/cycle at 0.4V, 4MHz, room temperature, and MRAM disabled. With random inputs, which cause unrealistic, near worst-case toggling of data lines, MINERGY gets 45 MOPs/mW on dense matrix-matrix multiplication (DMM).

Vector-dataflow uses less energy than vector: To make a fair comparison between VDF and vector execution models, we also taped out an alternative SoC design with an optimized vector co-processor in the same process technology. Fig. 7 overlays

Size (KB)	256
Area (mm^2)	0.31
Voltage (V)	1.1
Leakage (µW)	663
32b Read Latency @ 50 MHz (ns)	170
32b Write Latency @ 50 MHz (µs)	8.4
32b Read Energy (pJ)	437
32b Write Energy (nJ)	29.7
Read Energy (pJ/bit)	13.7
Write Energy (pJ/bit)	929

(a) *MRAM characterization.*



(b) *MOPS/mW for DMM.*

1: Running from MRAM, DCache enabled, 48.9 MHz, 0.64V Core
2: Running from MRAM, DCache disabled, 231 MHz, 1.0V Core
3: Running from SRAM, MRAM enabled, DCache enabled, 48.9 MHz, 0.64V Core
4: Running from SRAM, MRAM enabled, DCache disabled, 166 MHz, 1.0V Core

Figure 8: MRAM characterization & case study on DMM.

the normalized energies of MINERGY (Blue), the vector design (Green), and the scalar baseline (Purple). Vector execution already achieves state-of-the-art energy-efficiency, using 54% less than the scalar MCU; MINERGY's VDF execution reduces energy by a further 12%.

MRAM characterization: Fig. 8 characterizes the embedded MRAM and presents a case study of designs with MRAM enabled. MRAM leakage is 663µW, reads take 170ns and 13.7pJ/bit, while writes take 8.4µs and 929pJ/bit. Write latency is independent of clock frequency. A case study of DMM puts these numbers into context. The figure includes several system configurations: 1) MINERGY running out of MRAM with the DCache enabled @49MHz, 2) MINERGY running out of MRAM as fast as possible @231MHz (this necessitates the DCache being disabled), 3) MINERGY running from SRAM, DCache enabled, and MRAM enabled @49MHz, and 4) MINERGY running as fast as possible @166MHz (w/o DCache) and MRAM enabled. Configuration 4 achieves max efficiency with 11MOPS/mW and configuration 2 achieves max efficiency for running from MRAM with 2.3MOPS/mW. As found in prior low-power systems, MRAM's high static power is a significant challenge for energy efficiency. Addressing this challenge is the subject of future work.

REFERENCES

[1] D. Bol *et al.*, "19.6 a 40-to-80mhz sub-4μw/mhz ulv cortex-m0 mcu soc in 28nm fdsoi with dual-loop adaptive back-bias generator for 20μs wake-up from deep fully retentive sleep mode," in *ISSCC*, 2019.

[2] G. Gobieski *et al.*, "Manic: A vector-dataflow architecture for ultra-low-power embedded systems," in *MICRO*, 2019.

[3] T. Karnik *et al.*, "A cm-scale self-powered intelligent and secure iot edge mote featuring an ultra-low-power soc in 14nm tri-gate cmos," in *ISSCC*, 2018.

[4] B. Keller *et al.*, "A risc-v processor soc with integrated power management at submicrosecond timescales in 28 nm fd-soi," *JSSC*, 2017.

[5] P. Prabhat *et al.*, "27.2 m0n0: A performance-regulated 0.8-to-38mhz dvfs arm cortex-m33 simd mcu with 10nw sleep power," in *ISSCC*, 2020.