

VMXDOTP: A RISC-V Vector ISA Extension for Efficient Microscaling (MX) Format Acceleration

Max Wipfli*, Gamze İslamoğlu*, Navaneeth Kunhi Purayil*, Angelo Garofalo[†] and Luca Benini*[†]

*IIS, ETH Zurich, Switzerland; [†]DEI, University of Bologna, Italy

mwwipfli@ethz.ch, {gislamoglu,nkunhi,lbenini}@iis.ee.ethz.ch, angelo.garofalo@unibo.it

Abstract—Compared to the first generation of deep neural networks, dominated by regular, compute-intensive kernels such as matrix multiplications (MatMuls) and convolutions, modern decoder-based transformers interleave attention, normalization, and data-dependent control flow. This demands flexible accelerators, a requirement met by scalable, highly energy-efficient shared-L1-memory vector processing element (VPE) clusters. Meanwhile, the ever-growing size and bandwidth needs of state-of-the-art models make reduced-precision formats increasingly attractive. Microscaling (MX) data formats, based on block floating-point (BFP) representations, have emerged as a promising solution to reduce data volumes while preserving accuracy. However, MX semantics are poorly aligned with vector execution: block scaling and multi-step mixed-precision operations break the regularity of vector pipelines, leading to underutilized compute resources and performance degradation. To address these challenges, we propose VMXDOTP, a RISC-V Vector (RVV) 1.0 instruction set architecture (ISA) extension for efficient MX dot product execution, supporting MXFP8 and MXFP4 inputs, FP32 and BF16 accumulation, and software-defined block sizes. A VMXDOTP-enhanced VPE cluster achieves up to 97% utilization on MX-MatMul. Implemented in 12 nm FinFET, it achieves up to 125 MXFP8-GFLOPS and 250 MXFP4-GFLOPS, with 843/1632 MXFP8/MXFP4-GFLOPS/W at 1 GHz, 0.8 V, and only 7.2% area overhead. Our design yields up to 7.0× speedup and 4.9× energy efficiency with respect to software-emulated MXFP8-MatMul. Compared with prior MX engines, VMXDOTP supports variable block sizes, is up to 1.4× more area-efficient, and delivers up to 2.1× higher energy efficiency.

Index Terms—Microscaling, Vector processors, Efficiency

I. INTRODUCTION

The growing memory, bandwidth, and compute requirements of modern artificial intelligence (AI) workloads present significant challenges. To address these, one effective approach is the use of narrow bit-width data formats, which significantly reduce storage and data movement costs while enabling more energy-efficient computation. However, as bitwidths decrease, preserving model accuracy becomes increasingly challenging due to the reduced dynamic range and precision [1].

To alleviate this trade-off, block-scaled data formats have emerged as a compelling solution. By associating a shared scale factor with a block of low-bitwidth elements, these formats preserve high dynamic range while retaining the benefits of a compact representation. In particular, the recently proposed Microscaling (MX) formats [2] couple a block-level exponent to a vector of narrow floating-point (FP) elements. Standardized by the Open Compute Project (OCP) and supported by key

This work was supported in part by the Swiss State Secretariat for Education, Research, and Innovation (SERI) under the SwissChips initiative, and by Huawei Zurich Research Center (ZRC).

industry players, MX formats have demonstrated high accuracy across a wide range of AI workloads, often serving as a drop-in replacement for wider formats [3].

While the memory savings of MX formats are a direct consequence of their compact design, their computational benefits are often overlooked. MX quantization is frequently treated as a storage-only compression approach to alleviate memory bottlenecks, requiring decompression to wider formats before computation [4], [5]. Fully exploiting the computational efficiency of MX formats, however, requires native hardware support. Recognizing this, both NVIDIA and AMD have recently added such support in their *Blackwell* and *CDNA 4* microarchitectures, respectively [6], [7].

The success of MX formats in specialized hardware has naturally led to growing interest in supporting them on more general-purpose, programmable architectures [8]. In particular, vector processors are a promising target as they combine data parallelism, programmability, and software portability. These features have led to their adoption in mainstream ISAs, notably through Arm SVE, SVE2, and the recently ratified RISC-V Vector Extension (RVV) 1.0. Among these, the open-source RVV, explicitly designed for high efficiency on data-parallel workloads pervasive in AI, offers a compelling framework for supporting emerging standards such as MX formats.

However, the optimal path for integrating MX support into RVV is not yet clear. To enable software emulation of MX operations, narrow FP elements must be cast to wider formats for computation. To this end, a set of vector conversion instructions is in the process of being standardized for RVV [9], [10]. Although an essential first step, this approach treats MX formats purely as a storage or transport medium. As our analysis will show, this fails to address the core computational bottlenecks and can leave substantial performance and efficiency gains on the table.

This paper argues that unlocking the full computational benefits of MX formats on vector processors requires a tightly integrated hardware approach. To demonstrate this, we extend Spatz, an open-source VPE [11], with a custom RVV ISA extension that enables direct MX dot product computation without prior decompression, and make the following contributions:

- We implement and analyze RVV kernels for software-emulated MX-MatMul, identifying fundamental performance limitations that cannot be addressed using only the existing ISA and FP conversion instructions.
- We propose VMXDOTP, a novel RVV ISA extension that

provides native, single-instruction support for MXFP8 and MXFP4 dot products with accumulation in FP32 or BF16 and flexible, software-defined block sizes.

- We integrate VMXDOTP into the Spatz VPE and implement the design in a 12 nm FinFET technology, incurring an area overhead of 12.6 % at the core level, and only 7.2 % at the cluster level.
- We demonstrate up to $7.0\times$ speedup and $4.9\times$ higher energy efficiency for MX-MatMul compared to software emulation on the original Spatz processor.

II. BACKGROUND

A. Microscaling (MX) Formats

The OCP Microscaling (MX) specification [2] defines a class of BFP formats. Each MX block contains k elements sharing a single 8-bit exponent scale (E8M0), increasing the dynamic range between blocks despite the compact representation. The specification defines several concrete data formats, all with a block size of $k = 32$. There are five formats with FP elements (MXFP8_{E5M2/E4M3}, MXFP6_{E3M2/E2M3}, and MXFP4_{E2M1}) as well as the MXINT8 format with 8-bit signed integer elements.

The fundamental operation on MX data is the dot product (MX-DP) between two MX blocks, A and B , defined as:

$$C = \text{Dot}(A, B) = X(A) \cdot X(B) \cdot \sum_{i=1}^k P_i(A) \cdot P_i(B), \quad (1)$$

where $X(A)$, $X(B)$ are the block scales, $P_i(A)$, $P_i(B)$ the individual elements, and the result C should be in FP32 format.

This work focuses on MXFP8 and MXFP4 formats. We omit MXFP6 as its 6-bit elements are ill-suited to byte-oriented general-purpose processors, and exclude MXINT8, as it can be efficiently emulated using integer arithmetic [12].

B. RISC-V Vector Extension (RVV)

The standardized RISC-V Vector Extension (RVV) adds a data-parallel programming model to the ISA. Instructions are configured via the `vtype` Control and Status Register (CSR), which includes Selected Element Width (SEW) to define operand widths. RVV supports widening and narrowing operations, where the result element width is twice or half the operand element width, respectively. For these operations, the wider data type uses an Effective Element Width (EEW) of $2 \times \text{SEW}$. To increase register utilization, RVV uses Length Multipliers (LMULs) to combine multiple VLEN-bit registers into longer vectors. To match the number of elements in all operands in mixed-width operations, an Effective LMUL (EMUL) of $2 \times \text{LMUL}$ is used for the wider operands.

C. Spatz

Spatz [11] is an open-source RVV processor designed for energy efficiency and embedded applications. It is coupled to a tiny 32-bit scalar integer core, forming a Spatz core complex (CC). The VPE consists of a centralized vector register file (VRF), a controller, and three parallel functional units: the vector arithmetic unit (VAU), the vector load-store unit (VLSU), and the vector slide unit (VSLDU). The VRF hosts the 512-bit wide vector registers distributed across 4 banks, which provide three

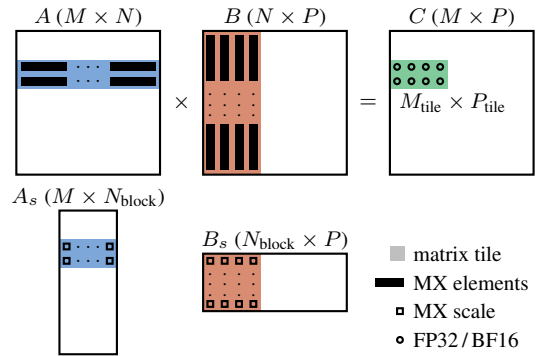


Fig. 1. Overview of MX-MatMul with $M_{\text{tile}} \times P_{\text{tile}}$ output tiles. The separate scale matrices (A_s , B_s) have reduced inner dimensions of $N_{\text{block}} = N/k$.

read ports and one write port (3R1W) each. The VAU handles most computational instructions through its integer processing unit (IPU) and four floating-point units (FPUs). Spatz also handles scalar FP operations, for which the controller hosts a separate register file (RF) and a load-store unit (LSU).

Two Spatz CCs form a cluster, sharing a 128 KiB L1 scratchpad memory. In total, the cluster can sustain 128 bits of integer or 512 bits of FP operations per cycle.

Spatz supports most of the standard RVV extension and two custom ISA extensions [13]: *MiniFloat-NN* adds full support for low-precision 16-bit (FP16, BF16) and 8-bit (FP8_{E5M2/E4M3}) FP formats. *ExSdotp* adds the `vwfwdotp` instruction to compute an “expanding sum of dot products.” It multiplies two FP operand pairs and accumulates the results into a double-width destination register, effectively doubling the throughput compared to regular widening fused multiply-add (FMA) instructions (`vwfmacc`).

III. SOFTWARE EMULATION

To show the limitations of emulating MX operations fully in software, we implement RVV kernels for MXFP8-MatMul. As illustrated in Fig. 1, data is quantized into MX blocks along the reduction axis (i.e. along rows for A , columns for B). The E8M0 block scales are stored separately (in A_s , B_s) from the FP elements (in A , B). For accumulation, we consider both the specified FP32 format and the more compact BF16 format.

A. FP8 Conversion Support

The kernels require vector and scalar instructions to expand FP8 operands to 16 bits for further processing. As this is not supported in standard RVV, we use the *MiniFloat-NN* extension [13] implemented by the baseline Spatz VPE.

The proposed *Zvfofp8min* standard extension [9] provides FP8-to-BF16 vector conversion instructions as an alternative, but lacks a scalar counterpart.

B. Baseline: MXFP8-MatMul Kernel

Our implementation uses an outer-product algorithm, which vectorizes computation along the output matrix’s row dimension, avoiding often inefficient reduction instructions (`vwfredusum.vs`). We assume that all input matrices are stored in row-major order, which avoids inefficient strided 8-bit loads on B and B_s . The pseudocode in Listing 1 illustrates the computation for a $1 \times P_{\text{tile}}$ output tile with FP32 accumulation.

The kernel iterates through each MX block along the reduction

Listing 1. Baseline RVV kernel for MXFP8-MatMul ($1 \times P_{\text{tile}}$ output tile) with FP32 accumulation.

```

size_t N, N_block = N / BLOCK_SIZE;
size_t P_tile = get_vlmax(SEW_32, LMUL_4);
fp8_t A[1][N];      e8m0_t As[1][N_block];
fp8_t B[N][P_tile]; e8m0_t Bs[N_block][P_tile];
float C[1][P_tile];
vsetvli(P_tile, SEW_32, LMUL_M4); v0..3 = vmv.v.i(0);
for (size_t block = 0; block < N_block; block++) {
    v4..7 = vmv.v.i(0);
    for (size_t elem = 0; elem < BLOCK_SIZE; elem++) {
        size_t idx = block * BLOCK_SIZE + elem;
        fp16_t a0 = fcvh.h.b(A[0][idx]);
        vsetvli(P_tile, SEW_8, LMUL_M1);
        v8 = vle8.v(B[idx][:]);
        v8..v9 = vfwcvt.f.f.v(v8);
        vsetvli(P_tile, SEW_16, LMUL_M2);
        v4..v7 = vfwmac.vf(v4..v7, v8..v9, a0);
    }
    int as0 = As[0][block] - 127; // remove bias
    v12 = vle8.v(Bs[block][:]);
    v12..v13 = vvcvtu.x.x.v(v12);
    vsetvli(P_tile, SEW_16, LMUL_M2);
    v16..v19 = vwadd.vx(v12..v13, as0);
    v16..v19 = vsll.vi(v16..v19, 23); // to FP32
    vsetvli(P_tile, SEW_32, LMUL_M4);
    v0..3 = vfmacc.vv(v0..v3, v4..v7, v16..v19);
}
c[0][:] = vse32.v(v0..v3); // store result

```

dimension, where it performs three steps: ① An inner loop iterates through the elements, producing an unscaled block dot product. Each iteration loads an FP8 element from A and the corresponding FP8 vector from B , expands them to FP16, and combines them using a widening FMA (`vfwmac.vf`). ② The E8M0 block scales are loaded from A_s (as a scalar) and B_s (as a vector). The 8-bit exponents are combined and converted to FP32 using a sequence of integer instructions [8]. ③ The unscaled dot products and the expanded scales are combined using a vector-vector FMA into the global accumulator vector. At the end, the result is written back to C .

To improve performance, we manually unroll the inner loop to parallelize loads with arithmetic operations. We also process multiple rows in parallel ($M_{\text{tile}} = 2$), maximizing data reuse within the VRF. Finally, we implement a similar kernel accumulating in BF16, where the widening FMA is replaced with a single-width instruction (`vfmacc.vf`).

C. Analysis

We evaluate our baseline MXFP8-MatMul kernels with a 64×64 output matrix and an inner dimension of $N = 128$ on the Spatz cluster (Section II-C), comparing them to standard FP32 and BF16 MatMul. As shown in Fig. 2, the MX kernels have runtimes of 63,162 cycles (FP32 accumulation) and 43,487 cycles (BF16). Compared to them, the regular FP32 and BF16 kernels are 88% and 155% faster, respectively.

To examine the overhead of our MX kernels, we analyze the utilization of the functional unit (VAU) and break down the execution time by instruction type. In the standard FP32 kernel, 97.7% of VAU cycles are spent on “useful” FMAs. While the MXFP8-to-FP32 kernel requires the same amount of time to perform (widening) FMAs, it performs significant additional work: 19.5% of runtime is used for vector and scalar FP conversions, while an additional 16.2% is spent converting

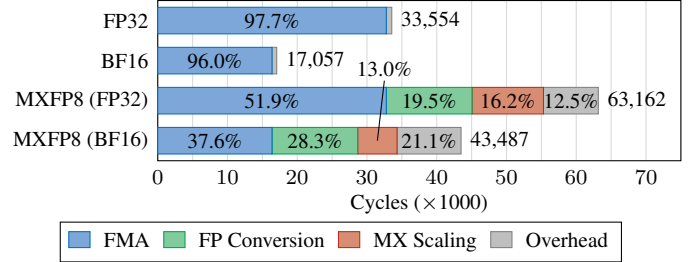


Fig. 2. VAU cycles spent executing different instruction types during MatMul kernels ($N = 128$).

and applying block scales. Furthermore, software emulation incurs significant additional overhead of around 12.5%, which has multiple reasons. First, the use of multi-step mixed-precision operations requires frequent `vtype` changes. Second, the large number of intermediate results increases register pressure. This requires the use of lower LMUL values, which in turn reduces the amount of data processed per instruction. Third, the increased loop nesting leads to more control-flow-related instructions.

The results for BF16 are similar in absolute terms. However, as BF16 FMAs have higher throughput, the FMA part of total runtime decreases to 37.6% and MX scaling approximately halves while FP conversion and overhead cycles stay approximately the same as in the FP32 case.

D. Discussion

Our analysis of MX software emulation reveals significant performance limitations inherent to the current RVV ISA. While these results were obtained on a specific implementation (i.e., Spatz), the identified bottlenecks, explicit FP conversions and software-managed scaling, are fundamental.

Consequently, the software-emulated approach fails to translate the compact representation of MX formats into a computational advantage. This introduces an undesirable trade-off: while MX formats reduce memory footprint and bandwidth, standard FP remains the more performant option. To resolve this and unlock the full potential of MX formats, native hardware support for MX operations is essential.

IV. THE VMXDOTP ISA EXTENSION

A. Design Goals

Motivated by the fundamental inefficiencies of software-emulated MX operations, we aim to design an RVV ISA extension that enables efficient MX-MatMul through a native MX-DP primitive. Our design is guided by several key goals:

- To eliminate the overhead of software scaling, the extension should **apply MX scales directly in hardware**.
- Similar to FMA, the MX-DP instructions must include a **fused accumulation** step. This avoids extra FP addition and normalization overhead.
- The extension should support **multiple formats**: This includes MXFP8_{E5M2/E4M3} and MXFP4_{E2M1} elements, with accumulation in both the specification-mandated FP32 and the more compact BF16 format.
- The new instructions should **integrate with the RVV programming model**, be vector-length agnostic, and include both vector-vector and vector-scalar variants.

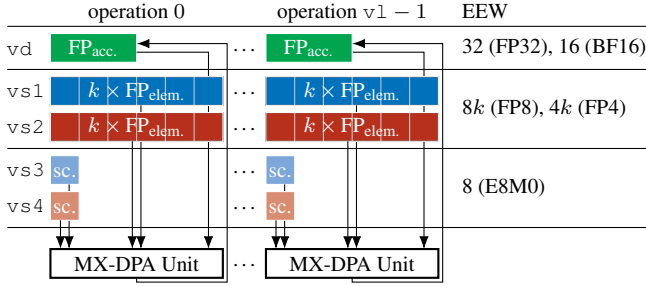


Fig. 3. Vector register layout for vector-vector VMXDOTP instructions with block size k . There are $v1$ independent MX-DPA operations.

- The design must allow **efficient microarchitectures**, achieving high computational throughput and targeting near-full FPU utilization at low complexity and cost.
- The extension should support **flexible block sizes** selected through software, and should not be architecturally constrained to the standard block size of 32.

B. Challenges

Given these design goals, we first consider a comprehensive MX Dot-Product-Accumulate (MX-DPA) instruction that computes a full 32-element dot product between two MX blocks, applies the block scales, and accumulates the result. Its vector data layout is illustrated in Fig. 3 (with $k = 32$).

There are several challenges to address to make this instruction conform to the design goals and feasible to implement:

- 1) The MX-DPA **operands vary greatly in bitwidth**, ranging in EEW from 8 (scales) to 256 bits (32 FP8 elements). This diverges significantly from standard RVV, where operand widths differ by a factor of two at most.
- 2) The MX-DPA unit's **inputs are very wide**, with an EEW of up to 256 bits for the element operands, compared with at most 32 or 64 bits for existing RVV instructions.
- 3) There are **5 source operands** for MX-DPA, while standard RVV instructions are limited to 3 (plus the special mask register). This creates challenges for both microarchitectures and instruction encoding.
- 4) This operation **fixes the block size** k , violating the requirement for flexibility in that regard.

Our solution to challenges 1 and 2 is based on a key insight: An MX-DP can be decomposed into the sum of multiple smaller dot products that reuse the same block scales. For example, a 32-wide MX-DP can be computed by summing the results of four 8-wide dot products.

This insight allows us to reduce the *hardware block size* from 32 to more manageable values. In particular, we reduce k until EEW of the MX elements equals the scalar FP register width (FLEN, either 32 or 64). This avoids introducing new EEWs not present in standard RVV. It also allows holding packed MX element operands in scalar FP registers, a requirement for vector-scalar instructions. As described previously, software can implement any MX block size which is a multiple of the hardware block size k (in particular, 32), solving challenge 4.

We do not address challenge 3 at the architectural level. Rather, as we will show later (Section V), it is possible to prefetch and buffer the narrow scale operands with minimal

TABLE I
VMXDOTP INSTRUCTION VARIANTS DEPENDING ON FLEN

FLEN	Instruction	Acc.	SEW	EMUL _{elem.}	EMUL _{sc.}	k_{FP8}	k_{FP4}
32	vmxdotp.vv/vf	FP32	32	$LMUL$	$LMUL/4$	4	8
	vmxdotp.wv/wf	BF16	16	$2 \cdot LMUL$	$LMUL/2$		
64	vmxdotp.wv/wf	FP32	32	$2 \cdot LMUL$	$LMUL/4$	8	16
	vmxdotp.qv/qf	BF16	16	$4 \cdot LMUL$	$LMUL/2$		

overhead, thereby avoiding the need for expensive additional read ports to the VRF banks.

C. Instruction Specification

The different VMXDOTP instructions for FLEN = 32 and 64 are listed in Table I. The accumulator precision is set via SEW (32 for FP32, 16 for BF16), while the element FP format (FP8_{E5M2/E4M3} or FP4_{E2M1}) is selected through a CSR. Based on the width ratio of the FLEN-bit element vectors and the FP accumulators, the instructions are classified as single-width (ratio 1, v), narrowing (2, w), or *quad-narrowing* (4, q).

For the vector-vector instructions (vv , wv , and qv suffixes), the i -th element of the accumulator vector is computed as:

$$vd[i] += vs3[i] \cdot vs4[i] \cdot \sum_{j=0}^{k-1} vs1[kj + j] \cdot vs2[kj + j], \quad (2)$$

where $vs1$ and $vs2$ are interpreted in element data format (FP8/FP4), $vs3$ and $vs4$ as E8M0 scales, vd contains the FP accumulators, and k is the hardware block size from Table I. The computation for the vector-scalar instructions (vf , wf , and qf suffixes) is similar, with the first and third operands being broadcast from scalar FP registers:

$$vd[i] += rs3 \cdot vs4[i] \cdot \sum_{j=0}^{k-1} rs1[j] \cdot vs2[kj + j], \quad (3)$$

The required 25 bits to encode the 5 register operands make it infeasible to encode the instructions within the 32-bit encoding space in a standard-compatible way. There are a number of approaches to reduce the number of bits required to encode the operands, e.g., restricting the number of addressable registers or grouping the scalar FP registers into pairs. However, all such schemes fail to achieve the required reduction in bits without placing severe restrictions on register allocation.

For future standardization, we propose using the longer 48-bit or 64-bit instruction encodings provided by RISC-V [14], which can easily accommodate 5 full register specifiers. However, to avoid the complexity of variable-length instruction decoding, prototypes and custom accelerators may recycle unused parts of the 32-bit encoding space instead. We use this second option for our implementation (Section V).

D. MX-MatMul Kernel Using VMXDOTP

We now implement accelerated RVV kernels for MX-MatMul leveraging the new VMXDOTP extension. Similar to the baseline, we use an outer-product algorithm. However, B is now stored in column-major order, such that elements of the same MX block are stored contiguously in memory. The pseudocode in Listing 2 illustrates the computation for a single output tile ($1 \times P_{\text{tile}}$) with MXFP8 inputs and accumulation in FP32.

```

Listing 2. VMXDOTP kernel for MXFP8-MatMul ( $1 \times P_{\text{tile}}$  output tile) with
FP32 accumulation. We use FLEN = 64, i.e., HW_BLOCK_SIZE =  $k_{\text{FP8}} = 8$ .
size_t N, P_tile = get_vlmax(SEW_32, LMUL_M2);
fp8_t A[1][N]; e8m0_t As[1][N_block];
fp8_t B[P_tile][N]; e8m0_t Bs[N_block][P_tile];
float C[1][P_tile]; double a0, as0;
vsetvli(P_tile, SEW_32, LMUL_M2); v0..1 = vmv.v.i(0);
for (size_t n = 0; n < N; n += HW_BLOCK_SIZE) {
  1 a0 = A[0][n:n+HW_BLOCK_SIZE]; // 8x FP8 packed
  v4..v7 = vlse64.v(B[:][n]);
  2 if (n % BLOCK_SIZE == 0) { // once per block
    as0 = As[0][n/BLOCK_SIZE]; // 1x E8M0
    v8 = vle8.v(B[n/BLOCK_SIZE][:]);
  }
  3 v0..v1 = vmxdotp.wf(v0..v1, a0, v4..v7, as0, v8);
}
c[0][:] = vse32.v(v0..v1); // store result

```

The code iterates block by block along the reduction dimension (step size k). 1 In each iteration, elements are loaded from A (packed into a scalar FP register) and B (using FLEN-bit strided loads). 2 For each block, scales are loaded from A_s and B_s as before. As the software block size ($\text{BLOCK_SIZE} = 32$) differs from the hardware block size ($\text{HW_BLOCK_SIZE} = 8$), they are reused across iterations. 3 Finally, the MX-DP is computed and accumulated using the `vmxdotp.wf` instruction.

As with the baseline, we unroll the loop and process multiple rows in parallel ($M_{\text{tile}} = 8$) to maximize performance.

To implement MXFP4-MatMul, only two modifications are required: we write the relevant CSR to select FP4 source format, and double HW_BLOCK_SIZE to $k_{\text{FP4}} = 16$.

V. HARDWARE IMPLEMENTATION

To evaluate our proposed VMXDOTP extension, we integrate it into the Spatz VPE. Based on Spatz’s FLEN of 64, we implement the narrowing (w^*) and quad-narrowing (q^*) instructions. Our modifications to Spatz are illustrated in Fig. 4.

For the datapath, we integrate the MXDOTP FPU [8], which includes an 8-wide MXFP8 dot product with FP32 accumulation. We extend the unit to support 16-wide MXFP4 dot products and BF16 accumulation. To provide the accumulator and FP elements, we reuse the existing infrastructure. The two MX scale operands need to be supplied to the FPUs separately, which we achieve by adding two read ports ($vs3, vs4$) to the VRF.

The comparatively low read bandwidth required for the scales (2×8 bits per operation) when compared to the elements (2×64 bits) prompts us to fetch a batch of scales at once, buffer them within the VAU, and consume them progressively over 8 cycles. This optimization allows us to multiplex the 5 logical read ports between VAU and VRF onto the 3 physical read ports of each VRF bank, avoiding the prohibitive area cost of additional read ports to the memory banks. In general, this introduces a cycle of overhead every 8 cycles, as the element read requests are stalled during scale prefetching. However, this overhead is avoided when the operands are mapped to different VRF banks, or in the case of vector-scalar instructions (`vmxdotp.*f`), which only use $vd, vs2$, and $vs4$.

We also adjust the operand shuffling to pack the accumulator and scale operands into a single 64-bit value as required by the FPUs, and modify the result selection to only write 32/16 bits of output per operation in the narrowing/quad-narrowing case.

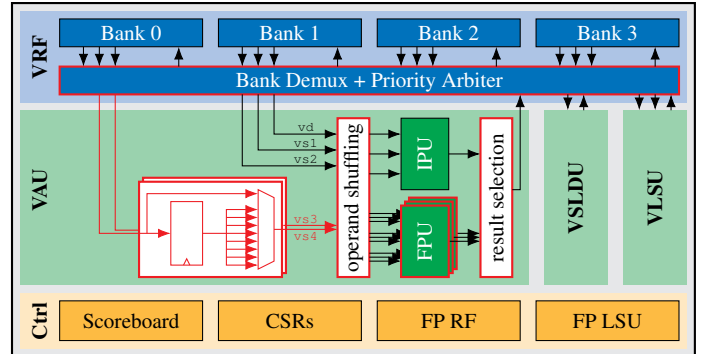


Fig. 4. Spatz VPE with datapath changes for VMXDOTP integration highlighted in red.

TABLE II
AREA IMPACT OF VMXDOTP AT DIFFERENT SPATZ HIERARCHY LEVELS

Hierarchy	Baseline (kGE)	This Work (kGE)	Change
Cluster	3995	4281	+ 7.2 %
Core Complex ($\times 2$)	2233	2515	+12.6 %
FPU ($\times 4$)	1264	1499	+18.6 %
VAU (w/o FPU/IPU)	74	97	+31.0 %
VRF	421	444	+ 5.5 %

VI. EVALUATION

A. Physical Implementation

We implement the baseline and VMXDOTP-enabled Spatz clusters using SYNOPSIS FUSION COMPILER 2022.03 in GLOBALFOUNDRIES 12 nm FinFET technology. We use a target frequency of 0.95 GHz in the worst-case corner (SS, 0.72 V, 125 °C). Our modified cluster successfully meets this target and reaches 1.27 GHz under typical conditions (TT, 0.80 V, 25 °C), matching the baseline without introducing a new critical path.

Our VMXDOTP-enabled Spatz cluster has a total area of 4.28 MGE, representing an increase of 7.2 % over the baseline (12.6 % at the CC level). A breakdown of the area overhead is presented in Table II. Most of the increase (82 %) is due to the added MX dot product unit within the FPUs, with the remaining overhead split evenly between VAU and VRF.

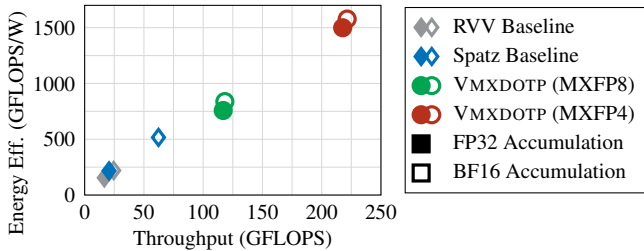
B. Software Benchmarks

We evaluate our VMXDOTP ISA extension on MX-MatMul with a 64×64 output matrix, varying inner dimensions, and FP32 or BF16 accumulation. This is compared with the kernels from Section III (*RVV baseline*) and enhanced versions using Spatz’s custom *MiniFloat-NN* and *ExSdotp* instructions (*Spatz baseline*), both executed on the unmodified Spatz cluster. All kernels read from and write to the cluster’s 128 KiB L1 scratchpad memory.

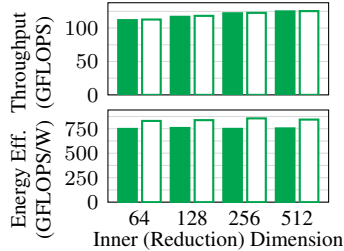
We use SYNOPSIS PRIME TIME 2022.03 for power estimation under typical conditions at 1 GHz, with switching activities extracted from post-layout simulation. We average power consumption over five different input samples, which are obtained from DeiT-Tiny [15] and quantized to MXFP8 and MXFP4 formats using Microsoft’s *Microxcaling* library [16].

C. Throughput and Energy Efficiency

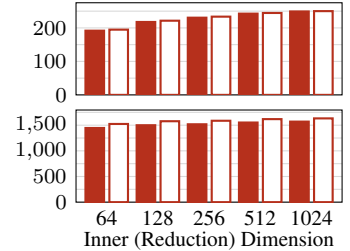
Fig. 5a compares our VMXDOTP-accelerated kernels with the RVV and Spatz baselines. Compared with RVV emulation,



(a) MXFP8 baseline and MXFP8/MXFP4 VMXDOTP kernels with inner dimension $N = 128$.



(b) MXFP8 VMXDOTP kernels.



(c) MXFP4 VMXDOTP kernels.

Fig. 5. Throughput and energy efficiency for MX-MatMul kernels with FP32 or BF16 accumulation.

the MXFP8 VMXDOTP kernels achieve a speedup of $7.0\times$ (FP32 accumulation) and $4.8\times$ (BF16) at $4.9\times$ and $3.8\times$ energy efficiency, respectively. Results are similar when compared to the FP32 Spatz baseline, while the BF16 Spatz baseline benefits heavily from full support for FP8 arithmetic. Despite this, our extension still provides a $1.9\times$ speedup at $1.6\times$ energy efficiency. As expected, the MXFP4 VMXDOTP kernels approximately double the throughput and efficiency of their MXFP8 counterparts. Compared to FP32, using BF16 accumulation with our extension increases energy efficiency by 5% to 10%, coupled with a small increase in throughput.

Figs. 5b and 5c show the performance of our VMXDOTP extension under various inner dimensions. For MXFP8, VMXDOTP achieves a throughput of up to 125.0 GFLOPS (FP32) and 125.4 GFLOPS (BF16) at an energy efficiency of 753 and 843 GFLOPS/W. These throughputs correspond to an FPU utilization of 97.6% and 97.9%, respectively. The results for MXFP4 inputs are similar, achieving a throughput of up to 249.1 GFLOPS (FP32, 97.3% utilization) and 250.1 GFLOPS (BF16, 97.7% utilization) at an energy efficiency of 1570 and 1632 GFLOPS/W, respectively.

D. Comparison with State of the Art

We compare VMXDOTP to state-of-the-art MX accelerators and a non-MX vector processor supporting FP8 arithmetic, as summarized in Table III.

VEGETA [17] and Cuyckens et al. [18] propose large-scale dataflow accelerators for MatMul using various MX formats. A direct comparison is challenging, however, as these works target fixed-function accelerators, whereas our design is a fully programmable VPE cluster. Their system-level figures are extrapolated from processing element (PE)-level synthesis results or simulator estimates, omitting the area, power, and timing overheads of system integration and physical implementation. In contrast, our cluster-level figures include the interconnect and 128 KiB shared-L1 memory, with energy efficiency results derived from back-annotated post-layout simulations. Despite this broader scope and full programmability, the energy efficiency of our design remains comparable, achieving $1.8\times$ that of Cuyckens et al. for MXFP8, and $0.94\times$ for MXFP4. Unlike VEGETA and Cuyckens et al., which both employ a fixed MX block size for quantization, our design supports software-defined block sizes. This flexibility is crucial given the rapidly evolving landscape of AI model quantization and recent work suggesting the use of smaller block sizes for optimal results [19].

TABLE III
COMPARISON OF VMXDOTP WITH STATE OF THE ART

Design	Tech. nm	Volt. V	Freq. GHz	Area mm ²	Input Format	Accum. Format	Area Eff. GFLOPS/mm ²	Energy Eff. GFLOPS/W
VEGETA ^{††} [17]	65	-	0.18	1.01	MXFP8 _{E5M2}	BF16	183	6460
				1.32	MXFP8 _{E4M3}		140	5680
				0.85	MXFP6 _{E3M2}		216	7912
Cuyckens et al. ^{*,‡} [18]	16	-	0.40	8.92	MXFP8 MXFP4	FP32	1469 2939	388-420 1667
MXDOTP [8]	12	0.8	1.00	0.59	MXFP8	FP32	173	356
MiniFloat-NN Spatz [13]	12	0.8	1.08	0.44	FP8	FP16	307	860
This Work	12	0.8	1.00	0.52	MXFP8 MXFP4	FP32/BF16	240/240 479/481	753/843 1570/1632

^{*}PE level. [†]System-level simulator estimates. [‡]Post-synthesis estimates.

Turning to programmable, core-based alternatives with instruction extensions, MXDOTP [8] proposes a scalar RISC-V instruction semantically similar to VMXDOTP. However, its reliance on Stream Semantic Registers (SSRs) to supply operands represents a significant architectural departure from standard RISC-V. In contrast, we resolve read port contention microarchitecturally through time-multiplexed RF accesses. Our design is $1.4\times$ more area-efficient and delivers $2.1\times$ higher energy efficiency for MXFP8 compared to MXDOTP despite more comprehensive format support. These results highlight the advantages of vector architectures over scalar processors.

VMXDOTP extends the MiniFloat-NN Spatz [13] baseline with MX dot product instructions, trading a small reduction in area and energy efficiency for the superior numerical robustness of MX formats compared to scalar minifloats. The added logic for scale manipulation and multi-operand accumulation accounts for our lower area efficiency and slight decrease (2% to 12%) in energy efficiency.

VII. CONCLUSION

We presented VMXDOTP, a RISC-V Vector ISA extension for efficient MXFP8 and MXFP4 dot products, with support for FP32 and BF16 accumulator precisions and software-defined block sizes. Integrated into Spatz and implemented in a 12 nm technology, VMXDOTP achieves up to 125 MXFP8-GFLOPS at up to 843 MXFP8-GFLOPS/W, and up to 250 MXFP4-GFLOPS at up to 1632 MXFP4-GFLOPS/W. Compared to software emulation, this represents a speedup of $7.0\times$ and $4.8\times$ for FP32 and BF16 accumulation, respectively, while improving energy efficiency by $4.9\times$ and $3.8\times$. These results highlight the need for dedicated block-scaled dot-product-accumulate instructions in RVV.

REFERENCES

- [1] B. D. Rouhani, R. Zhao, V. Elango, *et al.*, “With Shared Microexponents, A Little Shifting Goes a Long Way,” in *50th Annual International Symposium on Computer Architecture (ISCA '23)*, Jun. 2023.
- [2] B. D. Rouhani, N. Garegrat, T. Savell, *et al.*, *OCP Microscaling Formats (MX) Specification*, version 1.0, Sep. 2023.
- [3] B. D. Rouhani, R. Zhao, A. More, *et al.* “Microscaling Data Formats for Deep Learning.” arXiv: 2310.10537. (Oct. 19, 2023).
- [4] G. Gerogiannis, S. Eyerman, E. Georganas, W. Heirman, and J. Torrellas, “DECA: A Near-Core LLM Decompression Accelerator Grounded on a 3D Roofline Model,” in *58th IEEE/ACM International Symposium on Microarchitecture (MICRO' 25)*, Oct. 2025.
- [5] C. Verrilli. “Qualcomm Cloud AI 100 Accelerates Large Language Model Inference by ~2x Using Microscaling (Mx) Formats.” Qualcomm Technologies. (Jan. 9, 2024), [Online]. Available: <https://www.qualcomm.com/developer/blog/2024/01/qualcomm-cloud-ai-100-accelerates-large-language-model-inference-2x-using-microscaling-mx> (visited on 07/28/2025).
- [6] NVIDIA. “NVIDIA Blackwell Architecture.” (2025), [Online]. Available: <https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/> (visited on 05/18/2025).
- [7] Advanced Micro Devices, “Introducing AMD CDNA 4 Architecture,” Jun. 2025. [Online]. Available: <https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/white-papers/amd-cdna-4-architecture-whitepaper.pdf> (visited on 07/25/2025).
- [8] G. İslamoğlu, L. Bertaccini, A. S. Prasad, F. Conti, A. Garofalo, and L. Benini, “MXDOTP: A RISC-V ISA Extension for Enabling Microscaling (MX) Floating-Point Dot Products,” in *36th IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP '25)*, Jul. 2025.
- [9] A. Waterman. “OFP8 conversion extension Zvfopf8min, Version 0.2.1.” (May 27, 2025), [Online]. Available: <https://github.com/aswaterman/riscv-misc/blob/e1e20a75c9a9fa797519fcc1ee997c7a7be4503/isa/zvfopf8min.adoc> (visited on 05/30/2025).
- [10] A. Waterman. “OFP4 conversion extension Zvfopf4min, Version 0.1.” (May 27, 2025), [Online]. Available: <https://github.com/aswaterman/riscv-misc/blob/e1e20a75c9a9fa797519fcc1ee997c7a7be4503/isa/zvfopf4min.adoc> (visited on 05/30/2025).
- [11] M. Perotti, S. Riedel, M. Cavalcante, and L. Benini, “Spatz: Clustering Compact RISC-V-Based Vector Units to Maximize Computing Efficiency,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 44, no. 7, pp. 2488–2502, Jul. 2025.
- [12] N. Satya Murthy, F. Catthoor, and M. Verhelst, “Optimization of block-scaled integer GeMMs for efficient DNN deployment on scalable in-order vector processors,” *Journal of Systems Architecture*, vol. 154, p. 103 236, Sep. 2024.
- [13] L. Bertaccini, G. Paulin, M. Cavalcante, T. Fischer, S. Mach, and L. Benini, “MiniFloats on RISC-V Cores: ISA Extensions With Mixed-Precision Short Dot Products,” *IEEE Transactions on Emerging Topics in Computing*, vol. 12, no. 4, pp. 1040–1055, Oct. 2024.
- [14] RISC-V International, *The RISC-V Instruction Set Manual, Volume I: Unprivileged Architecture*, version 20250508, May 2025.
- [15] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, “Training data-efficient image transformers & distillation through attention,” in *38th International Conference on Machine Learning (ICML '21)*, Jul. 2021, pp. 10 347–10 357.
- [16] Microsoft, *MX Pytorch Emulation Library*, version 1.1.0, Aug. 2024. [Online]. Available: <https://github.com/microsoft/microscaling>.
- [17] K. B. Nine, C. Talley, A. S. Mandadi, T. Krishna, and A. Raychowdhury, “Optimizing Sparse/Dense VEGETA Accelerator Performance with Microscaling Quantization,” in *2025 IEEE International Symposium on Circuits and Systems (ISCAS '25)*, May 2025.
- [18] S. Cuyckens, X. Yi, N. Satya Murthy, C. Fang, and M. Verhelst, “Efficient Precision-Scalable Hardware for Microscaling (MX) Processing in Robotics Learning,” in *2025 IEEE/ACM International Symposium on Low Power Electronics and Design (ISPLED '25)*, Aug. 2025.
- [19] B. Chmiel, M. Fishman, R. Banner, and D. Soudry, “FP4 All the Way: Fully Quantized Training of LLMs,” in *39th Conference on Neural Information Processing Systems (NeurIPS '25)*, Dec. 2025.