RoadBlock: Rethinking GPU Tensor Core Microarchitecture for Emerging Microscaling Format Support

Nikhil Rout
Computer Science
University of California, Los Angeles
Los Angeles, California, USA
nikhilrout97@gmail.com

Abstract

While multiple recent custom Microscaling (MX) format proposals preserve higher model quality over OCP MX, their hardware evaluations remain confined to standalone MAC-unit prototypes, ignoring the system-level integration overheads inside real GPU Tensor Cores. To this end, I present RoadBlock, an open-source framework for accurate end-to-end modeling of custom MX formats in a programmable Tensor Core, built on the RISC-V Vortex GPGPU. RoadBlock exposes data size, scale-factor size, and block size as configurable parameters and automatically derives per-thread metadata tiling for arbitrary formats, enabling format-specific Perplexity-PPA tradeoff evaluation before silicon commitment.

CCS Concepts

• **Computer systems organization** → **Parallel architectures**.

Keywords

GPU Microarchitecture, Tensor Core, Quantization, Microscaling

ACM Reference Format:
Nikhil Rout. 2026. RoadBlock: Rethinking GPU Tensor Core Microarchitecture for Emerging Microscaling Format Support. In *International Conference on Parallel Architectures and Compilation Techniques (PACT '26), October 19–22, 2026, Chicago, IL, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3838684.3847309

1 Introduction

Emerging Microscaling (MX) formats such as 4over6 [2], IF4 [3], and RaZeR [1] are typically evaluated as standalone MAC-unit prototypes intended for systolic-array accelerators, silently ignoring the system-level integration overheads in NVIDIA Blackwell and AMD CDNA4-like GPU Tensor Cores that production ML systems stacks actually use. To bridge this gap, I propose ***RoadBlock***[1], an open-source framework for end-to-end custom MX format modeling and Perplexity-PPA trade-off analysis, built on the programmable RISC-V Vortex GPGPU [5]. It exposes data size, scale-factor size, and block size as configurable parameters for MX format design-space exploration, and automatically derives per-thread scale-factor metadata tiling and register requirements for arbitrary formats.

[1] https://github.com/vortexgpgpu/vortex/tree/roadblock

PACT '26, Chicago, IL, USA

ACM ISBN 979-8-4007-2915-7/2026/10
https://doi.org/10.1145/3838684.3847309

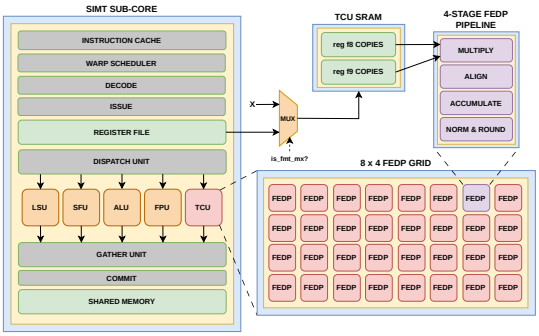


Figure 1: RoadBlock-extended Vortex GPGPU SIMT Sub-Core

2 RoadBlock Microarchitecture

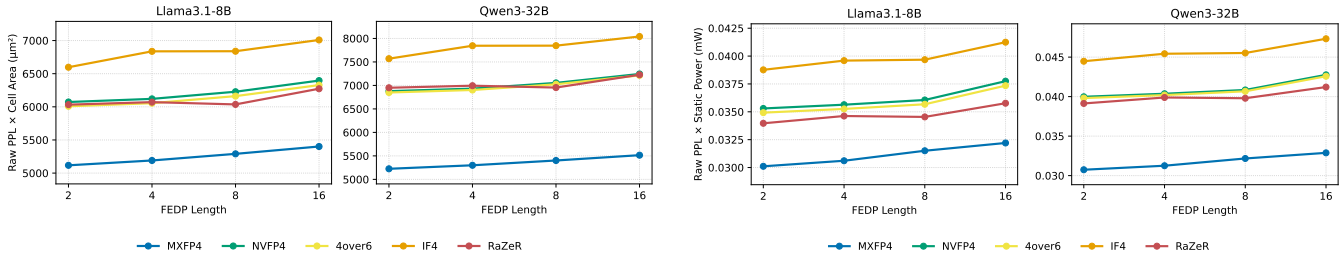
MX-enhanced Fused Dot Product (FEDP). The dot product for two MX-compliant vectors A and B of length k is given by:

$$\mathrm{Dot}(A,B) = X^{(A)}X^{(B)} \sum_{i=1}^{k} \left(P_i^{(A)} \times P_i^{(B)} \right) \tag{1}$$

where $X^{(A)}, X^{(B)}$ are block scale factors and $P_i^{(A)}, P_i^{(B)}$ are their corresponding elements. RoadBlock adopts and enhances the 4-cycle Ten-Four [4] mixed-precision FEDP microarchitecture as illustrated in Fig. 1. Since Ten-Four performs early accumulation of the "C" addend alongside the vector products, the required scale-factor computations and shifts need to be applied to each lane's significand multiplication result in the first `MULTIPLY` stage itself.

MX Scale Factor Metadata Tiling. Vortex's RISC-V-extended SIMT programming model provides a 32 registers/thread budget that is divided between matrices A, B, and C/D (8 each). Given that contiguous low-precision elements are packed along `K` into 32-bit bundles, Fig. 3 derives the required per-thread MX metadata for an arbitrarily designed format. Assuming NVFP4 (`d=4,w=8,k=16`) as the finest granularity, these formulae indicate a maximum requirement of 64 bits/thread (two extra registers) for MX metadata.

MX Metadata Delivery and Storage. Since Vortex TCU WMMA instructions only expose three source operands, MX metadata cannot be supplied as an additional explicit source. RoadBlock instead duplicates registers `f8,f9` into a lightweight TCU-SRAM through a MUX activated when the WMMA's format field indicates an MX format. The scoreboard treats these as "hidden source registers," to preserve dependency tracking. Additional metadata required by unorthodox format proposals can also be stored in this TCU-SRAM.



(a) Perplexity × Cell Area (↓) **(b) Perplexity × Static Power (↓)**

Figure 2: WikiText-2 Perplexity – Hardware Efficiency Tradeoff in Emerging 4-bit MX Formats Modeled on RoadBlock

Given

$$\text{No. Regs/Operand } (NR) = 8,\ \text{No. Threads/Warp } (NT) \in \{4, 8, 16, 32\}$$
$$d = \text{element bits},\ w = \text{scale bits},\ k = \text{block size}$$

Tiling (per-warp work chunk)

$$tile\_cap = NT \times NR = 8\,NT \quad \ell_g = \log_2(tile\_cap) = 3 + \log_2(NT) \tag{2}$$

$$tileM = 2^{\ell_g - \lfloor \ell_g/2 \rfloor} = \sqrt{8\,NT} \quad tileK = \frac{tile\_cap}{\max(tileM, tileN)} = \sqrt{8\,NT} \tag{3}$$

MX block scaling

$$tileK_{fmt} = tileK \times \frac{32}{d} = \frac{32\sqrt{8\,NT}}{d} \quad num\_blocks_k = \frac{tileK_{fmt}}{k} = \frac{32\sqrt{8\,NT}}{dk} \tag{4}$$

$$scale\_bits_k = num\_blocks_k \times w = \frac{32\,w\sqrt{8\,NT}}{dk} \tag{5}$$

MX Metadata

$$A\_metadata = tileM \times scale\_bits_k = \sqrt{8\,NT} \times \frac{32\,w\sqrt{8\,NT}}{dk} = \frac{256\,wNT}{dk} \tag{6}$$

$$Since,\ tileN = \left\{tileM, \frac{tileM}{2}\right\}, \quad B\_metadata \in \left\{\frac{256\,wNT}{dk}, \frac{128\,wNT}{dk}\right\} \tag{7}$$

$$total\_metadata = A\_metadata + B\_metadata \in \left\{\frac{512\,wNT}{dk}, \frac{384\,wNT}{dk}\right\} \tag{8}$$

$$per\_thread\_metadata = \frac{total\_metadata}{NT} \in \left\{\frac{512\,w}{dk}, \frac{384\,w}{dk}\right\}\ bits \tag{9}$$

Figure 3: Required Per-Thread MX Metadata Derivation

3 RoadBlock Host-Kernel ABI

RoadBlock extends the Vortex WMMA ABI with pointers to the A/B scale-factor buffers produced during host-side quantization, as shown in Fig. 4. For two-level formats such as NVFP4, block-wise E4M3 scaling is handled within the TCU, while tensor-wise FP32 dequantization is executed as WMMA epilogue on the FPU lanes.

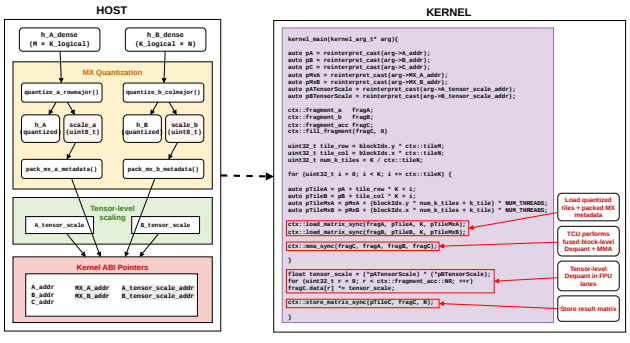


Figure 4: RoadBlock Host-Kernel ABI Flow

4 Experiments & Evaluation

To evaluate the framework's effectiveness, I modeled MXFP8, MXFP4, NVFP4, 4over6, IF4, and RaZeR on RoadBlock, and synthesized format-specific FEDP units using ASAP7 LVT/TT in Yosys under a 1.5 GHz timing constraint. Fig. 2 characterizes Llama3.1-8B and Qwen3-32B perplexity against the formats' hardware efficiency across standard GPU Tensor Core scaling points ($N \in \{2, 4, 8, 16\}$).

I find that, while MXFP4 has the worst perplexity among the group, its simple single-level E8M0 scale-factor and 32-element block size still make it the most efficient datatype when accounting for both accuracy and hardware cost together. On the other hand, while IF4 preserves model quality quite well, its extra range-alignment correction multiplier makes it an impractical addition to existing Tensor Core FEDP datapaths. As a middle-ground, RaZeR delivers the strongest Perplexity-PPA tradeoff among its NVFP4 counterparts, becoming especially attractive at larger FEDP lengths, exactly where modern GPU Tensor Core designs are headed.

5 Future Work & Conclusion

Next, I plan on prototyping more unconventional emerging MX formats on RoadBlock and investigating warp-specialized kernels for multi-level block-scaled formats. In conclusion, RoadBlock provides a flexible, open-source, hardware-grounded means to evaluate whether a custom format proposal's algorithmic advantage survives integration into a realistic, programmable GPU Tensor Core setting.

References

[1] Yuzong Chen, Xilai Dai, Jake Hyun, Chi-Chih Chang, Wonsuk Jang, Yuheng Wu, Thierry Tambe, Jae sun Seo, and Mohamed S. Abdelfattah. 2026. RaZeR: Pushing the Limits of NVFP4 Quantization with Redundant Zero Remapping. arXiv:2501.04052 [cs.LG] https://arxiv.org/abs/2501.04052

[2] Jack Cook, Junxian Guo, Guangxuan Xiao, Yujun Lin, Keith Wyss, Mahdi Nazemi, Asit Mishra, Carlo del Mundo, Tijmen Blankevoort, and Song Han. 2026. Four Over Six: More Accurate NVFP4 Quantization with Adaptive Block Scaling. arXiv:2512.02010 [cs.CL] https://arxiv.org/abs/2512.02010

[3] Jack Cook, Hyemin S. Lee, Kathryn Le, Junxian Guo, Giovanni Traverso, Anantha P. Chandrakasan, and Song Han. 2026. Adaptive Block-Scaled Data Types. arXiv:2603.28765 [cs.CL] https://arxiv.org/abs/2603.28765

[4] Nikhil Rout and Blaise Tine. 2026. Ten-Four: An Open-Source Fused Dot Product Unit for Mixed-Precision GPGPU Tensor Cores. arXiv:2512.00053 [cs.AR] https://arxiv.org/abs/2512.00053

[5] Blaise Tine, Krishna Praveen Yalamarthy, Fares Elsabbagh, and Kim Hyesoon. 2021. Vortex: Extending the RISC-V ISA for GPGPU and 3D-Graphics. In *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture* (Virtual Event, Greece) *(MICRO '21)*. Association for Computing Machinery, New York, NY, USA, 754–766. doi:10.1145/3466752.3480128