Accelerating Quantized LLM Inference for Embedded RISC-V CPUs with Vector Extension (RVV)

Andes Technology

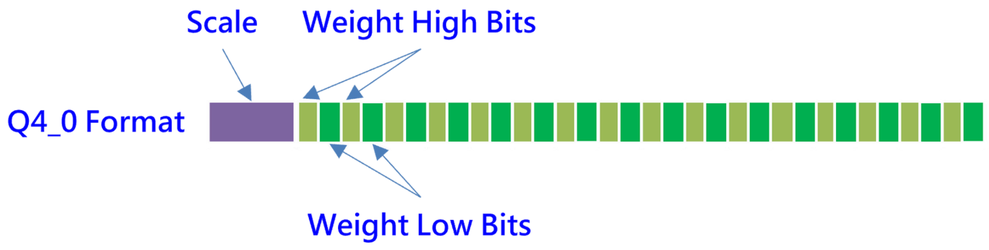
Frank Yueh-Feng Lee, Yi-Jui Chu, Chih-Chung Huang, Heng-Kuan Lee

➢llama.cpp

- Open source project for LLM inference
- Plain C/C++ implementation
- x86 / ARM / RISC-V / GPU …
- Embedded device friendly
- State-of-the-art GGUF models on Hugging Face

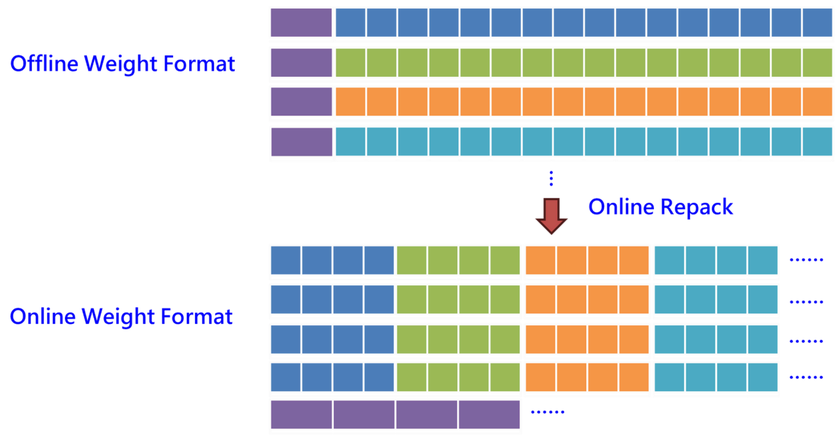
➢llama.cpp quantized GEMM / GEMV computation

- Block-wise quantization
- 2-bit to 8-bit formats
- GGML: micro-kernel design for various quantization formats
- Support RISC-V scalar and basic form of RVV optimization



➢RVV Acceleration

- RVV Q4 online repack with Andes VDOT instructions
- RVV FP32 / FP16 acceleration
- Nonlinear functions acceleration (Andes libnn)
  - Softmax, SiLU, GeLU
- Vector functions acceleration(Andes libvec) – Mul, Add



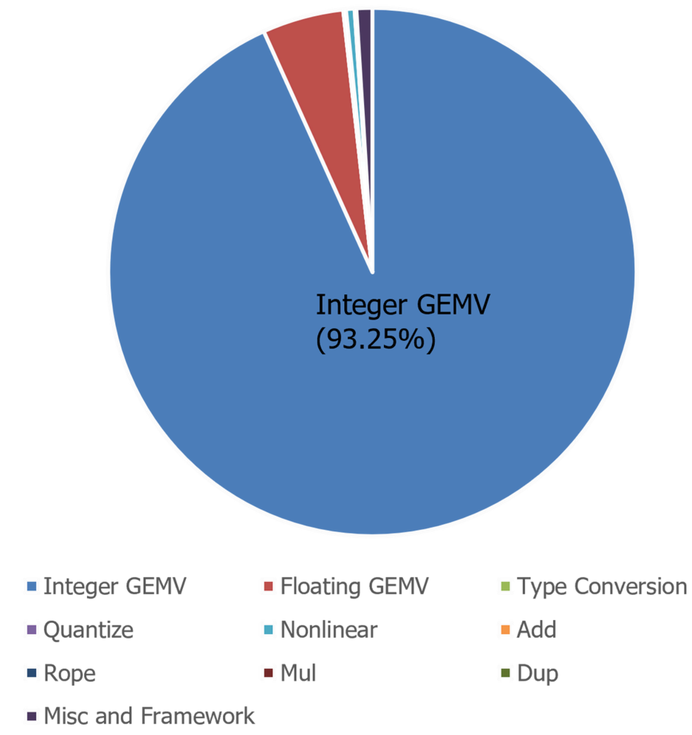
➢Nonlinear Functions Usage

Model	Nonlinear Functions
TinyLlama 1.1B	Softmax / SiLU
Llama 3.2 1B	Softmax / SiLU
Gemma 3 1B	Softmax / GeLU
DeepSeek R1 distill Qwen 1.5B	Softmax / SiLU
DeepSeek v2 Lite Chat 16B	Softmax / SiLU

➢Perplexity

Model	Platform	SW Option	PPL (wiki.test First 32 Chunk)
TinyLlama 1.1B Q4	x86	Scalar	8.0166 +/- 0.22516
	RISC-V	Scalar	8.0199 +/- 0.22519
		Andes RVV with libnn and libvec optimization	8.0225 +/- 0.22534

➢TinyLlama 1.1B Q4 Scalar Token Generation



➢TinyLlama 1.1B Q4 RVV Op Breakdown

Operator	Percentage
Integer GEMV	84.26%
Floating GEMV	5.19%
Type Conversion	1.58%
Quantize	0.74%
Softmax	0.82%
Add	0.47%
Rope	0.33%
Mul	0.21%
Dup	0.18%
SiLU	0.16%
Misc and Framework	5.04%
Total	100%

➢RVV Acceleration Results

Model	ISA	TG128 Token/Sec Scaleup @1GHz	Speedup
TinyLLaMA 1.1B Q4	RISC-V Scalar	0.2625	1x
	RISC-V Vector	6.1550	23.45x
LLaMA 2 7B Q4	RISC-V Scalar	0.0400	1x
	RISC-V Vector	1.0625	26.56x
LLaMA 3 1B Q4	RISC-V Scalar	0.1950	1x
	RISC-V Vector	5.4250	27.82x
DeepSeek R1 Distill Qwen 1.5B	RISC-V Scalar	0.1550	1x
	RISC-V Vector	4.3300	27.94x
DeepSeek v2 Lite Chat Q4	RISC-V Scalar	0.0975	1x
	RISC-V Vector	2.5200	25.85x
Gemma 3 1B Q4	RISC-V Scalar	0.2375	1x
	RISC-V Vector	6.4750	27.26x
Gemma 3 4B Q4	RISC-V Scalar	0.0625	1x
	RISC-V Vector	1.7800	28.48x

AX45MPV Single Core FPGA
VLEN 512, DLEN 512, 8MB L2, DDR 40T
TG128: llama-bench generate 128 tokens