Hardware support in RISC-V for ternary LLMs

Sergio Martínez-Arribas, David Aledo, Pablo Prieto, Pablo Abad and Valentín Puente

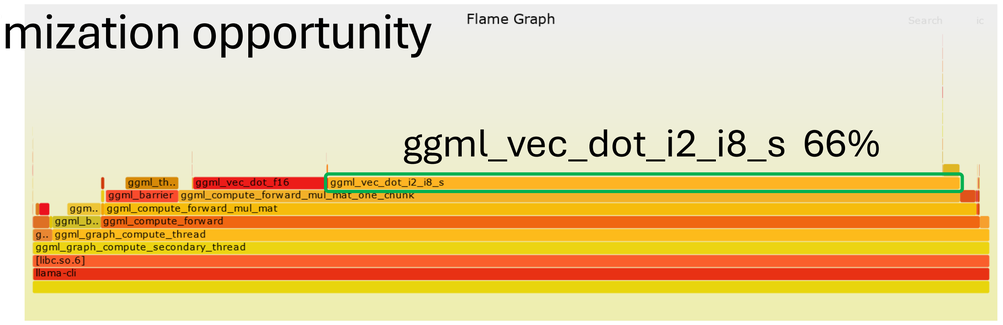
LLM on Edge

- Constantly bigger models.
- Memory bounded.
- Memory capacity and bandwidth to the limit.
- Data privacy requires local compute.

Memory footprint reduction

- Quantization → reduce number of bits per parameter.
- Small Language Models (SLM) → reduce number of parameters.

Optimization opportunity



BitNet

- Defines 3 kernels for ternary weights {-1, 0, 1} on top of Llama.cpp:
  - Kernel i2_s:
    - Quantizes weights (32 or 16 bits FP) to ternary into 2bits.
    - Lack of ternary or 2bit arithmetic support:
      - Need for dequantization + full-precision arithmetic.

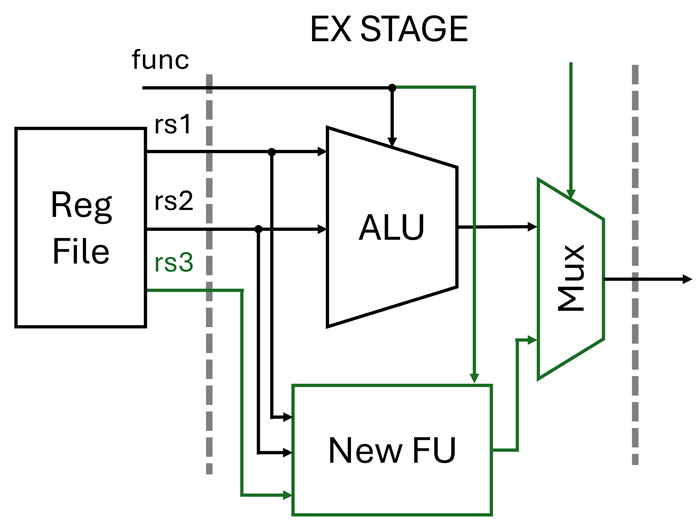
Our solution

- SIMD Ternary FU (Functional Unit).
- Masked summatory.
- 2x registers for 8*8bit activations.
- 1x register for 32*2bit weights → 1 control bit for High or Low half.

Custom-0 R-Type 4 instruction

RS3	F2	RS2	RS1	F3	RD	OPCODE
31-27	x x	24-20	19-15	0 0 F	11-7	0 0 0 1 0 1 1

- F: control bit for High (1) or Low (0) half of the weights register.



Note: rs3 adds extra logic to the datapath and the hazard detection unit.

Easy to use

- On C/C++: Macro hiding the use of ***intrinsics***

```
void ggml_vec_dot_i2_i8_s(int n, int* s, const int8_t* x, const int8_t* y) {
  int accu = 0;
  for (int i = 0; i < n; i+=8) {
    QUANT_MAD(accu, &x[i], &y[i*4], &y[i*4+1], &y[i*4+2], &y[i*4+3]);
  }
  *s = accu;
}
```

```
#define QUANT_MAD(sum, weights, act0, act1, act2, act3) \
  int64_t rd = 0; \
  asm volatile(".insn r4 0xb, 0, 0, %0, %1, %2, %3\n" \
    : "=r"(rd) \
    : "r"(weights), "r"(act0), "r"(act1) \
    : "memory"); \
  sum += rd; \
  asm volatile(".insn r4 0xb, 1, 1, %0, %1, %2, %3\n" \
    : "=r"(rd) \
    : "r"(weights), "r"(act2), "r"(act3) \
    : "memory"); \
  sum += rd;
```

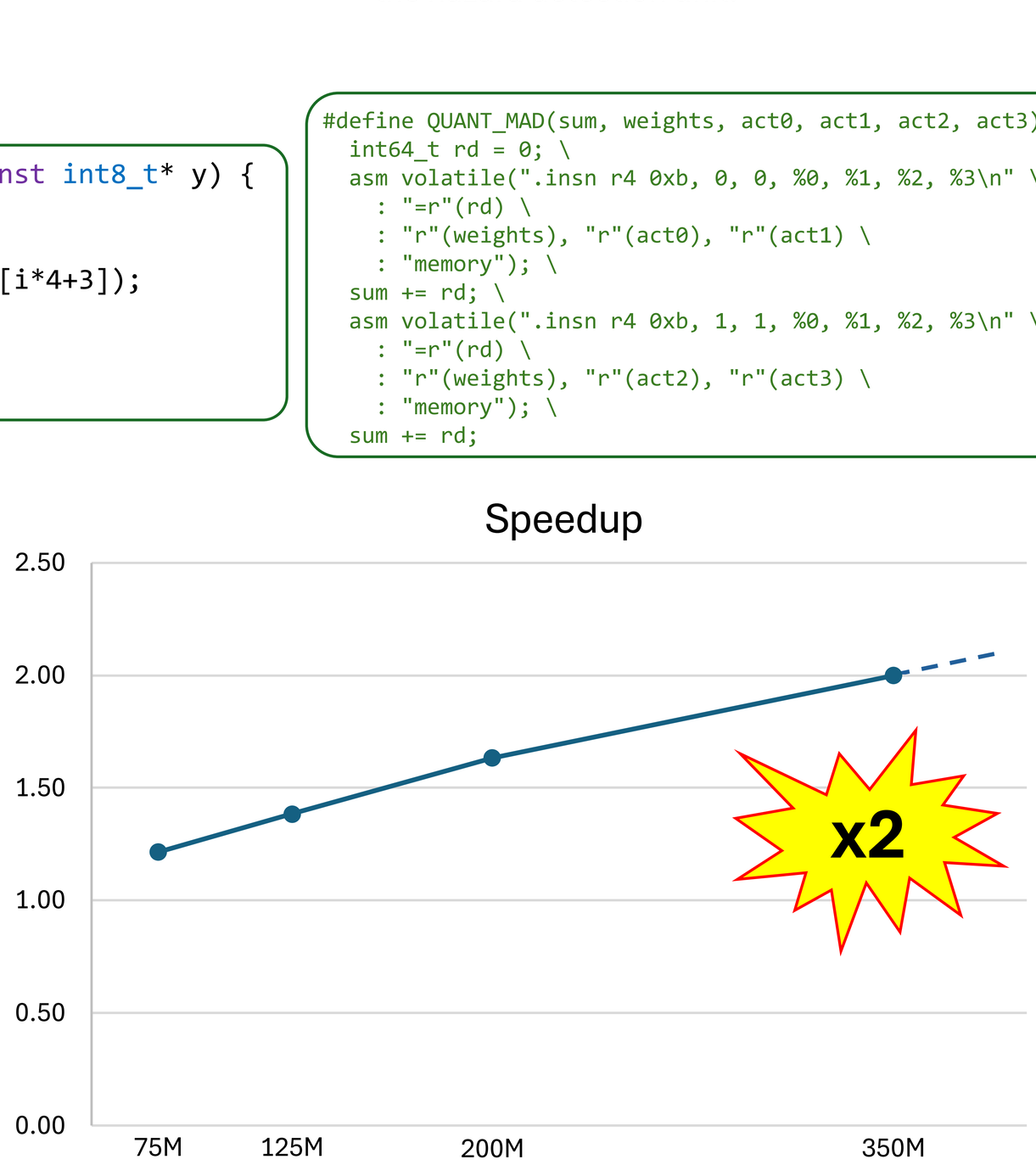
Resource utilization on FPGA

- Rocket 64bits on Nexys Video development board.

resource	without	with FU	increment
LUT	50188	50813	1.25%
FF	31608	31676	0.22%
BRAM	27	27	0%
DSP	15	15	0%

- New FU alone consumes only 287 LUTs.

Speedup



- Tokens/s with FU / tokens/s without it.

Funded by

Project PID2022-139664NB-I00

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Cofinanciado por la Unión Europea

AGENCIA ESTATAL DE INVESTIGACIÓN