

MMM macro PR — speed-gain attribution

Paired-build measurement of the gains from this PR's substitutions alone. Both binaries are built from the same source tree (this PR's branch); the only difference is the `+relaxed-simd` target-feature flag, which flips the macro's emission between `mul+add` and `f32x4.relaxed_madd`.

Method

- **Host:** M1 Pro, macOS, wasmtime 44.0.0, rustc 1.95.0
- **Build:** `cargo build --release --target wasm32-wasip1`
- **RUSTFLAGS:**
 - Baseline: `-C target-feature=+simd128`
 - Relaxed: `-C target-feature=+simd128,+relaxed-simd`
- All other variables (model, runtime, host) held constant within each comparison.

Kernel-level — microbench_8x8 GEMM

Shape	Baseline ns	Relaxed ns	Δ	Speedup
DFN3 m=64 k=64 n=8	2,547	1,845	-27.6%	1.38x
m=64 k=64 n=64	18,298	13,233	-27.7%	1.38x
m=128 k=128 n=8	8,679	6,039	-30.4%	1.44x
m=128 k=128 n=64	67,005	45,533	-32.1%	1.47x
m=256 k=256 n=8	31,581	21,164	-33.0%	1.49x
m=256 k=256 n=64	249,803	165,875	-33.6%	1.51x
m=384 k=1536 n=8	267,932	172,875	-35.5%	1.55x

Speedup grows with K (FMA dependency-chain reduction).

E2E models — wasm-model-bench

Model	Baseline ms	Relaxed ms	Δ	Speedup	Spread
Inception v3 (NNEF)	427.3	293.2	-31.4%	1.46x	1-9%
SqueezeNet 1.1 (ONNX)	30.9	24.2	-21.9%	1.28x	2-3%
modnet 512×512 (ONNX)	1,875.9	1,586.6	-15.4%	1.18x	3-10%
all-MiniLM-L6-v2 b=1 s=128 (ONNX)	103.0	93.8	-8.9%	1.10x	6-14%
DFN3 erb_dec T=100 (ONNX)	14.65	13.41	-8.5%	1.09x	1-3%
MobileNet v2 (ONNX)	75.2	69.1	-8.0%	1.09x	7-8%
DFN3 df_dec T=100 (NNEF)	11.9	11.0	-7.4%	1.08x	1-5%

DFN3 df_dec T=100 (ONNX)	12.58	11.60	-7.8%	1.08×	1-13%
--------------------------	-------	-------	-------	--------------	-------

Vision CNNs (8x8-dominated): 1.18-1.46x. Transformer attention: 1.10x. RNN-style audio (GEMV-heavy at M=256, bandwidth-bound) and depthwise-heavy CNNs: 1.08-1.10x. All consistent with the per-kernel speedup table.

Quality — L2 norms

TRACT_BENCH_QUALITY=1 runs:

Model	Output shape	Baseline L2	Relaxed L2	xor differs
Inception v3	[1, 1001]	6.477089e-2	6.477089e-2	yes (FMA bit-pattern)
DFN3 df_dec	[1, 100, 96, 10]	1.080686e-2	1.080686e-2	yes

Per-element values diverge in the 7th-8th decimal place — exactly FMA single-rounding. Within `Approximation::Close` (1e-4).

Real-world chained inference (DFN3 + libDF)

VoiceBank+DEMAND, 10 clips, 48 kHz, libDF ported to canonical tract. **This number includes #2195's sigmoid/tanh contribution as well** (already merged on main; reported separately in #2195 as ~1% E2E).

variant	mean RTF	min clip	max clip	Δ vs Rikorose
Rikorose 0.21.4 (chronological baseline)	0.1371	0.118	0.159	--
0.22.1 + relaxed-simd flag, no macro	0.1336	0.119	0.146	-2.6%
Canonical main + this PR	0.0537	0.044	0.083	-60.8% / 2.55×

Of the 60.8% RTF reduction, ~1.0 percentage point traces to #2195 (per #2195's bench), the remainder to this PR + the kernel kit (#2164/#2173/#2192).

Quality on chained DFN3 (vs Rikorose baseline)

metric	value
max abs Δ	3.05e-5 (= 1 LSB of 16-bit PCM, i.e. quantisation floor)
RMSE	4-6e-7
sample-level bit-equality	100.0% on every clip
SNR (signal/diff)	96-103 dB
DNSMOS OVR / SIG / BAK / P808	2.7442 / 3.1185 / 3.6900 / 3.5413 — identical to Rikorose

Bytecode op counts

`wasm-objdump -d tract_linalg-*.wasm` on the test binary:

Build	f32x4.relaxed_madd
+simd128	0
+simd128,+relaxed-simd	1044

The 1044 figure breaks down as ~972 from #2195's sigmoid/tanh (already merged) plus 70 from this PR's MMM substitutions plus ~2 LLVM opportunistic.