

C++ SIMD

The support for these instructions is wide but not universal. Both Intel and AMD support the compatible version of FMA, called FMA 3, in their CPUs released since 2012-2013. See hardware support section for more info.

Another caveat, the latency of FMA is not great, 4-5 CPU cycles on modern CPUs. If you're computing dot product or similar, have an inner loop which updates the accumulator, the loop will throttle to 4-5 cycles per iteration due to data dependency chain. To resolve, unroll the loop by a small factor like 4, use 4 independent accumulators, and sum them after the loop. This way each iteration of the loop handles 4 vectors independently, and the code should saturate the throughput instead of stalling on latency. See this [stackoverflow answer](#) for the sample code which computes dot product of two FP32 vectors.

<http://const.me/articles/simd/simd.pdf>

Revision #1

Created 11 January 2025 09:46:27 by Colin

Updated 12 January 2025 06:32:22 by Colin