Hybrid Custom Floating-Point Multiply-Accumulate Unit

IEEE 754 single-precision floating-point

Sign Exponent Mantissa

8-bit 23-bit

S E E E E E E E E M M M M M M M M M M M M M M M M M M M M M M M M

Float4

4-bit

E E E E

E_H

M_H

E_F
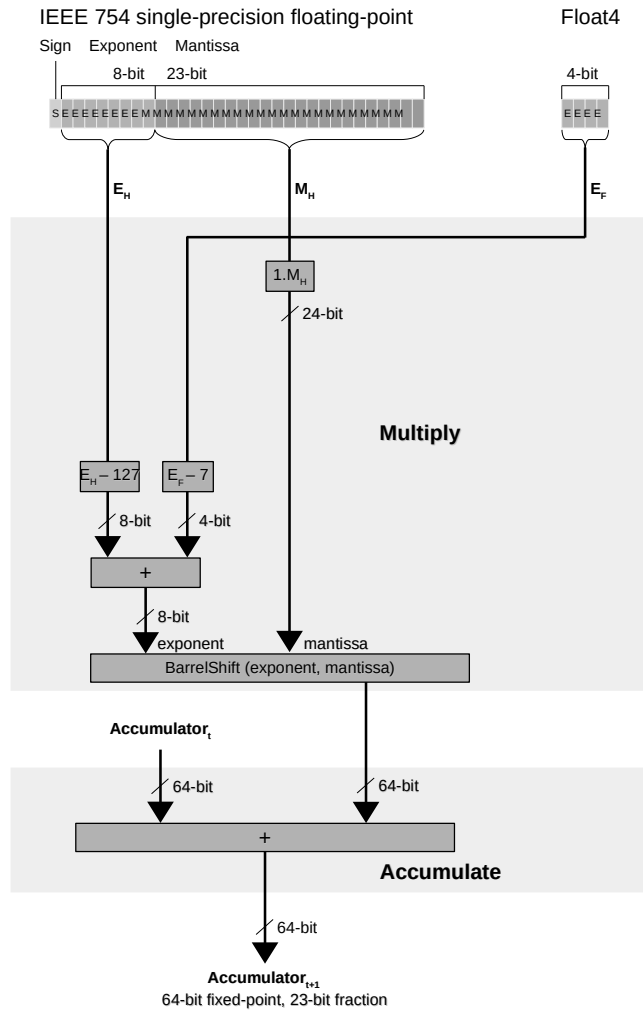
For **noise-robust applications** with **non-negativity constraints**.

$1.M_H$

24-bit

Non-negative Tensor Factorization

Non-negative Sparse Coding

Non-negative Matrix Factorization

Spike-by-Spike Neural Networks

Etc.

Multiply

$E_H - 127$ $E_F - 7$

8-bit   4-bit

+

8-bit

exponent   mantissa

BarrelShift (exponent, mantissa)

Accumulator$_t$

64-bit   64-bit

+

Accumulate

64-bit

Accumulator$_{t+1}$
64-bit fixed-point, 23-bit fraction