Hybrid Custom Floating-Point Multiply-Accumulate Unit

IEEE 754 single-precision floating-point

Float4

Sign Exponent Mantissa

8-bit   23-bit

4-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

E E E E

E_H

M_H

E_F

$1.M_H$

24-bit

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