HFP8 MAC

IEEE 754 single-precision floating-point   FP 8

Sign   Exponent   Mantissa

8-bit   23-bit

5-bit   2-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   S E E E E E M M

S_H E_H M_H S_F E_F M_F

$1.M_H$

a) Multiply

24-bit

Mantissa multiplication

>> 1   >> 2

2^{-1} 2^{-2}

2-bit

$E_H - 127$ $E_F - 15$

8-bit   5-bit   24-bit   24-bit   24-bit

+   +

8-bit   24-bit
exponent   mantissa

BarrelShift (exponent, mantissa)

Accumulator$_t$

64-bit   64-bit

+ $-/+$ $S_H \oplus S_F$

b) Accumulate

64-bit

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