Topics

26 Floating Point Formats

All MSP430 floating-point formats consist of three fields: an exponent field (e), a single-bit sign field (s), and a fraction field (f). The sign field and fraction field may be considered as one unit and referred to as the mantissa field. The fraction contains an implied most-significant bit, which is always 1 for a correctly represented floating-point constant. This provides an additional bit of precision. The exponent is bias 128; that is, subtract 128 from the unsigned value of the 8 exponent bits to arrive at an actual value for the exponent. A sign, exponent and fraction of zero is used as a special representation of value zero.

26.1 Single Precision Format

In the single precision format, the floating-point number is represented by an 8-bit exponent, a sign bit and a 23-bit fraction.

The format is as follows:

e	e	e	e	e	e	e	e		s	f	f	f	f	f	f	f		f	f	f	f	f	f	f	f		f	f	f	f	f	f	f	f

The fraction contains 23 actual bits plus an implied bit f_0, always representing a 1. The value of each f; is arrived at through this formula:

$$f_i = \frac{1}{2^i} \quad \Rightarrow \quad f = \sum_{i=0}^{23} f_i = \sum_{i=0}^{23} \frac{1}{2^i}$$

Therefore, the layout in terms of values is

$$1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}, \frac{1}{256}, \frac{1}{512}, \ldots$$

Example: Calculating the fraction (80 in those examples is the exponent)

Floating Point Value	**f_i**	**Fraction Decimal Equivalent**
80100000	$1, \frac{1}{8}$	$1 + \frac{1}{8} = 1.125$
80310000	$1, \frac{1}{4}, \frac{1}{8}, \frac{1}{128}$	$1 + \frac{1}{4} + \frac{1}{8} + \frac{1}{128} = 1.3828125$

Given the above format, some examples of acceptable floating-point values are shown in the following examples.

Example: Calculating Floating-Point Values

81d00000

Exponent									Sign									Fraction																
1	0	0	0	0	0	0	1		1	1	0	0	0	0	0	0		0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0

The encoded exponent equals 129; the real exponent equals 129-128 = 1.
The fraction equals 1 (implied f_0) $+ \frac{1}{2} + \frac{1}{8} = 1.625$.

The following formula expresses the actual value of the floating-point number:

$$s \times f \times 2^{e-128}$$

where s is the sign of the number (either 1 or -1), f is the value of the fraction ($1.0 \leq f < 2.0$) and e is the represented value of the exponent.

Therefore, the floating-point value is

$$-1 \times 1.625 \times 2^{129-128} = -3.25$$

The following list gives other examples of proper floating-point values derived from the above formulas.

80000000h	=>	1.0	81500000h	=>	3.25
80800000h	=>	- 1.0	8f3b8000h	=>	4.8e4
00000000h	=>	0.0	840c0000h	=>	1.75e1
83200000h	=>	1.0e1	79937500h	=>	- 9.0e-3

26.2 Double Precision Format

The only difference to the single precision format is the length of the fraction:

e e e e e e e e	s f f f f f f f	f f f f f f f f	f f f f f f f f	f f f f f f f f	f f f f f f f f

Here it contains 39 actual bits plus an implied bit fo; so the summation formation for the fraction changes to:

$$f = \sum_{i=0}^{39} f_i = \sum_{i=0}^{39} \frac{1}{2^i}$$