Floating-point

Van Duc NGUYEN

Viettel IC Design

Over View

1. Introduce
2. Floating-point Representation
3. IEEE 754 floating-point standard for binary
4. Floating-point Arithmetic
5. Implementation in FPGA

1.Introduce

Why Floating-point?

1. Advantages
   - Dynamic range
     with fixed-point $DR_{fxpt} = r^n - 1$.
     with floating-point $DR_{flpt} = \frac{M_{max} * b^{E_{max}}}{M_{min} * b^{E_{min}}}$
2. Disadvantages
   - Precision
   - Roundoff error
   - Complex implementation

2.Floating-point Representation

Form represent a floating-point number:

With three fields:

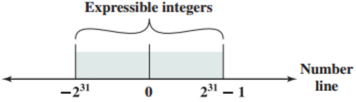
- Sign bit:S (0 is positive and 1 is negative)
- Fraction:F (Significand or mantissa)
- Exponent:E

Value of floating-point number:$(-1)^S * F * B^{\pm E}$
with B [1]

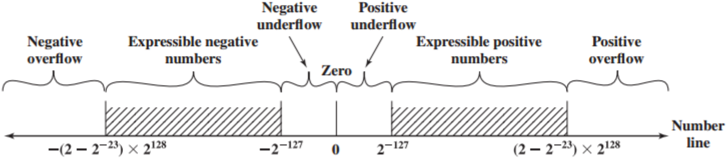
[1] The base B is implicit.(base is 2,10..)

Normalized and Denormalized representation

- Normalized:$\pm 1.mmm...m * B^{\pm E}$
- Denormalized:$\pm 0.mmm...m * B^{E_{min}}$



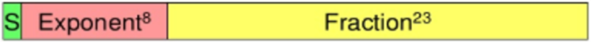
(a) Twos complement integers



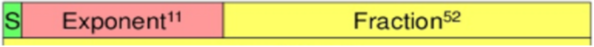
(b) Floating-point numbers

3.IEEE 754 Floating-point Standard for binary

The three basic format have bit lengts of 32,64 and 128 bits:



a) Binary32 format



b) Binary64 format

For a normalized floating-point number:

Value of floating-point: $\pm 1.f_1 f_2 f_3 f_4 ... f_l * 2^{\pm E}$

IEEE 754 Format Paremeter

Parameter	Format		
	Binary32	**Binary64**	**Binary128**
Storage width (bits)	32	64	128
Exponent width (bits)	8	11	15
Exponent bias	127	1023	16383
Maximum exponent	127	1023	16383
Minimum exponent	-126	-1022	-16382
Approx normal number range (base 10)	$10^{-38}, 10^{+38}$	$10^{-308}, 10^{+308}$	$10^{-4932}, 10^{+4932}$
Trailing significand width (bits)*	23	52	112
Number of exponents	254	2046	32766
Number of fractions	2^{23}	2^{52}	2^{112}
Number of values	1.98×2^{31}	1.99×2^{63}	1.99×2^{128}
Smallest positive normal number	2^{-126}	2^{-1022}	2^{-16362}
Largest positive normal number	$2^{128} - 2^{104}$	$2^{1024} - 2^{971}$	$2^{16384} - 2^{16271}$
Smallest subnormal magnitude	2^{-149}	2^{-1074}	2^{-16494}

Biased Exponent Representation

IEEE 754 use biased representation for the exponent:

- Value of exponet= val(E)=E - Bias(Bias is a constant)
- Bias is computed base on $bias = 2^{k-1} - 1$ (with k is lengths of bit)
- For signle precision,k=8 and bias=127,value of E(biased)=val(E)+127.

Special value

IEEE 754 define some special value as NaN,Infinity to represent for underflow,overflow and not a number...

Single-Precision	Exponent = 8	Fraction = 23	Value
Normalized Number	1 to 254	Anything	$\pm (1.F)_2 \times 2^{E-127}$
Denormalized Number	0	nonzero	$\pm (0.F)_2 \times 2^{-126}$
Zero	0	0	± 0
Infinity	255	0	$\pm \infty$
NaN	255	nonzero	NaN

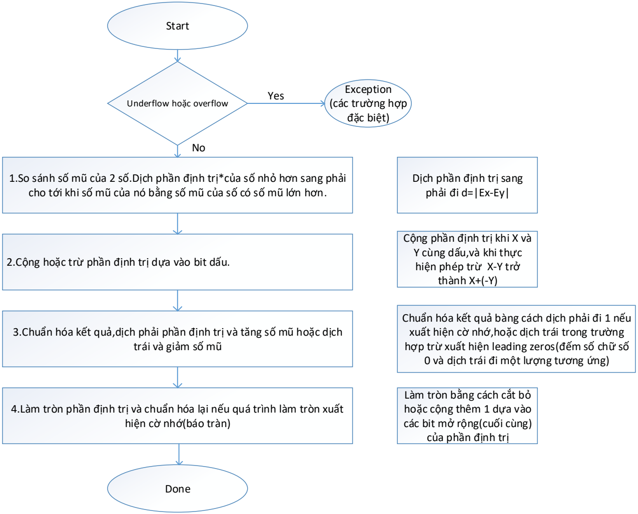
Double-Precision	Exponent = 11	Fraction = 52	Value
Normalized Number	1 to 2046	Anything	$\pm (1.F)_2 \times 2^{E-1023}$
Denormalized Number	0	nonzero	$\pm (0.F)_2 \times 2^{-1022}$
Zero	0	0	± 0
Infinity	2047	0	$\pm \infty$
NaN	2047	nonzero	NaN

3.Floating-point Arithmetic

Basic opreations for floating-point arithmetic:

Floating-Point Numbers	**Arithmetic Operations**
$X = X_S \times B^{X_E}$ $Y = Y_S \times B^{Y_E}$	$\left.\begin{array}{l} X + Y = (X_S \times B^{X_E - Y_E} + Y_S) \times B^{Y_E} \\ X - Y = (X_S \times B^{X_E - Y_E} - Y_S) \times B^{Y_E} \end{array}\right\} X_E \leq Y_E$ $X \times Y = (X_S \times Y_S) \times B^{X_E + Y_E}$ $\frac{X}{Y} = \left(\frac{X_S}{Y_S}\right) \times B^{X_E - Y_E}$

Addition and Subtration



4.Implementation in FPGA

5.Conclusion