Floating Point Unit

Lecture By Dr. Guy Even

Summarized By Guy Cohen

Table of contents

I will try to relate as close as possible to the Article " *On The Design of IEEE Compliant Floating Point Unit* " By Dr. Guy Even and Wolfgang Paul.

1 IEEE Standard – Representable Numbers

1.1 Factoring

Every real number X can be Factored into a Floating point number using three Factors:

1. s – Sign bit {0,1}.
2. e – A scale Factor Exponent {Integer}.
3. f – A Significant {None negative Real number which is usually between [0,2)

$$X = \text{Val}(s,e,f) = (-1)^s \bullet 2^e \bullet f$$

There are two more values which have to be defined $\pm\infty$:

- The Factoring of $+\infty$ is $(0, e_\infty, f_\infty)$
- The Factoring of $-\infty$ is $(1, e_\infty, f_\infty)$

We usually relate to two different unique representations:

1. $f \in [0,1)$ – Every real number has only one unique factoring excluding Zero which has many different factoring.
2. $f \in [1,2)$ – The problem with this representation is the factoring of Zero which doesn't exists and the representation of very small numbers (The exponents would be very small).

The standard tries to combine these two representations and take the good out of the two, using the Normalized Factoring, which will be introduced later on.

1.2 Standard Representable Real Numbers

In this section we define which numbers can be represented when using the standard.

The Exponent is defined using three parameters

- Two extremes values which represent the rainbow of different values between: $[e_{min}, e_{max}]$.
- A parameter n which defines the length of the exponent string (in bits).

e_{min} and e_{max} are determined as follows:

- $e_{min} \equiv 1 - bias$
- $e_{max} \equiv 2^n - 2 - bias$

Where $bias \equiv 2^{n-1} - 1$

This representation may seem a little strange at first but the next example will show its advantages:

$n = 3$ yields $e_{min} = -2$ and $e_{max} = 3$, a range of numbers different from the straight forward method without negative numbers.

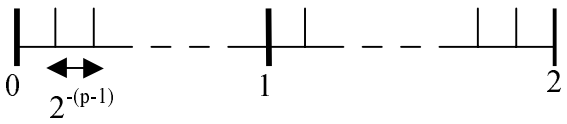
The Significant has only one parameter

- A parameter p which defines the length of significant (in bits).

The representable value of the significant is:

$$\left\{ i \cdot 2^{-(p-1)} \right\}_{i=0}^{(2^{(p-1)}-1)\cdot 2}$$

This gives us a number, which is a multiple of $2^{-(p-1)}$ in the half open interval [0,2).



[*denormalized* – זעיר)[*normalized* – מנורמל)

- There is a certain connection between the Exponent and the Significant, Not all exponents can go with every significant, There is a division into two different parts:
  1. *Normalized f* with any *e*.
  2. Denormalized f with e_{min}.

1.3 The Geometry of representable numbers

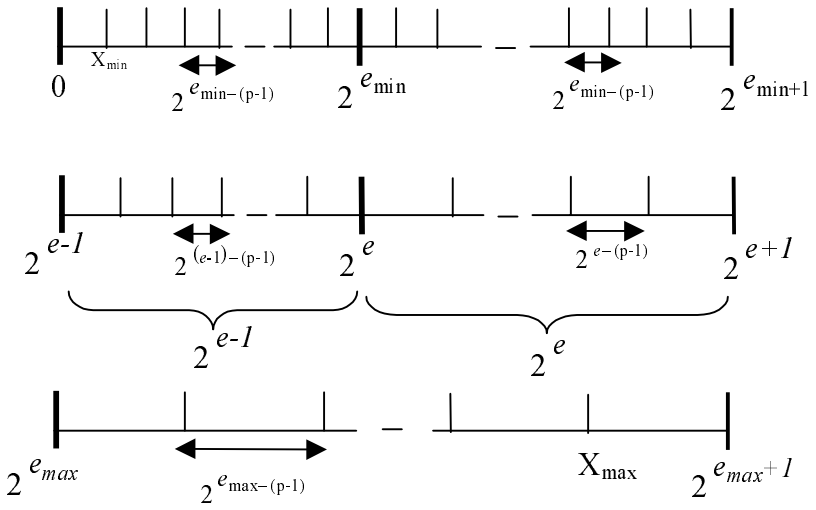
The set of numbers can be represented in geometric way (only positive numbers will be presented here since the negatives are symmetrical).
The set will be partitioned into three parts:

1. $[0, 2^{(e_{min}+1)}]$:
   - The gaps between two consecutive numbers are $2^{e_{min}-(p-1)}$.
   - The gaps between the two interval $[0, 2^{e_{min}})$ and $[2^{e_{min}}, 2^{e_{min}+1})$ are equal, since we have the same exponent for all the section ($2^{e_{min}}$).
   - $X_{min} = 2^{e_{min}-(p-1)}$,since the *Unit in the last position* is $2^{-(p-1)}$ which is multiplied by $2^{e_{min}}$.
   - There is a large gap between Zero and $2^{e_{min}}$,which is filled with *denormalized* values, this property is called *gradual underflow*.

2. $[e_{min}, e_{max}]$:
   - It is noted that as the exponent value is increased by one the gaps between two representable numbers double (the gap between two representable numbers is $2^{e-(p-1)}$ for ant given e between $[e_{min}, e_{max}]$, when the exponent is increases by one the gap changes to $2^{e+1-(p-1)}$ which is double the latter). Although the number of representable numbers is fixed and equal to $2^{(p-1)}$. This phenomenon causes Rounding problems around numbers that are near the power of two.

3. $[2^{e_{max}}, 2^{e_{max}+1})$:
   - $X_{max} = 2^{e_{max}+1} - 2^{e_{max}-(p-1)} = 2^{e_{max}}\underbrace{(2 - 2^{-(p-1)})}_{f_{max}}$

The graphic representation of the three parts:



2 Rounding

The article deals with only one kind of rounding but actually there are four different types of rounding.

1. **Rounding to Zero** – Rounds to the nearest representable number closest to Zero.
2. **Rounding to +∞ -** Rounds to the nearest representable number closest to + .
3. **Rounding to -∞ -** Rounds to the nearest representable number closest to - .
4. **Rounding to nearest (even) -** Rounds to the nearest even representable number

These rounding techniques come in handy in the significant rounding part.
There are four different stages when rounding a number:

1. Normalized Factoring
2. Significant Rounding
3. Post Normalization
4. Exponent Rounding

2.1 Normalized Factoring

If we didn't have a restriction on the Exponent we would define Normalized Factoring as:

$\eta` : \mathbf{R} \longrightarrow$ Factoring

$\eta`(x) = (s,e,f)$

With the following conditions:

1. Val $(s,e,f) = x$
2. $f \in [1,2)$

There are several problems with this definition (as said before)

- Large Exponent (we deal with this later in Exponent Rounding)
- Small Exponents.
- Zero is not representable.

So the definition will have to change a little bit:

$$\eta(x) = \begin{cases} \eta`(x) & \text{if} \quad \text{abs}(x) \geq 2^{e_{\min}} \text{ (Not to small)} \\ (s,e_{\min},f) & \text{if} \quad \text{abs}(x) < 2^{e_{\min}} \\ & \text{Where } f \in [0,1) \text{ and Val } (s,e,f) = x \\ & \rightarrow \quad f = (\text{abs}(x)/\, 2^{e_{\min}}) \bullet (-1)^s \end{cases}$$

We will also define two other values:

- $\eta(+\infty) = (0, e_\infty, f_\infty)$
- $\eta(-\infty) = (1, e_\infty, f_\infty)$

2.2 *Significant Rounding*

The rounding of the Significant takes the Significant and rounds it to its binary representation with at most p-1 bits to the right of the binary point.
First we sandwich f between two representable numbers:

$$q \cdot 2^{-(p-1)} \le f < (q+1) \cdot 2^{-(p-1)}$$

And we copy to one of these two values.
The copying depends on three parameters:

1. The value of f.
2. The rounding type.
3. The Sign bit.

We will define the rounding for the four different rounding types that were mentioned earlier.

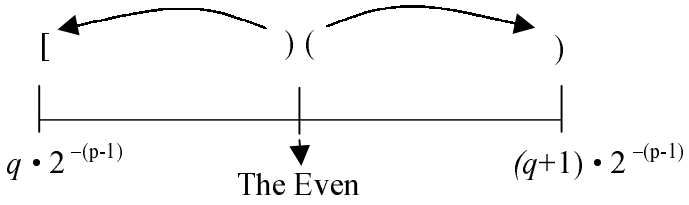
$$Sig\_rnd_0(f) = q \cdot 2^{-(p-1)}$$

$$Sig\_rnd_{+\infty}(f) = \begin{cases} q \cdot 2^{-(p-1)} & \text{if s=1} \\ (q+1) \cdot 2^{-(p-1)} & \text{if s=0} \end{cases}$$

$$Sig\_rnd_{-\infty}(f) = \begin{cases} q \cdot 2^{-(p-1)} & \text{if s=0} \\ (q+1) \cdot 2^{-(p-1)} & \text{if s=1} \end{cases}$$

$$Sig\_rnd_{ne}(f) = \begin{cases} q \cdot 2^{-(p-1)} & \text{if } q \cdot 2^{-(p-1)} \le f < (q+0.5) \cdot 2^{-(p-1)} \\ q' \cdot 2^{-(p-1)} & \text{if } f = (q+0.5) \cdot 2^{-(p-1)} \\ (q+1) \cdot 2^{-(p-1)} & \text{if } (q+0.5) \cdot 2^{-(p-1)} < f < (q+1) \cdot 2^{-(p-1)} \end{cases}$$

Where q' is the even integer in $\{q, q+1\}$



The article deals with rounding to the Nearest Even.

2.3 Post Normalization

After Significant rounding there is a possibility that the rounding gave us a significant with the value of two. That is we started from $f \in [0,2)$ and after Significant rounding we ended up with $f \in [0,2]$.
When $Sig\_rnd_{ne}(f) = 2$ it is called *Significant Overflow*.
The Post Normalization comes to correct this problem and its defined as follows:

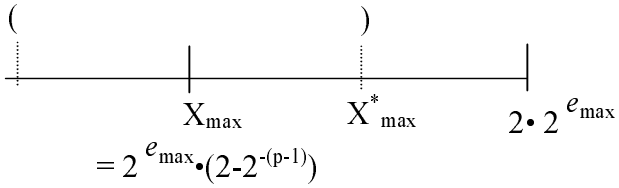
$$Post\_norm(s,e,f) = \begin{cases} (s, e+1, f) & \text{if } f=2 \\ (s,e,f) & \text{Otherwise} \end{cases}$$

2.4 Exponent Rounding

Exponent Rounding deals with the case that the absolute value of the factoring is too large (or too small) and is defined as follows:

$$exp\_rnd(s,e,f) = \begin{cases} (s, e, f) & \text{if } 2^e \cdot f \geq X^*_{max} \\ \eta(Val(s,e,f)) & \text{otherwise} \end{cases}$$

Where $X^*_{max} = 2^{e_{max}} \cdot (2-2^{p})$ which is the middle point in the last gap.



We note two kinds of overflow:
1. *Big Overflow* – When the overflow is detected in the first stage $e>e+1$
2. *Small Overflow* – When the overflow is detected after Post Normalization.

- If $f \in [1,2)$ is an integral multiple of $2^{-(p-1)}$ (which is the case after Post Normalization) then we simplify Exponent rounding to:

$$exp\_rnd(s,e,f) = \begin{cases} (s, e_\infty, f_\infty) & \text{if } e \geq e_{max} \\ (s,e,f) & \text{otherwise} \end{cases}$$

2.5 Rounding

The mapping of reals into factoring.
Let x be a real number and the normalized factoring of x is (s,e,f) ($\eta(x)=(s,e,f)$) then the rounding of x is:

$$r(x)=exp\_rnd(post\_norm(s,e, sig\_rnd(f)))$$

3 Rounding Computation