# Berkeley Hard-Float 解析

Berkeley HardFloat 是二进制浮点运算的硬件实现，符合 IEEE 浮点运算标准。HardFloat 支持多种浮点格式，使用模块参数独立确定指数字段和有效位字段的宽度。可能的格式集包括 16 位半精度、32 位单精度、64 位双精度和 128 位四精度的标准格式。一些历史扩展格式，如英特尔的老 80 位双扩展精度浮点，不直接支持。（但是，HardFloat 可以用于实现这些和其他基于 IEEE 的格式，并添加了在编码之间转换的模块。）

Float Point 有三种主要的形式，IEEE 标准形式(简称 fN)，Recoder 形式，Raw Deconstructions。

## IEEE 标准形式(fN)

需要满足条件

$$p \leq 2^{(w-2)} + 3$$

其中 w 为 expWidth，p 为 sigWidth，一个 IEEE 标准形式浮点数长度为 w+p，其中 p 中包含一位符号位。

常见的 IEEE 浮点格式如下

| | expWidth(w) | sigWidth(p) |
|---|---|---|
| 16-bits half-precision | 5 | 11 |
| 32-bits single-precision | 8 | 24 |
| 64-bits double-precision | 11 | 53 |
| 128-bits quadruple | 15 | 113 |

## Recoded Formats(RecFN)

对于每种标准形式，HardFloat 都定义了一个对应的 recoded format，以稍微不同的表示形式对同一组值进行了编码。RecFN 和标准形式一样具有符号位、指数位和有效数字位，其中指数位扩展了一位，因此标准的 32bit 但精度浮点数的 RecFN 形式是 33 位，1 个符号位、9 个指数位(扩展一位)和 23 个有效数字位。RecFN 能够在一些方面简化浮点数运算，但是最重要的方面是 RecFN 能够将 subnormal 类型的数值正规化，因此可以将这类数值当作正规浮点数进行处理。下表总结了 recoding

| | Standard format | | | HardFloat's recoded format | | |
|---|---|---|---|---|---|---|
| | sign | exponent | significand | sign | exponent | significand |
| zeros | s | 0 | 0 | s | 000xx...xx | 0 |
| Subnormal numbers | s | 0 | F | s | $2^k$+2-n | normalized F<<n |
| Normal numbers | s | E | F | s | E+$2^k$+1 | F |
| infinities | s | 1111...11 | 0 | s | 110xx...xx | xxxx.xxx |
| NaNs | s | 1111...11 | F | s | 111xx...xx | F |

其中参数 k 是 expWidth-1，这里的 expWidth 是标准形式中的指数位数。x 代表 dont care，都与特殊的值 0，无穷大，NaNs，只有三位有意义，分别为 000，110，111，因此可以通过指数高三位来进行快速判断。否则这个值就是一个正常有限数值。其次，如果 recoderd 之后，指数部分大于等于 $2^k + 2$，那么该值就是一个正常数值，如果小于则该值就是一个正规化后的 subnormal 值。normalized F<<n 意思是 F 左移 n bit 进行正则化之后得到的值。

对于大部分浮点数(0，正常数值，NaNs)不同之处仅仅是指数部分进行了编码，其余部分一致。Infinities 特殊对待，仅仅有 3bit 有效。所以只有 subnormal 数值指数部分和有效数字部分因为正则化都进行了转换。

## 近似模式

2008 IEEE 浮点标准定义了 5 各近似模式，HardFloat 添加了一种近似模式，round to odd。

| | |
|---|---|
| round_near_even | round to nearest, with ties to even |
| round_near_maxMag | round to nearest, with ties to maximum magnitude (away from zero) |
| round_minMag | round to minimum magnitude (toward zero) |
| round_min | round to minimum (down) |
| round_max | round to maximum (up) |
| round_odd | round to odd (jamming) |

## 异常结果

HardFloat 支持 IEEE 浮点标准的所有 5 各异常标志。

{invalid, infinite, overflow, underflow, inexact}

Infinite 被定义为除以 0，意味着从有限操作数中得到了一个无穷大结果。