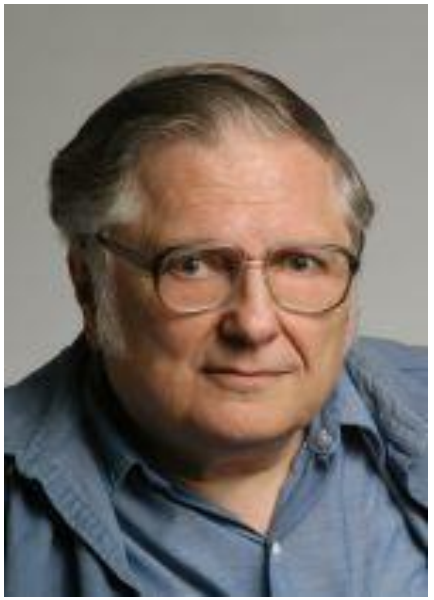
IEEE 754浮点数

IEEE 754 浮点数标准

1989
图灵奖获得者



Kahan教授

www.cs.berkeley.edu/~wkahan/

…/ieee754status/754story.html

浮点数标准 IEEE754

- 单精度 32位

- 双精度 64 位

- 扩展精度 80 位 (Intel)

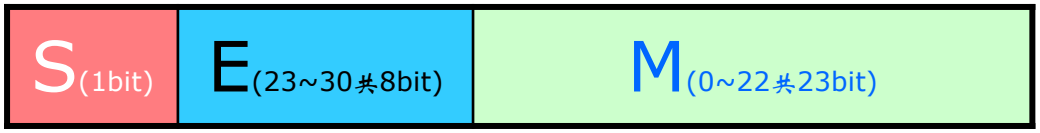
单精度浮点数编码格式

符号位S，阶码E，尾数M

符号位	阶码	尾数	表示
0/1	255	非零1xxxx	NaN Not a Number
0/1	255	非零0xxxx	sNaN 发信号的 NaN
0	255	0	+∞
1	255	0	- ∞
0/1	1~254	f	$(-1)^S \times (1.f) \times 2^{(e-127)}$
0/1	0	f(非零)	$(-1)^S \times (0.f) \times 2^{(-126)}$
0/1	0	0	+0/-0

IEEE754单精度浮点数标准

- 规格化数:



代表数值: $(-1)^S \times 1.m \times 2^{e-127}$

- $e_{min}=1$, $e_{max}=254$
- 阶: e减去偏移量 127, 表达范围: -126 ~ +127
- 尾数: 采用原码表示。IEEE754 将小数点前的1作为隐藏位缺省存储, 使得尾数表示范围比实际存储多一位。
- 规格化数的最高数字位总是1, 1.m 是规格化数。

举例：将如下十进制数用IEEE754 单精度浮点数用表示：

$v = (-1)^s \times 1.m \times 2^{e-127}$

- 数值 F = 15213.0;
  - $15213_{10} = 11101101101101_2$
    $= 1.1101101101101_2 \times 2^{13}$
- 尾数：**1.1101101101101_2**
  $M = 11011011011010000000000_2$
- 阶: 13
- E = 13+127 = 140 = **10001100_2**
- 单精度浮点数15213.0 的IEEE754编码 ：

0	10001100	11011011011010000000000
S	**E**	**M**

将如下IEEE754 单精度浮点数用十进制数表示:

1 10000001 01000000000000000000000

解: $s=1$, $e=129$, $f = 1/4 = 0.25$,

$(-1)^1 \times (1+0.25) \times 2^{129-127}$

$= -1 \times 1.25 \times 2^2$

$= -1.25 \times 4$

$= -5.0$

IEEE754 规格化浮点数表示范围

格式	最小值	最大值
单精度	$E_{min}=1$, M=0, $1.0\times2^{1-127} = 2^{-126}$	$E_{max}=254$, f=1.1111…, 1.111…1$\times2^{254-127}$ $= 2^{127}\times(2-2^{-23})$
双精度	$E_{min}=1$, M=0, $1.0\times2^{1-1023} =2^{-1022}$	$E_{max}=2046$, f=1.1111…,1.111…1$\times2^{2046-1023}$ $=2^{1023}\times(2-2^{-52})$

单精度: （有效尾数24位，相当于7位十进制有效位数）
双精度: （有效尾数53位，相当于17位十进制有效位数）

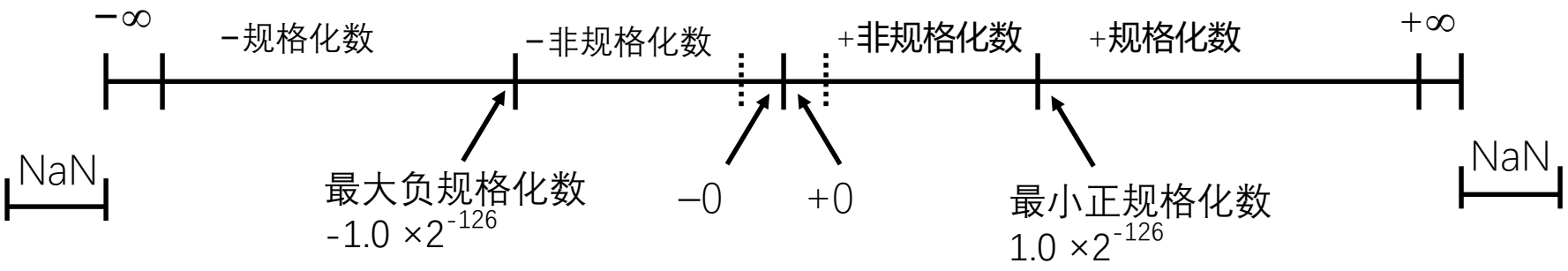
IEEE754 标准：零的表达

- **E = 000…0, M = 000…0**
  - 表达： 0
  - 注意不同的值 ：+0 和 –0

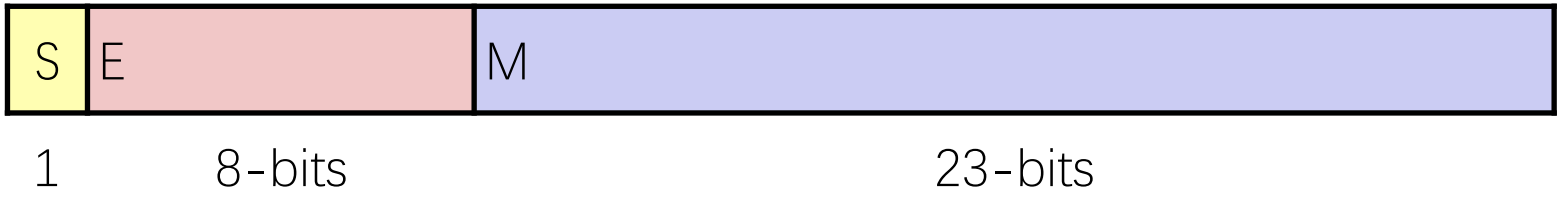
非规格化数

- 非规格化数(Subnormal)
- **E = 000…0, M ≠ 000…0** (E =0，M 非零)

代表： $(-1)^S \times 0.m \times 2^{-126}$



∞ (infinity)



E = 111…1, M = 000…0

- 表示 ∞ (infinity)
- 注意不同的值 : + ∞ and – ∞

- 一般是运算溢出 后得到的结果
- 例如: 1.0/0.0 = −1.0/−0.0 = +∞, 1.0/−0.0 = −∞

NaN

E = 111…1, M ≠ 000…0

- 不是一个数 Not-a-Number (NaN)
- 表达当数值无法确定时，
- 例如： sqrt(–1), ∞ − ∞, ∞ × 0

IEEE 754的特点

- 规格化尾数隐藏位缺省存储，使得尾数表示范围更大
- 提供了非规格化数、NaN, ∞ 等多样化的数据表达
- 浮点0的形式和 整数0的表达形式相同
- 几乎可以用Unsigned Integer 比较器直接比较大小
  - 非规格化 vs. 规格化
  - 规格化 vs. 无穷大
  - 除了：
    - 要先比较符号位，负数的符号位是1
    - 必须考虑 −0 = 0
    - NaNs 比任何其他数值都大