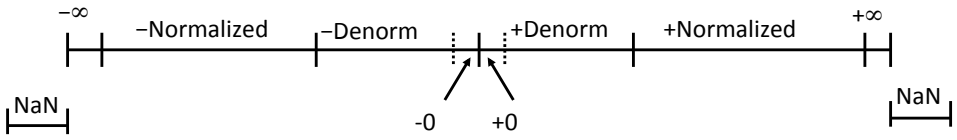
IEEE Floating Point (IEEE 754)

$$value = (-1)^s \; M \;\; 2^E$$

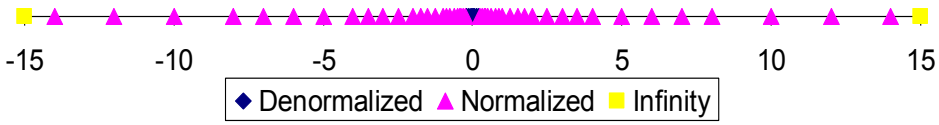
1 k = 8 or 11 23 or 52

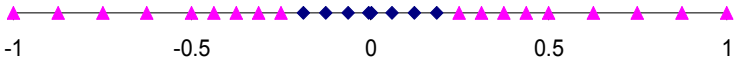
- Normalized
  - E = expU - bias
  - bias = 2^(k-1) - 1
  - M = 0b1.frac
- Denormalized
  - exp = 00...00
  - E = 1 - bias
  - M = 0b0.frac
- Special Values
  - exp = 11...11
  - Infinity (frac = 00...00)
  - NaN (frac ≠ 00...00)

Floating Point Visualization



6-bit values (3 exp, 2 frac)





Close-up (central values)