Computer Basics
IEEE Floats*

Rubin H Landau
With
Sally Haerer and Scott Clark

Computational Physics for Undergraduates
BS Degree Program: Oregon State University

"Engaging People in Cyber Infrastructure"
Support by EPICS/NSF & OSU

IEEE Standard Float (how)*

$$x_{\text{Float}} = (-1)^s \times 1.f \times 2^{\text{e - bias}} \quad (1)$$

- Sign s = single bit = 0 (+), 1 (-)
- f = fractional part after *binary* point
  - assume 1st bit = 1 (phantom)
  - maintains same relative precision
- e = stored exponent always > 0
- bias: fixed, e < bias $\Rightarrow$ p = true exp < 0
- *Normal* numbers: 0 < e < 255
- *Subnormal* numbers: e=0, e=255
  - special cases & numbers (table)

IEEE Special Cases

Number Name	*Values of s, e & f*	*Value of Single*
Normal	$0 < e < 255$	$(-1)^s \times 2^{e-127} \times 1.f$
Subnormal	$e = 0, \;\; f \neq 0$	$(-1)^s \times 2^{-126} \times 0.f$
Signed Zero	$e = 0, \;\; f = 0$	$(-1)^s \times 0.0$
$+\infty$ ($\neq$ math)	$s = 0, \;\; e = 255, \;\; f = 0$	+INF
$-\infty$ ($\neq$ math)	$s = 1, \;\; e = 255, \;\; f = 0$	-INF
Not a Number	$s = u, \;\; e = 255, \;\; f \neq 0$	NaN

Implementation: IEEE Single (float)*

Position	*s*	*e*		*f*	
32 Bit word	31	30	23	22	0

- **Conversion of Exponent *e***
  - biased exponent *e*: 8 bits

$$(-1)^s \times 1.f \times 2^{e-127} \quad (1)$$

  - normal: $0 < e < 255$
  - $\Rightarrow \quad 1 \le e \le 254$
  - bias = $127_{10} \quad \Rightarrow p = e_{10} - 127$
  - $-126 \le p \le 127$ (see limits)
- **Specials**
  - $e = f = 0: \quad \pm 0$

$$(-1)^s \times 0.f \times 2^{e-126} \quad (2)$$

  - $e = 0, \ f \neq 0$: mantissa = $0.f$

E.G.: Largest, Normal, 32-bit Float*

$$
\begin{aligned}
e_{max} \text{ (normal)} &= 254 \quad \Rightarrow \quad p = e - 127 = 127 && (1)\\
s &= 0 && (2)\\
f\text{max} &= 1.1111\,1111\,1111\,1111\,1111\,111 && (3)\\
&= 1 + 0.5 + 0.25 + \ldots \simeq 2 && (4)\\
\Rightarrow \quad & (-1)^s \times 1.f \times 2^{p=e-127} \simeq 2 \times 2^{127} && (5)\\
&\simeq 3.4 \times 10^{38} && (6)
\end{aligned}
$$

Time for Exercises in Lab

Exercise: IEEE Representation*
(by hand)

Consider the 32--bit single-precision floating point number

	s	e	f
Bit position	31	30 23	22 0
	0	0000 1110	1010 0000 0000 0000 0000 000

1. What are binary values for
   a. sign *s*
   b. exponent *e*
   c. fractional mantissa *f ?*

2. What are decimal values for
   a. for stored exponent *e*
   b. actual exponent *p*

3. Show that mantissa = 1.625 000

4. What is the full decimal value of *number?*

Exercise: Overflows & Underflows

Determine experimentally: underflow, overflow limits

- *IEEE* $\Rightarrow$ *Overflow:* $x_{SP} > 2^{128}$
- *IEEE* $\Rightarrow$ *Underflow:* $x_{SP} < 2^{-128}$
- $\Rightarrow$ x_c = *NAN, INF, ?*
- *Repeat for singles, doubles, ints, negatives*

```
under = 1.
over = 1.
begin do N times
        under = under/2.
        over = over * 2.
        write out: loop number, under, over
end do
```

