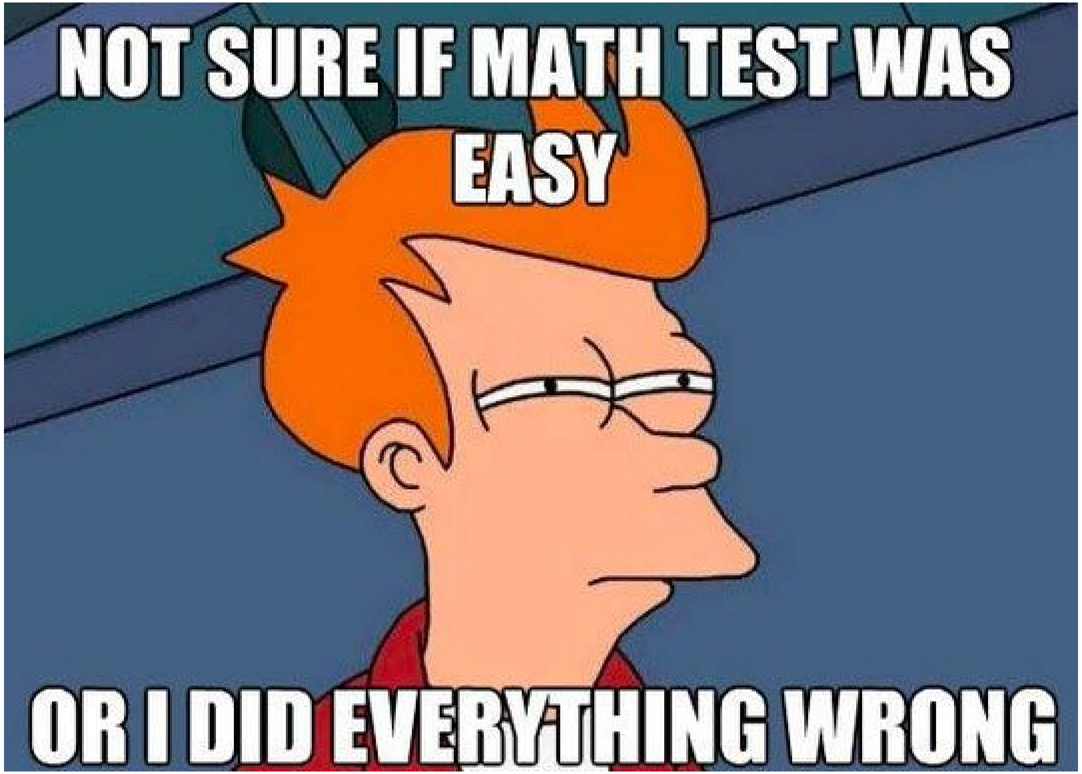
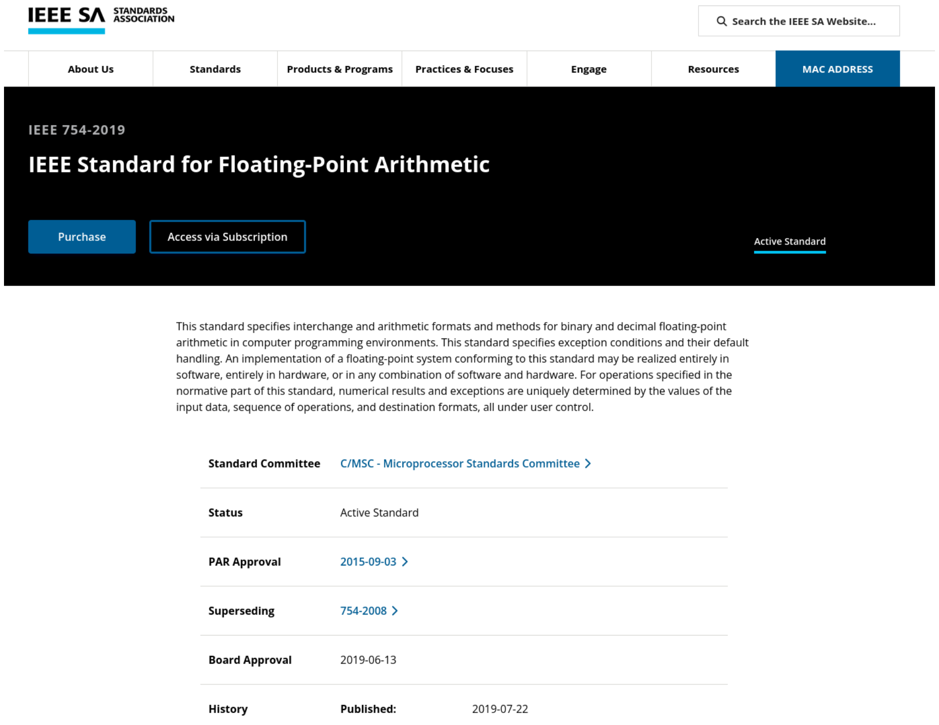
Surprise Quiz 1 Review



IEEE Floating Point Types

- IEEE Standard 754



Precision Types

- Wikipedia

			Significand		Exponent		Properties[b]				
Name	Common name	Radix	Digits[c]	Decimal digits[d]	Min	Max	*MAXVAL*	$\log_{10}$ *MAXVAL*	*MINVAL*>0 (normal)	*MINVAL*>0 (subnormal)	Notes
binary16	Half precision	2	11	3.31	−14	15	65504	4.816	6.10×10^{-5}	5.96×10^{-8}	Interchange
binary32	Single precision	2	24	7.22	−126	127	3.40×10^{38}	38.532	1.18×10^{-38}	1.40×10^{-45}	Basic
binary64	Double precision	2	53	15.95	−1022	1023	1.80×10^{308}	308.255	2.23×10^{-308}	4.94×10^{-324}	Basic
binary128	Quadruple precision	2	113	34.02	−16382	16383	1.19×10^{4932}	4932.075	3.36×10^{-4932}	6.48×10^{-4966}	Basic
binary256	Octuple precision	2	237	71.34	−262142	262143	$1.61 \times 10^{78\,913}$	78913.207	2.48×10^{-78913}	2.25×10^{-78984}	Interchange
decimal32		10	7	7	−95	96	1.0×10^{97}	$97 - 4.34 \times 10^{-8}$	1×10^{-95}	1×10^{-101}	Interchange
decimal64		10	16	16	−383	384	1.0×10^{385}	$385 - 4.34 \times 10^{-17}$	1×10^{-383}	1×10^{-398}	Basic
decimal128		10	34	34	−6143	6144	1.0×10^{6145}	$6145 - 4.34 \times 10^{-35}$	1×10^{-6143}	1×10^{-6176}	Basic

C Header Definition Review

- */usr/include/bits/types.h*
  - Defines all the Integer Types
- */usr/include/bits/floatn-common.h*
  - Doesn’t clearly definite float types
  - Floats are “Complex” types.

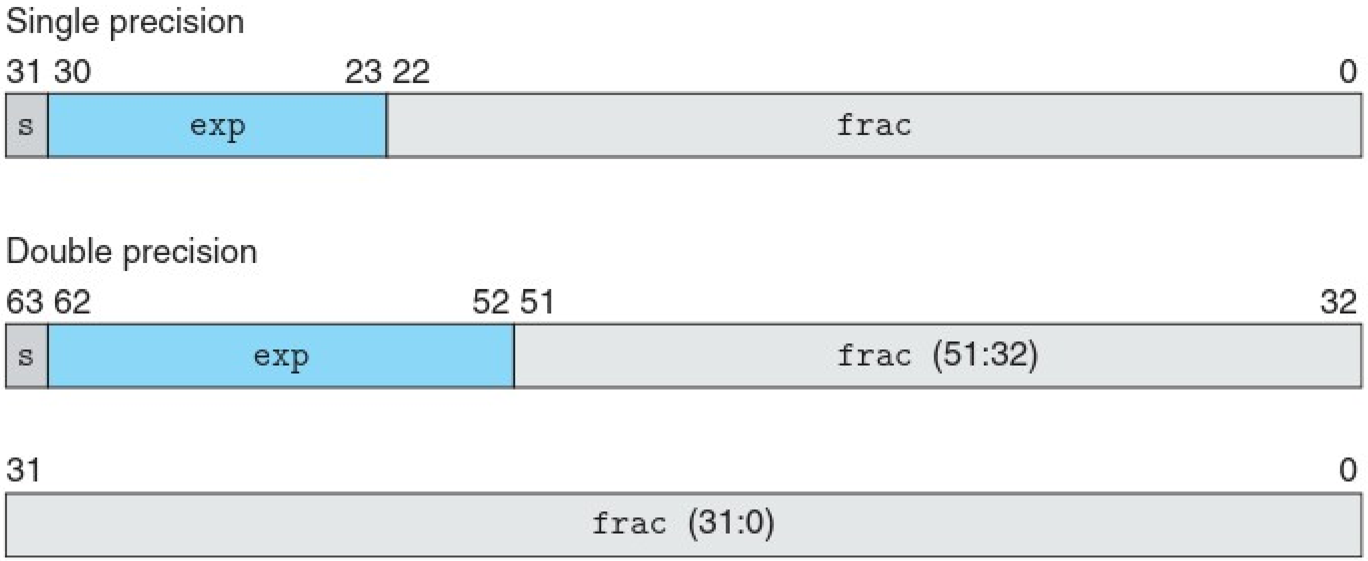
Floats Represent Binary Fractions

- Power of 2 fractions
- Surprise Quiz?

Representation	Value	Decimal
0.0_2	$\frac{0}{2}$	0.0_{10}
0.01_2	$\frac{1}{4}$	0.25_{10}
0.010_2	$\frac{2}{8}$	0.25_{10}
0.0011_2	$\frac{3}{16}$	0.1875_{10}
0.00110_2	$\frac{6}{32}$	0.1875_{10}
0.001101_2	$\frac{13}{64}$	0.203125_{10}
0.0011010_2	$\frac{26}{128}$	0.203125_{10}
0.00110011_2	$\frac{51}{256}$	0.19921875_{10}

The chunks and bits

- Higher precision means bigger



Single Precision

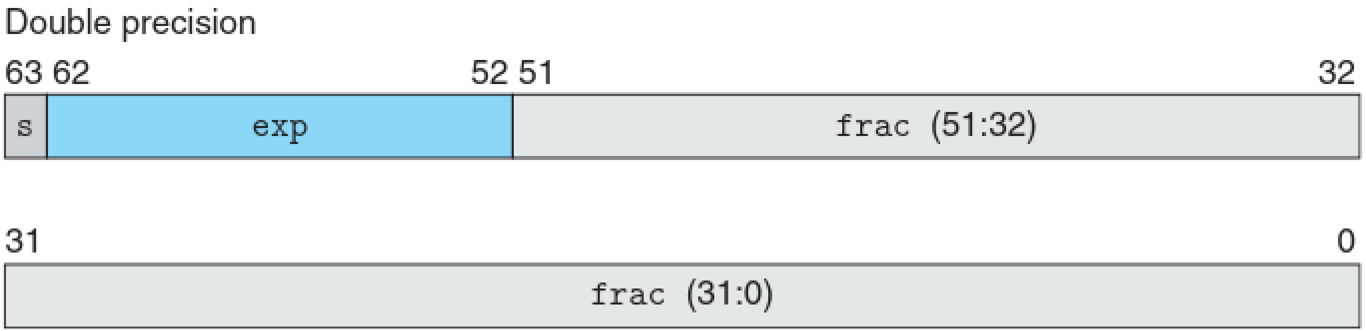
- MSbit is a simple sign +/-
- EXP is how you move the decimal point
  - Powers of 2 indicator → 2^exp

Single Precision

- Fraction is your binary number
  - Multiply by 2^exp to get the real number

Double Precision

- Can store more precision.



What is going on here (base 10)

- Binary scientific notation … sort of.

$$2469/200 = 12.345 = \underbrace{12345}_{\text{significand}} \times \underbrace{10}_{\text{base}}^{\overbrace{-3}^{\text{exponent}}}$$