Floating Point

15-213: Introduction to Computer Systems – Recitation
January 24, 2011

Today: Floating Point

- Data Lab
- Floating Point Basics
  - Representation
  - Interpreting the bits
  - Rounding
- Floating Point Examples
  - Number to Float
  - Float to Number

Data Lab

- Due tomorrow at 11:59pm
- Any questions?

Representation

- Basic format of bit representation (single precision):

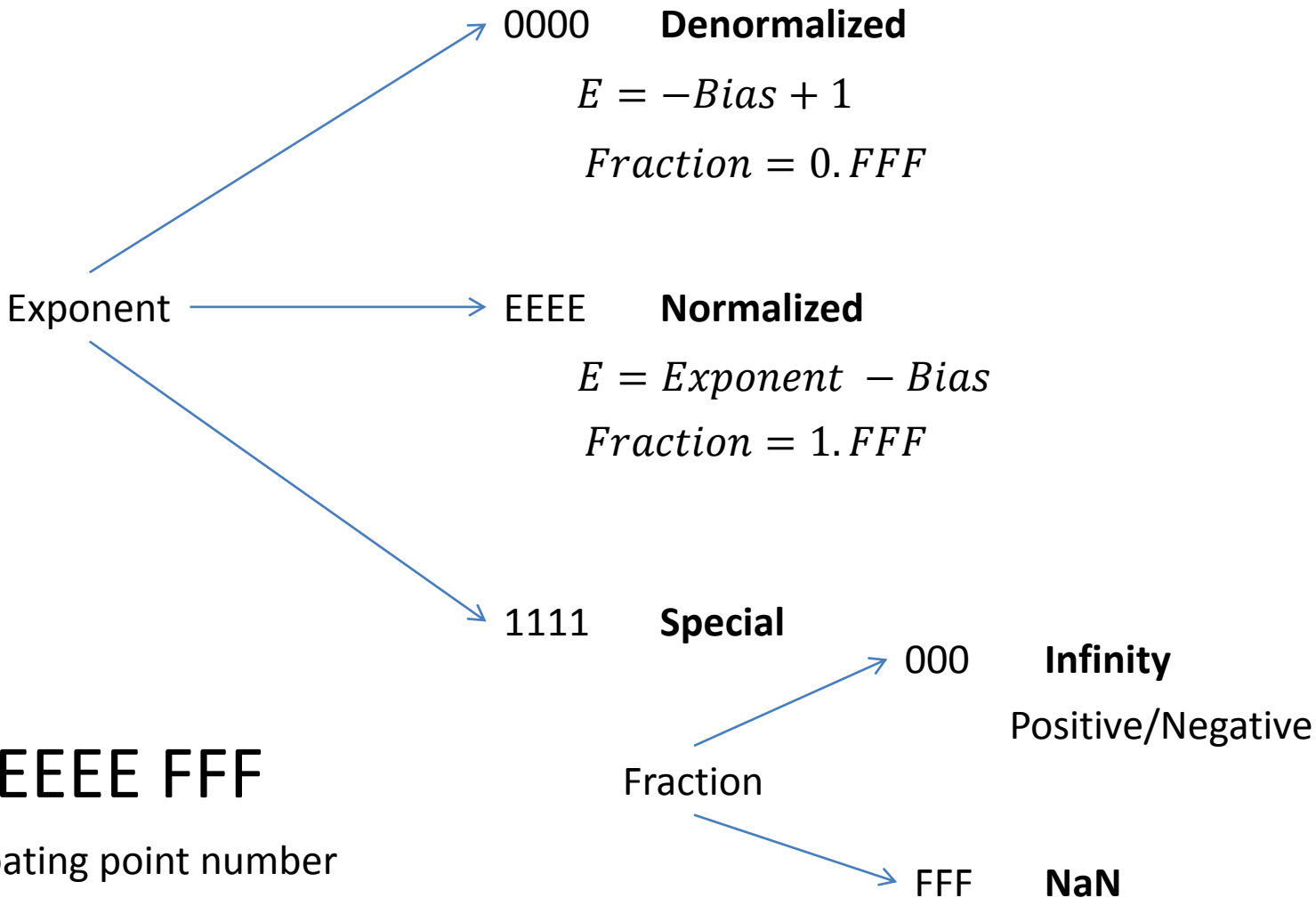
$$-1^{S} * M * 2^{E}$$

Where E is based on the Bias

$$Bias = 2^{k-1} - 1 = 2^{8-1} - 1 = 127$$

k = exponent bits

Interpreting the Bits



S EEEE FFF

8-bit floating point number

Rounding

- Round to even
  - Like regular rounding except for the exactly half case
  - If the last rounded bit is 1 round up, else round down

1.10 1001	Greater then 0.5, round up	1.11
1.10 0110	Less than 0.5, round down	1.10
1.11 1000	Round to even up	10.00
1.10 1000	Round to even down	1.10

Number to Float

- Convert: -5

S EEEE FFF

8-bit floating point number

Number to Float

- Convert: -5

$$Bias = 2^{4-1} - 1 = 2^3 - 1 = 7$$

- Negative so we know $S = 1$
- Turn 5 to bits:

$$5_{10} = 101_2$$

- Normalized value so lets fit in the leading 1

$$1.01_2$$

- Thus $F = 010$
- Now we figure out the exponent:

$$101_2 \rightarrow 1.01_2 \; x \; 2^2$$

Number to Float

- Calculate the exponent bits:

$$E = Exponent - Bias \rightarrow 2 = Exponent - 7 \rightarrow Exponent = 2 + 7 = 9$$

- So the exponent is 9, in bits:

$$Exponent = 1001_2$$

- Answer: $1\ 1001\ 010_2$

Number to Float

- Convert: 6/512

Number to Float

- Convert: 6/512
- Is it denormalized? Check the largest denormalized:

$$0\ 000\ 111_2 = 0.111_2\ x\ 2^{-6} = 111_2\ x\ 2^{-9} = 7\ x\ 2^{-9} = 7/512$$

- Denormalized, we know the exponent is going to be:

$$E = -Bias + 1 = -7 + 1 = -6$$

- So we know the form of the answer is going to be:

$$0.FFF\ x\ 2^{-6}$$

- Lets remove the decimal point to make it a bit easier:

$$FFF\ x\ 2^{-9}$$

- The fraction bits are the top of the fraction:

$$6_{10} = 110_2$$

Number to Float

- Why does that work? Lets remove the decimal point to make it a bit easier to see:

$$0.110\ x\ 2^{-6} = 110\ x\ 2^{-9}$$

- Remembering that these are fractions:

2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{256}$	$\frac{1}{512}$

- We can see the table in terms on 512ths:

2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}
$\frac{256}{512}$	$\frac{128}{512}$	$\frac{64}{512}$	$\frac{32}{512}$	$\frac{16}{512}$	$\frac{8}{512}$	$\frac{4}{512}$	$\frac{2}{512}$	$\frac{1}{512}$

- We want 6 512ths which are the bits we used.

Number to Float

- Putting it all together: 0 0000 110

Number to Float

- Convert: 27

Number to Float

- Convert: 27
- Positive so we know $S = 0$
- Turn 27 to bits:

$$27_{10} = 11011_2$$

- Normalized value so lets fit in the leading 1

$$1.1011_2$$

- But we only have 3 fraction bits so we must round. Digits after rounding equal half, last rounding digit is 1 so we round up.

$$1.1011_2 = 1.110_2$$

- Thus $F = 110$

Number to Float

- Calculate the exponent :

$$11100_2 \rightarrow 1.1100_2 \; x \; 2^4$$

- Calculate the exponent bits:

$$E = Exponent - Bias \rightarrow 4 = Exponent - 7 \rightarrow Exponent = 4 + 7 = 11$$

- So the exponent is 11, in bits:

$$Exponent = 1011_2$$

- Answer: $0\ 1011\ 110_2$

Float to Number

- Convert: $1\ 1001\ 010_2$

Float to Number

- Convert: $1\ 1001\ 010_2$
- Sign bit tells us it is negative
- We know it is normalized (non-zero exponent) so lets figure out the exponent:

$$1001_2 = 9_{10}$$

$$E = Exponent - Bias \rightarrow 9 - 7 = 2$$

- Now the fraction (remember the leading 1):

$$1.010_2$$

- Put it all together:

$$1.010_2\ x\ 2^2 = 101_2 = 5_{10}$$

- Answer: -5

Float to Number

- Convert: 0 0000 110

Float to Number

- Convert: $0\ 0000\ 110$
- Sign bit tells us its positive
- It is denormalized because of the 0 exponent so lets figure out the exponent:

$$E = -Bias + 1 \rightarrow -7 + 1 = -6$$

- Now the fraction (remember the leading 0):

$$0.110\ x\ 2^{-6}$$

- Put it all together:

$$0.110_2\ x\ 2^{-6} = 0.00000110_2$$

Float to Number

- Put it all together:

$$0.110_2 \; x \; 2^{-6} = 0.000000110_2$$

- Now lets examine our fraction chart:

2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}
$\frac{256}{512}$	$\frac{128}{512}$	$\frac{64}{512}$	$\frac{32}{512}$	$\frac{16}{512}$	$\frac{8}{512}$	$\frac{4}{512}$	$\frac{2}{512}$	$\frac{1}{512}$

- Answer: 6/512

Float to Number

- Convert: $0\ 1011\ 110_2$

Float to Number

- Convert: $0\ 1011\ 110_2$
- Sign bit tells us it is positive
- We know it is normalized (non-zero exponent) so lets figure out the exponent:

$$1011_2 = 11_{10}$$

$$E = Exponent - Bias \rightarrow 11 - 7 = 4$$

- Now the fraction (remember the leading 1):

$$1.110_2$$

- Put it all together:

$$1.110_2\ x\ 2^4 = 11100_2 = 28_{10}$$

- Answer: 28

Questions?