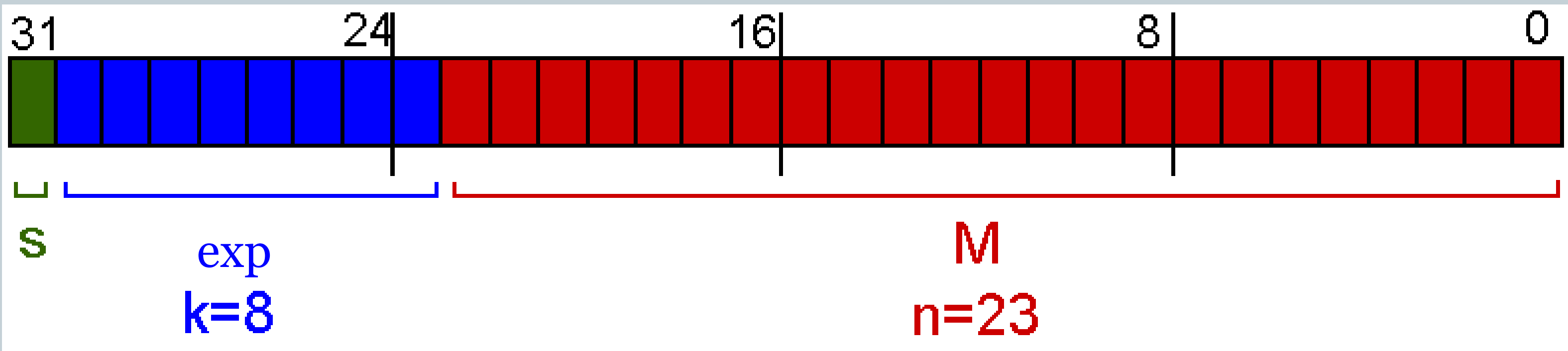
Floating Point and Datalab

MONDAY JANUARY 25TH, 2010

Announcements

- Datalab due Thursday Jan 28th at 11:59pm.
- You must make an official handin through Autolab. An unofficial handin will not be graded and your most recent official handin will be graded.
- Office Hours in WEH 5207 Sun-Thurs 6pm-9pm.
- IRC Chat Room ##213.

Floating Point IEEE Standard



E = exp - bias
Bias = 2^(k-1) -1
Value = (-1)^s * M * (2^E)

Note: exponent boundary is NOT aligned with byte boundary e.g. 0xFF7FFFFF has lowest exponent bit zero (is normalized v.)

For a double precision floating point number, k = 11, n= 52

Floating Point Normalization

- If exp is > 0 (and not all ones), then the mantissa has an assumed leading 1.
- 0x81200000
- Sign bit = ?
- Exponent = ?
- Mantissa = ?
- Final value = ?

Denormalization

- When exp = 0, it is a denormalized number.
- Formula for denormalized:
  - Value = (-1)^s * M * 2^(exp-bias+1)
- In a denormalized number, there is no assumed leading 1 in the Mantissa.

Representing Zero

- Since there is a sign bit, there exist both a positive and negative representation for zero.
- 0x80000000 and 0x00000000

Not a Number (NaN) and Infinity

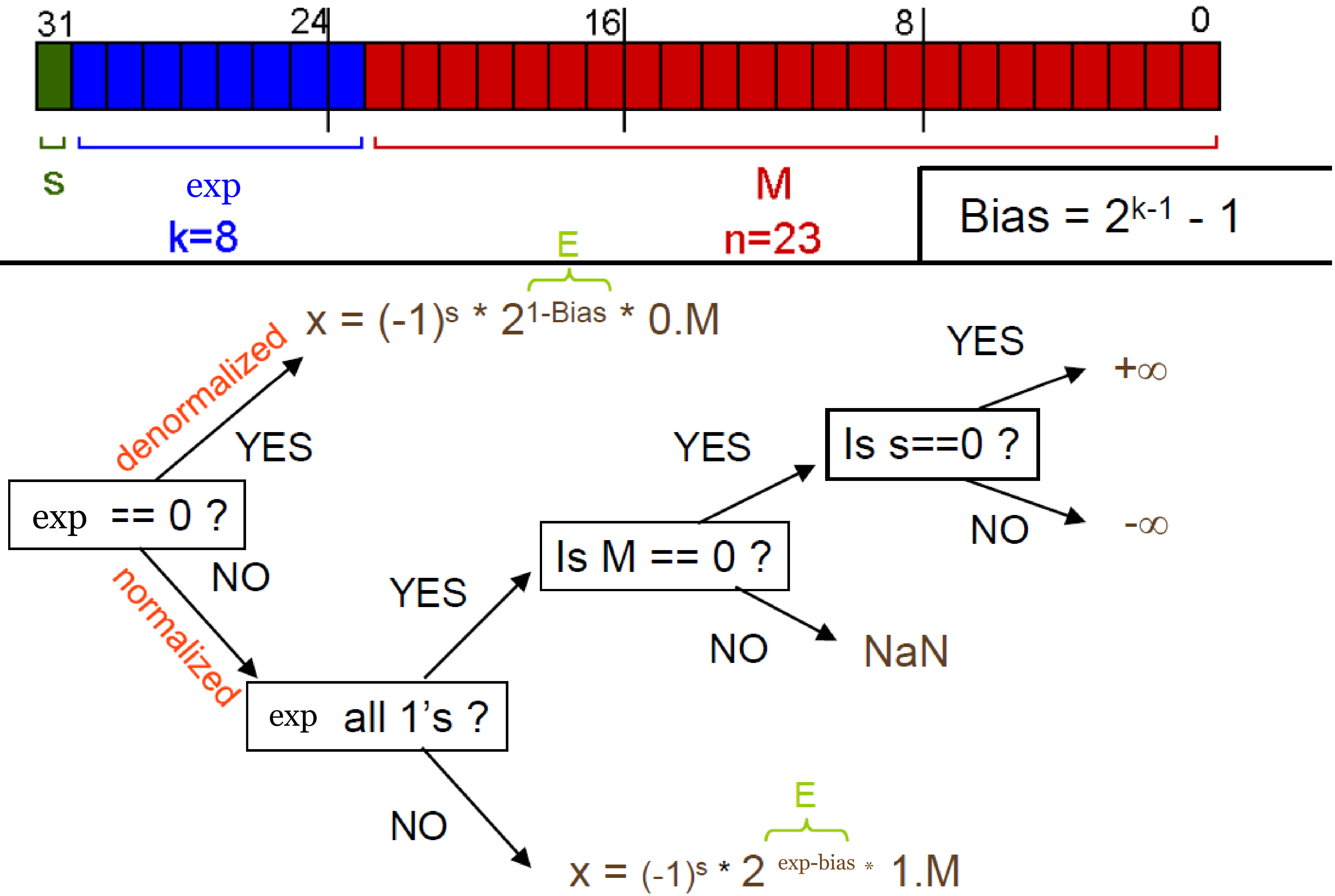
- Definition of NaN: exp = 0xFF and M>0
- The IEEE floating point standard gives both a positive representation of infinity and a negative one.
  - 0xFF800000 = Negative infinity.
  - 0x7F800000 = Positive infinity.

- Note the difference between infinity and NaN.

Round to Even

- How do we make rounding unbiased?
  - Answer: Round to nearest even number
- “Even” when least significant bit is 0
- “Half way” when bits to right of rounding position = 100…
- Only apply rule when number is halfway. Otherwise, closest number.

Floating point decision diagram



Example

3/8 = 1.1 * 2^-2

Bias = 2^(8-1) -1= 127

s = 0, e = -2 + Bias = 125, M = 0x400000

0	01111101	10000000000000000000000
S	Exponent	Mantissa

Old Exam Question

- **Problem 3. (12 points):**
- Consider the following two 8-bit floating point representations based on the IEEE floating point format.
- Neither has a sign bit—they can only represent nonnegative numbers.
- 1. Format A
- • There are k = 3 exponent bits. The exponent bias is 3.
- • There are n = 5 fraction bits.
- 2. Format B
- • There are k = 5 exponent bits. The exponent bias is 15.
- • There are n = 3 fraction bits.

Old Exam Question cont.

Format A		Format B	
Bits	Value	Bits	Value
`011 00000`	1	`01111 000`	1
			15
	$\frac{53}{16}$		
		`10100 110`	
`000 00001`			

Old Exam Question Answer

Format A		Format B	
Bits	Value	Bits	Value
011 00000	1	01111 000	1
110 11100	15	10010 111	15
100 10101	$\frac{53}{16}$	10000 101	13/4
111 00000	+ Inf	10100 110	56
000 00001	1/128	01000 000	1/128

Example Datalab puzzle

greatestBitPos - return a mask that marks the position of the most significant 1 bit. If x == 0, return 0

Example: greatestBitPos(96) = 0x40

Legal ops: ! ~ & ^ | + << >>

Max ops: 70   Rating: 4

greatestbitpos(int x)

Solve on blackboard…

Good Coding Style

- Consistent indentation
- Avoid long sequences of commands without a comment
- Each source file should have an appropriate header
- Have a brief comment at the beginning of each function