

Applied Discrete Mathematics
Floating Point Numbers

William Shoaff

Fall 2007

1 FLOATING POINT NUMBERS

Every well-rounded programmer ought to have a knowledge of what goes on during the elementary steps of floating point arithmetic.

Don Knuth, “The Art of Computer Programming:
Seminumerical Algorithms American computer
scientist (1938 –)

RATIONAL NUMBERS

RATIONAL NUMBERS

- Rational numbers are those that can be written as ratios of integers

$$\mathbb{Q} = \left\{ r : r = \frac{a}{b} : a \in \mathbb{Z}, b \in \mathbb{Z}, b \neq 0 \right\}$$

RATIONAL NUMBERS

- Rational numbers are those that can be written as ratios of integers

$$\mathbb{Q} = \left\{ r : r = \frac{a}{b} : a \in \mathbb{Z}, b \in \mathbb{Z}, b \neq 0 \right\}$$

- Rational numbers could be represented as a data structure of two signed integer

RATIONAL NUMBERS

- Rational numbers are those that can be written as ratios of integers

$$\mathbb{Q} = \left\{ r : r = \frac{a}{b} : a \in \mathbb{Z}, b \in \mathbb{Z}, b \neq 0 \right\}$$

- Rational numbers could be represented as a data structure of two signed integer

type Rational { int numerator; int denominator; }

RATIONAL NUMBERS

- Rational numbers are those that can be written as ratios of integers

$$\mathbb{Q} = \left\{ r : r = \frac{a}{b} : a \in \mathbb{Z}, b \in \mathbb{Z}, b \neq 0 \right\}$$

- Rational numbers could be represented as a data structure of two signed integer
 type Rational { int numerator; int denominator; }
- Some problems with this idea include

RATIONAL NUMBERS

- Rational numbers are those that can be written as ratios of integers

$$\mathbb{Q} = \left\{ r : r = \frac{a}{b} : a \in \mathbb{Z}, b \in \mathbb{Z}, b \neq 0 \right\}$$

- Rational numbers could be represented as a data structure of two signed integer
type Rational { int numerator; int denominator; }
- Some problems with this idea include
 - Multiple representations of the same rational

RATIONAL NUMBERS

- Rational numbers are those that can be written as ratios of integers

$$\mathbb{Q} = \left\{ r : r = \frac{a}{b} : a \in \mathbb{Z}, b \in \mathbb{Z}, b \neq 0 \right\}$$

- Rational numbers could be represented as a data structure of two signed integer

type Rational { int numerator; int denominator; }

- Some problems with this idea include
 - Multiple representations of the same rational
 - Restricted representational range: Using 64 bits

RATIONAL NUMBERS

- Rational numbers are those that can be written as ratios of integers

$$\mathbb{Q} = \left\{ r : r = \frac{a}{b} : a \in \mathbb{Z}, b \in \mathbb{Z}, b \neq 0 \right\}$$

- Rational numbers could be represented as a data structure of two signed integer

type Rational { int numerator; int denominator; }

- Some problems with this idea include
 - Multiple representations of the same rational
 - Restricted representational range: Using 64 bits
 - the largest positive rational is only $2^{63} - 1 \approx 9.2 \times 10^{18}$

RATIONAL NUMBERS

- Rational numbers are those that can be written as ratios of integers

$$\mathbb{Q} = \left\{ r : r = \frac{a}{b} : a \in \mathbb{Z}, b \in \mathbb{Z}, b \neq 0 \right\}$$

- Rational numbers could be represented as a data structure of two signed integer

type Rational { int numerator; int denominator; }

- Some problems with this idea include
 - Multiple representations of the same rational
 - Restricted representational range: Using 64 bits
 - the largest positive rational is only $2^{63} - 1 \approx 9.2 \times 10^{18}$
 - the smallest positive rational is only $1/(2^{63} - 1) \approx 1.1 \times 10^{-19}$

IEEE 754 FLOATING POINT NUMBERS

IEEE 754 FLOATING POINT NUMBERS

- Floating point numbers are a subset of rationals

$$\mathbb{F} \subset \mathbb{Q}$$

IEEE 754 FLOATING POINT NUMBERS

- Floating point numbers are a subset of rationals

$$\mathbb{F} \subset \mathbb{Q}$$

- The set of floating point numbers depends on the architecture of a computer

IEEE 754 FLOATING POINT NUMBERS

- Floating point numbers are a subset of rationals

$$\mathbb{F} \subset \mathbb{Q}$$

- The set of floating point numbers depends on the architecture of a computer
- The IEEE 754 Floating Point standard specifies expectations of computer architectures

IEEE 754 FLOATING POINT NUMBERS

- Floating point numbers are a subset of rationals

$$\mathbb{F} \subset \mathbb{Q}$$

- The set of floating point numbers depends on the architecture of a computer
- The IEEE 754 Floating Point standard specifies expectations of computer architectures
- A `float` number type is 32 bits

IEEE 754 FLOATING POINT NUMBERS

- Floating point numbers are a subset of rationals

$$\mathbb{F} \subset \mathbb{Q}$$

- The set of floating point numbers depends on the architecture of a computer
- The IEEE 754 Floating Point standard specifies expectations of computer architectures
- A `float` number type is 32 bits
- A `double` number type is 64 bits

IEEE 754 FLOATING POINT NUMBERS

- Floating point numbers are a subset of rationals

$$\mathbb{F} \subset \mathbb{Q}$$

- The set of floating point numbers depends on the architecture of a computer
- The IEEE 754 Floating Point standard specifies expectations of computer architectures
- A `float` number type is 32 bits
- A `double` number type is 64 bits
- IEEE 754 floating point has too many bits and other intricacies to describe fully

IEEE 754 FLOATING POINT NUMBERS

- Floating point numbers are a subset of rationals

$$\mathbb{F} \subset \mathbb{Q}$$

- The set of floating point numbers depends on the architecture of a computer
- The IEEE 754 Floating Point standard specifies expectations of computer architectures
- A `float` number type is 32 bits
- A `double` number type is 64 bits
- IEEE 754 floating point has too many bits and other intricacies to describe fully
- The basics of floating point numbers can be learned by studying more simple systems

DECIMAL SCIENTIFIC NOTATION

DECIMAL SCIENTIFIC NOTATION

- Scientific notation should be a familiar

DECIMAL SCIENTIFIC NOTATION

- Scientific notation should be a familiar
- Here's some examples as a brief review

Decimal Number	Scientific Notation
----------------	---------------------

DECIMAL SCIENTIFIC NOTATION

- Scientific notation should be a familiar
- Here's some examples as a brief review

Decimal Number	Scientific Notation
----------------	---------------------

DECIMAL SCIENTIFIC NOTATION

- Scientific notation should be a familiar
- Here's some examples as a brief review

Decimal Number	Scientific Notation
0.0000031415	

DECIMAL SCIENTIFIC NOTATION

- Scientific notation should be a familiar
- Here's some examples as a brief review

Decimal Number	Scientific Notation
0.0000031415	3.1415×10^{-6}

DECIMAL SCIENTIFIC NOTATION

- Scientific notation should be a familiar
- Here's some examples as a brief review

Decimal Number	Scientific Notation
0.0000031415	3.1415×10^{-6}
161803.399	

DECIMAL SCIENTIFIC NOTATION

- Scientific notation should be a familiar
- Here's some examples as a brief review

Decimal Number	Scientific Notation
0.0000031415	3.1415×10^{-6}
161803.399	1.61803399×10^5

DECIMAL SCIENTIFIC NOTATION

- Scientific notation should be a familiar
- Here's some examples as a brief review

Decimal Number	Scientific Notation
0.0000031415	3.1415×10^{-6}
161803.399	1.61803399×10^5
-0.0005772	

DECIMAL SCIENTIFIC NOTATION

- Scientific notation should be a familiar
- Here's some examples as a brief review

Decimal Number	Scientific Notation
0.0000031415	3.1415×10^{-6}
161803.399	1.61803399×10^5
-0.0005772	-5.772×10^{-4}

DECIMAL SCIENTIFIC NOTATION

- Scientific notation should be a familiar
- Here's some examples as a brief review

Decimal Number	Scientific Notation
0.0000031415	3.1415×10^{-6}
161803.399	1.61803399×10^5
-0.0005772	-5.772×10^{-4}
-2.71828183	

DECIMAL SCIENTIFIC NOTATION

- Scientific notation should be a familiar
- Here's some examples as a brief review

Decimal Number	Scientific Notation
0.0000031415	3.1415×10^{-6}
161803.399	1.61803399×10^5
-0.0005772	-5.772×10^{-4}
-2.71828183	-2.71828183×10^0

DECIMAL SCIENTIFIC NOTATION

- Scientific notation should be a familiar
- Here's some examples as a brief review

Decimal Number	Scientific Notation
0.0000031415	3.1415×10^{-6}
161803.399	1.61803399×10^5
-0.0005772	-5.772×10^{-4}
-2.71828183	-2.71828183×10^0
0.0060221415	6.0221415×10^{-3}

DECIMAL SCIENTIFIC NOTATION

- Scientific notation should be a familiar
- Here's some examples as a brief review

Decimal Number	Scientific Notation
0.0000031415	3.1415×10^{-6}
161803.399	1.61803399×10^5
-0.0005772	-5.772×10^{-4}
-2.71828183	-2.71828183×10^0
$0.0060221415 \times 10^{26}$	6.0221415×10^{23}

DECIMAL SCIENTIFIC NOTATION

DECIMAL SCIENTIFIC NOTATION

- When written in decimal scientific notation, a rational number r has the form

$$r = \pm(i.f) \times 10^e$$

where:

DECIMAL SCIENTIFIC NOTATION

- When written in decimal scientific notation, a rational number r has the form

$$r = \pm(i.f) \times 10^e$$

where:

- i is a non-zero digit (1, 2, 3, ..., 9)

DECIMAL SCIENTIFIC NOTATION

- When written in decimal scientific notation, a rational number r has the form

$$r = \pm(i.f) \times 10^e$$

where:

- i is a non-zero digit (1, 2, 3, ..., 9)
 - When i is non-zero are said to be **normalized**

DECIMAL SCIENTIFIC NOTATION

- When written in decimal scientific notation, a rational number r has the form

$$r = \pm(i.f) \times 10^e$$

where:

- i is a non-zero digit (1, 2, 3, ..., 9)
 - When i is non-zero are said to be **normalized**
- f is an integer representing a fractional part

DECIMAL SCIENTIFIC NOTATION

- When written in decimal scientific notation, a rational number r has the form

$$r = \pm(i.f) \times 10^e$$

where:

- i is a non-zero digit (1, 2, 3, ..., 9)
 - When i is non-zero are said to be **normalized**
- f is an integer representing a fractional part
- e is an integer power chosen such that

$$r = \pm(i.f) \times 10^e$$

SCIENTIFIC NOTATION IN BINARY

SCIENTIFIC NOTATION IN BINARY

- When 2 is the base, scientific notation is similar, but only bits and powers of 2 are used

Binary Number	Scientific Notation
---------------	---------------------

SCIENTIFIC NOTATION IN BINARY

- When 2 is the base, scientific notation is similar, but only bits and powers of 2 are used

Binary Number	Scientific Notation

SCIENTIFIC NOTATION IN BINARY

- When 2 is the base, scientific notation is similar, but only bits and powers of 2 are used

Binary Number	Scientific Notation
0.0000011011	

SCIENTIFIC NOTATION IN BINARY

- When 2 is the base, scientific notation is similar, but only bits and powers of 2 are used

Binary Number	Scientific Notation
0.0000011011	1.1011×2^{-6}

SCIENTIFIC NOTATION IN BINARY

- When 2 is the base, scientific notation is similar, but only bits and powers of 2 are used

Binary Number	Scientific Notation
0.0000011011	1.1011×2^{-6}
101000.011	

SCIENTIFIC NOTATION IN BINARY

- When 2 is the base, scientific notation is similar, but only bits and powers of 2 are used

Binary Number	Scientific Notation
0.0000011011	1.1011×2^{-6}
101000.011	1.01000011×2^5

SCIENTIFIC NOTATION IN BINARY

- When 2 is the base, scientific notation is similar, but only bits and powers of 2 are used

Binary Number	Scientific Notation
0.0000011011	1.1011×2^{-6}
101000.011	1.01000011×2^5
– 0.0001001	

SCIENTIFIC NOTATION IN BINARY

- When 2 is the base, scientific notation is similar, but only bits and powers of 2 are used

Binary Number	Scientific Notation
0.0000011011	1.1011×2^{-6}
101000.011	1.01000011×2^5
-0.0001001	-1.001×2^{-4}

SCIENTIFIC NOTATION IN BINARY

- When 2 is the base, scientific notation is similar, but only bits and powers of 2 are used

Binary Number	Scientific Notation
0.0000011011	1.1011×2^{-6}
101000.011	1.01000011×2^5
- 0.0001001	$- 1.001 \times 2^{-4}$
- 1.11001110	

SCIENTIFIC NOTATION IN BINARY

- When 2 is the base, scientific notation is similar, but only bits and powers of 2 are used

Binary Number	Scientific Notation
0.0000011011	1.1011×2^{-6}
101000.011	1.01000011×2^5
- 0.0001001	$- 1.001 \times 2^{-4}$
- 1.11001110	$- 1.11001110 \times 2^0$

SCIENTIFIC NOTATION IN BINARY

- When 2 is the base, scientific notation is similar, but only bits and powers of 2 are used

Binary Number	Scientific Notation
0.0000011011	1.1011×2^{-6}
101000.011	1.01000011×2^5
- 0.0001001	$- 1.001 \times 2^{-4}$
- 1.11001110	$- 1.11001110 \times 2^0$
$0.0010001111 \times 10^{26}$	

SCIENTIFIC NOTATION IN BINARY

- When 2 is the base, scientific notation is similar, but only bits and powers of 2 are used

Binary Number	Scientific Notation
0.0000011011	1.1011×2^{-6}
101000.011	1.01000011×2^5
- 0.0001001	$- 1.001 \times 2^{-4}$
- 1.11001110	$- 1.11001110 \times 2^0$
$0.0010001111 \times 10^{26}$	1.0001111×2^{23}

DECIMAL AND BINARY FRACTIONAL PARTS

DECIMAL AND BINARY FRACTIONAL PARTS

- Decimal and binary fractional parts are written in **positional** notation

DECIMAL AND BINARY FRACTIONAL PARTS

- Decimal and binary fractional parts are written in **positional** notation
- The decimal 0.278 means “2 tenths” plus “7 hundredths” plus “8 thousandths”

$$\begin{aligned} 0.278 &= \frac{2}{10} + \frac{7}{100} + \frac{8}{1000} \\ &= 2 \cdot 10^{-1} + 7 \cdot 10^{-2} + 8 \cdot 10^{-3} = \frac{278}{1000} \end{aligned}$$

DECIMAL AND BINARY FRACTIONAL PARTS

- Decimal and binary fractional parts are written in **positional** notation
- The decimal 0.278 means “2 tenths” plus “7 hundredths” plus “8 thousandths”

$$\begin{aligned}0.278 &= \frac{2}{10} + \frac{7}{100} + \frac{8}{1000} \\ &= 2 \cdot 10^{-1} + 7 \cdot 10^{-2} + 8 \cdot 10^{-3} = \frac{278}{1000}\end{aligned}$$

- The binary 0.1001 means “1 half” plus “0 quarters” plus “0 eighths” plus “1 sixteenth”

$$\begin{aligned}0.1001 &= \frac{1}{2} + \frac{0}{4} + \frac{0}{8} + \frac{1}{16} \\ &= 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 0 \cdot 2^{-3} + 1 \cdot 2^{-4} = \frac{9}{16}\end{aligned}$$

DECIMAL AND BINARY FRACTIONAL PARTS

DECIMAL AND BINARY FRACTIONAL PARTS

- Here are some examples fractions in binary and their equivalent decimal representations

BINARY	SUM	DECIMAL FRACTION
0.1000	$\frac{1}{2}$	8/16
0.1100	$\frac{1}{2} + \frac{1}{4}$	12/16
0.1010	$\frac{1}{2} + \frac{1}{8}$	10/16
0.1110	$\frac{1}{2} + \frac{1}{4} + \frac{1}{8}$	14/16
0.1001	$\frac{1}{2} + \frac{1}{16}$	9/16
0.1101	$\frac{1}{2} + \frac{1}{4} + \frac{1}{16}$	13/16
0.1011	$\frac{1}{2} + \frac{1}{8} + \frac{1}{16}$	11/16
0.1111	$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16}$	15/16

HORNER'S METHOD FOR RATIONAL NUMBERS

HORNER'S METHOD FOR RATIONAL NUMBERS

- Horner's method should be used to convert a binary number to decimal

HORNER'S METHOD FOR RATIONAL NUMBERS

- Horner's method should be used to convert a binary number to decimal
- First recall a decimal number such as 3.1415 means

$$\frac{31415}{10^4}$$

HORNER'S METHOD FOR RATIONAL NUMBERS

- Horner's method should be used to convert a binary number to decimal
- First recall a decimal number such as 3.1415 means

$$\frac{31415}{10^4}$$

- Similarly binary number such as 1.1011 means

$$\frac{11011}{2^4}$$

HORNER'S METHOD FOR RATIONAL NUMBERS

- Horner's method should be used to convert a binary number to decimal
- First recall a decimal number such as 3.1415 means

$$\frac{31415}{10^4}$$

- Similarly binary number such as 1.1011 means

$$\frac{11011}{2^4}$$

- To convert a binary number $i.f$ to decimal

HORNER'S METHOD FOR RATIONAL NUMBERS

- Horner's method should be used to convert a binary number to decimal
- First recall a decimal number such as 3.1415 means

$$\frac{31415}{10^4}$$

- Similarly binary number such as 1.1011 means

$$\frac{11011}{2^4}$$

- To convert a binary number $i.f$ to decimal
 - ❶ Ignore the binary point

HORNER'S METHOD FOR RATIONAL NUMBERS

- Horner's method should be used to convert a binary number to decimal
- First recall a decimal number such as 3.1415 means

$$\frac{31415}{10^4}$$

- Similarly binary number such as 1.1011 means

$$\frac{11011}{2^4}$$

- To convert a binary number $i.f$ to decimal
 - 1 Ignore the binary point
 - 2 Convert if to decimal by Horner's method

HORNER'S METHOD FOR RATIONAL NUMBERS

- Horner's method should be used to convert a binary number to decimal
- First recall a decimal number such as 3.1415 means

$$\frac{31415}{10^4}$$

- Similarly binary number such as 1.1011 means

$$\frac{11011}{2^4}$$

- To convert a binary number $i.f$ to decimal
 - 1 Ignore the binary point
 - 2 Convert if to decimal by Horner's method
 - 3 Divide the result by 2^n where n is the number of bits in f

HORNER'S METHOD FOR RATIONAL NUMBERS

HORNER'S METHOD FOR RATIONAL NUMBERS

- Convert the binary rational 1100.1001 to a decimal fraction

HORNER'S METHOD FOR RATIONAL NUMBERS

- Convert the binary rational 1100.1001 to a decimal fraction
- Ignoring the binary point, convert 11001001 to decimal by Horner's algorithm

1	1	0	0	1	0	0	1
	2	6	12	24	50	100	200
1	3	6	12	25	50	100	201

HORNER'S METHOD FOR RATIONAL NUMBERS

- Convert the binary rational 1100.1001 to a decimal fraction
- Ignoring the binary point, convert 11001001 to decimal by Horner's algorithm

1	1	0	0	1	0	0	1
	2	6	12	24	50	100	200
1	3	6	12	25	50	100	201

- The binary rational 1100.1001 equals

$$\frac{201}{2^4} = \frac{201}{16} = 12 + \frac{9}{16}$$

HORNER'S METHOD FOR RATIONAL NUMBERS

- Convert the binary rational 1100.1001 to a decimal fraction
- Ignoring the binary point, convert 11001001 to decimal by Horner's algorithm

1	1	0	0	1	0	0	1
	2	6	12	24	50	100	200
1	3	6	12	25	50	100	201

- The binary rational 1100.1001 equals

$$\frac{201}{2^4} = \frac{201}{16} = 12 + \frac{9}{16}$$

- Note: the string 1100 from 1100.1001 is 12

HORNER'S METHOD FOR RATIONAL NUMBERS

- Convert the binary rational 1100.1001 to a decimal fraction
- Ignoring the binary point, convert 11001001 to decimal by Horner's algorithm

1	1	0	0	1	0	0	1
	2	6	12	24	50	100	200
1	3	6	12	25	50	100	201

- The binary rational 1100.1001 equals

$$\frac{201}{2^4} = \frac{201}{16} = 12 + \frac{9}{16}$$

- Note: the string 1100 from 1100.1001 is 12
- Note: the string 1001 from 1100.1001 is 9

HORNER'S METHOD FOR RATIONAL NUMBERS

HORNER'S METHOD FOR RATIONAL NUMBERS

- Convert the binary rational 110.01001 to a decimal fraction

HORNER'S METHOD FOR RATIONAL NUMBERS

- Convert the binary rational 110.01001 to a decimal fraction
- This is the same example as in the previous slides, except the binary point has been moved

HORNER'S METHOD FOR RATIONAL NUMBERS

- Convert the binary rational 110.01001 to a decimal fraction
- This is the same example as in the previous slides, except the binary point has been moved
- Now, we'll divide the result by 2^5

1	1	0	0	1	0	0	1
	2	6	12	24	50	100	200
1	3	6	12	25	50	100	201

HORNER'S METHOD FOR RATIONAL NUMBERS

- Convert the binary rational 110.01001 to a decimal fraction
- This is the same example as in the previous slides, except the binary point has been moved
- Now, we'll divide the result by 2^5

1	1	0	0	1	0	0	1
	2	6	12	24	50	100	200
1	3	6	12	25	50	100	201

- The binary rational 110.01001 equals

$$\frac{201}{2^5} = \frac{201}{32} = 6 + \frac{9}{32}$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Recall to convert decimal integers to binary you repeatedly divide by 2 and collect remainders

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Recall to convert decimal integers to binary you repeatedly divide by 2 and collect remainders
- To convert fractions, repeatedly multiply by 2 extracting a bit on each multiplication

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Recall to convert decimal integers to binary you repeatedly divide by 2 and collect remainders
- To convert fractions, repeatedly multiply by 2 extracting a bit on each multiplication

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Recall to convert decimal integers to binary you repeatedly divide by 2 and collect remainders
- To convert fractions, repeatedly multiply by 2 extracting a bit on each multiplication

$$0.703125 \times 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Recall to convert decimal integers to binary you repeatedly divide by 2 and collect remainders
- To convert fractions, repeatedly multiply by 2 extracting a bit on each multiplication

$$0.703125 \times 2 = 1.40625 \quad (0.1 \dots)_2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Recall to convert decimal integers to binary you repeatedly divide by 2 and collect remainders
- To convert fractions, repeatedly multiply by 2 extracting a bit on each multiplication

$$0.703125 \times 2 = 1.40625 \quad (0.1 \dots)_2$$

$$0.40625 \times 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Recall to convert decimal integers to binary you repeatedly divide by 2 and collect remainders
- To convert fractions, repeatedly multiply by 2 extracting a bit on each multiplication

$$0.703125 \times 2 = 1.40625 \quad (0.1 \dots)_2$$

$$0.40625 \times 2 = 0.8125 \quad (0.10 \dots)_2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Recall to convert decimal integers to binary you repeatedly divide by 2 and collect remainders
- To convert fractions, repeatedly multiply by 2 extracting a bit on each multiplication

$$0.703125 \times 2 = 1.40625 \quad (0.1 \dots)_2$$

$$0.40625 \times 2 = 0.8125 \quad (0.10 \dots)_2$$

$$0.8125 \times 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Recall to convert decimal integers to binary you repeatedly divide by 2 and collect remainders
- To convert fractions, repeatedly multiply by 2 extracting a bit on each multiplication

$$0.703125 \times 2 = 1.40625 \quad (0.1 \dots)_2$$

$$0.40625 \times 2 = 0.8125 \quad (0.10 \dots)_2$$

$$0.8125 \times 2 = 1.625 \quad (0.101 \dots)_2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Recall to convert decimal integers to binary you repeatedly divide by 2 and collect remainders
- To convert fractions, repeatedly multiply by 2 extracting a bit on each multiplication

$$0.703125 \times 2 = 1.40625 \quad (0.1 \dots)_2$$

$$0.40625 \times 2 = 0.8125 \quad (0.10 \dots)_2$$

$$0.8125 \times 2 = 1.625 \quad (0.101 \dots)_2$$

$$0.625 \times 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Recall to convert decimal integers to binary you repeatedly divide by 2 and collect remainders
- To convert fractions, repeatedly multiply by 2 extracting a bit on each multiplication

$$0.703125 \times 2 = 1.40625 \quad (0.1\dots)_2$$

$$0.40625 \times 2 = 0.8125 \quad (0.10\dots)_2$$

$$0.8125 \times 2 = 1.625 \quad (0.101\dots)_2$$

$$0.625 \times 2 = 1.25 \quad (0.1011\dots)_2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Recall to convert decimal integers to binary you repeatedly divide by 2 and collect remainders
- To convert fractions, repeatedly multiply by 2 extracting a bit on each multiplication

$$0.703125 \times 2 = 1.40625 \quad (0.1 \dots)_2$$

$$0.40625 \times 2 = 0.8125 \quad (0.10 \dots)_2$$

$$0.8125 \times 2 = 1.625 \quad (0.101 \dots)_2$$

$$0.625 \times 2 = 1.25 \quad (0.1011 \dots)_2$$

$$0.25 \times 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Recall to convert decimal integers to binary you repeatedly divide by 2 and collect remainders
- To convert fractions, repeatedly multiply by 2 extracting a bit on each multiplication

$$0.703125 \times 2 = 1.40625 \quad (0.1\dots)_2$$

$$0.40625 \times 2 = 0.8125 \quad (0.10\dots)_2$$

$$0.8125 \times 2 = 1.625 \quad (0.101\dots)_2$$

$$0.625 \times 2 = 1.25 \quad (0.1011\dots)_2$$

$$0.25 \times 2 = 0.5 \quad (0.10110\dots)_2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Recall to convert decimal integers to binary you repeatedly divide by 2 and collect remainders
- To convert fractions, repeatedly multiply by 2 extracting a bit on each multiplication

$$0.703125 \times 2 = 1.40625 \quad (0.1 \dots)_2$$

$$0.40625 \times 2 = 0.8125 \quad (0.10 \dots)_2$$

$$0.8125 \times 2 = 1.625 \quad (0.101 \dots)_2$$

$$0.625 \times 2 = 1.25 \quad (0.1011 \dots)_2$$

$$0.25 \times 2 = 0.5 \quad (0.10110 \dots)_2$$

$$0.5 \times 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Recall to convert decimal integers to binary you repeatedly divide by 2 and collect remainders
- To convert fractions, repeatedly multiply by 2 extracting a bit on each multiplication

$$0.703125 \times 2 = 1.40625 \quad (0.1 \dots)_2$$

$$0.40625 \times 2 = 0.8125 \quad (0.10 \dots)_2$$

$$0.8125 \times 2 = 1.625 \quad (0.101 \dots)_2$$

$$0.625 \times 2 = 1.25 \quad (0.1011 \dots)_2$$

$$0.25 \times 2 = 0.5 \quad (0.10110 \dots)_2$$

$$0.5 \times 2 = 1.0 \quad (0.101101)_2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Notice, writing these equations in reverse order, we have

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Notice, writing these equations in reverse order, we have

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Notice, writing these equations in reverse order, we have

1.0

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Notice, writing these equations in reverse order, we have

$$1.0 = 0.5 \times 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Notice, writing these equations in reverse order, we have

$$\begin{aligned} 1.0 &= 0.5 \times 2 \\ &= (0.25 \times 2) \times 2 \end{aligned}$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Notice, writing these equations in reverse order, we have

$$\begin{aligned}1.0 &= 0.5 \times 2 \\ &= (0.25 \times 2) \times 2 \\ &= ((0.625 \times 2 - 1) \times 2) \times 2\end{aligned}$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Notice, writing these equations in reverse order, we have

$$1.0 = 0.5 \times 2$$

$$= (0.25 \times 2) \times 2$$

$$= ((0.625 \times 2 - 1) \times 2) \times 2$$

$$= (((0.8125 \times 2 - 1) \times 2 - 1) \times 2) \times 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Notice, writing these equations in reverse order, we have

$$1.0 = 0.5 \times 2$$

$$= (0.25 \times 2) \times 2$$

$$= ((0.625 \times 2 - 1) \times 2) \times 2$$

$$= (((0.8125 \times 2 - 1) \times 2 - 1) \times 2) \times 2$$

$$= (((((0.40625 \times 2) \times 2 - 1) \times 2 - 1) \times 2) \times 2) \times 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Notice, writing these equations in reverse order, we have

$$1.0 = 0.5 \times 2$$

$$= (0.25 \times 2) \times 2$$

$$= ((0.625 \times 2 - 1) \times 2) \times 2$$

$$= (((0.8125 \times 2 - 1) \times 2 - 1) \times 2) \times 2$$

$$= (((((0.40625 \times 2) \times 2 - 1) \times 2 - 1) \times 2) \times 2) \times 2$$

$$= (((((((0.703125 \times 2 - 1) \times 2) \times 2 - 1) \times 2 - 1) \times 2 - 1) \times 2) \times 2) \times 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- These calculations can then be unrolled as a binary fraction.

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- These calculations can then be unrolled as a binary fraction.

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- These calculations can then be unrolled as a binary fraction.

$$\frac{1}{2}$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- These calculations can then be unrolled as a binary fraction.

$$\frac{1}{2} = (((((0.703125 \cdot 2 - 1) \cdot 2) \cdot 2 - 1) \cdot 2 - 1) \cdot 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- These calculations can then be unrolled as a binary fraction.

$$1 + \frac{1}{2^2} = \frac{1}{2} = (((((0.703125 \cdot 2 - 1) \cdot 2) \cdot 2 - 1) \cdot 2 - 1) \cdot 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- These calculations can then be unrolled as a binary fraction.

$$\frac{1}{2} = (((((0.703125 \cdot 2 - 1) \cdot 2) \cdot 2 - 1) \cdot 2 - 1) \cdot 2 - 1) \cdot 2$$
$$1 + \frac{1}{2^2} = (((((0.703125 \cdot 2 - 1) \cdot 2) \cdot 2 - 1) \cdot 2 - 1) \cdot 2 - 1) \cdot 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- These calculations can then be unrolled as a binary fraction.

$$\frac{1}{2} = (((((0.703125 \cdot 2 - 1) \cdot 2) \cdot 2 - 1) \cdot 2 - 1) \cdot 2$$

$$1 + \frac{1}{2^2} = (((0.703125 \cdot 2 - 1) \cdot 2) \cdot 2 - 1) \cdot 2$$

$$1 + \frac{1}{2} + \frac{1}{2^3}$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- These calculations can then be unrolled as a binary fraction.

$$\frac{1}{2} = (((((0.703125 \cdot 2 - 1) \cdot 2) \cdot 2 - 1) \cdot 2 - 1) \cdot 2$$

$$1 + \frac{1}{2^2} = (((0.703125 \cdot 2 - 1) \cdot 2) \cdot 2 - 1) \cdot 2$$

$$1 + \frac{1}{2} + \frac{1}{2^3} = (0.703125 \cdot 2 - 1) \cdot 2 \cdot 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- These calculations can then be unrolled as a binary fraction.

$$\frac{1}{2} = (((((0.703125 \cdot 2 - 1) \cdot 2) \cdot 2 - 1) \cdot 2 - 1) \cdot 2$$

$$1 + \frac{1}{2^2} = (((0.703125 \cdot 2 - 1) \cdot 2) \cdot 2 - 1) \cdot 2$$

$$1 + \frac{1}{2} + \frac{1}{2^3} = (0.703125 \cdot 2 - 1) \cdot 2 \cdot 2$$

$$1 + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^5}$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- These calculations can then be unrolled as a binary fraction.

$$\frac{1}{2} = (((((0.703125 \cdot 2 - 1) \cdot 2) \cdot 2 - 1) \cdot 2 - 1) \cdot 2$$

$$1 + \frac{1}{2^2} = (((0.703125 \cdot 2 - 1) \cdot 2) \cdot 2 - 1) \cdot 2$$

$$1 + \frac{1}{2} + \frac{1}{2^3} = (0.703125 \cdot 2 - 1) \cdot 2 \cdot 2$$

$$1 + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^5} = 0.703125 \cdot 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- These calculations can then be unrolled as a binary fraction.

$$\frac{1}{2} = (((((0.703125 \cdot 2 - 1) \cdot 2) \cdot 2 - 1) \cdot 2 - 1) \cdot 2$$

$$1 + \frac{1}{2^2} = (((0.703125 \cdot 2 - 1) \cdot 2) \cdot 2 - 1) \cdot 2$$

$$1 + \frac{1}{2} + \frac{1}{2^3} = (0.703125 \cdot 2 - 1) \cdot 2 \cdot 2$$

$$1 + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^5} = 0.703125 \cdot 2$$

$$\frac{1}{2} + \frac{1}{2^3} + \frac{1}{2^4} + \frac{1}{2^6}$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- These calculations can then be unrolled as a binary fraction.

$$\frac{1}{2} = (((((0.703125 \cdot 2 - 1) \cdot 2) \cdot 2 - 1) \cdot 2 - 1) \cdot 2$$

$$1 + \frac{1}{2^2} = (((0.703125 \cdot 2 - 1) \cdot 2) \cdot 2 - 1) \cdot 2$$

$$1 + \frac{1}{2} + \frac{1}{2^3} = (0.703125 \cdot 2 - 1) \cdot 2 \cdot 2$$

$$1 + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^5} = 0.703125 \cdot 2$$

$$\frac{1}{2} + \frac{1}{2^3} + \frac{1}{2^4} + \frac{1}{2^6} = 0.703125$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Convert $1/3 = 0.3333\dots$ to binary

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Convert $1/3 = 0.3333\dots$ to binary

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Convert $1/3 = 0.3333\dots$ to binary

$$0.3333\dots \times 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Convert $1/3 = 0.3333\dots$ to binary

$$0.3333\dots \times 2 = 0.6666\dots \quad (0.0\dots)_2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Convert $1/3 = 0.3333\dots$ to binary

$$0.3333\dots \times 2 = 0.6666\dots \quad (0.0\dots)_2$$

$$0.6666\dots \times 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Convert $1/3 = 0.3333\dots$ to binary

$$0.3333\dots \times 2 = 0.6666\dots \quad (0.0\dots)_2$$

$$0.6666\dots \times 2 = 1.3333\dots \quad (0.01\dots)_2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Convert $1/3 = 0.3333\dots$ to binary

$$0.3333\dots \times 2 = 0.6666\dots \quad (0.0\dots)_2$$

$$0.6666\dots \times 2 = 1.3333\dots \quad (0.01\dots)_2$$

$$0.3333\dots \times 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Convert $1/3 = 0.3333\dots$ to binary

$$0.3333\dots \times 2 = 0.6666\dots \quad (0.0\dots)_2$$

$$0.6666\dots \times 2 = 1.3333\dots \quad (0.01\dots)_2$$

$$0.3333\dots \times 2 = 0.6666\dots \quad (0.010\dots)_2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Convert $1/3 = 0.3333\dots$ to binary

$$0.3333\dots \times 2 = 0.6666\dots \quad (0.0\dots)_2$$

$$0.6666\dots \times 2 = 1.3333\dots \quad (0.01\dots)_2$$

$$0.3333\dots \times 2 = 0.6666\dots \quad (0.010\dots)_2$$

$$0.6666\dots \times 2$$

CONVERSION OF DECIMAL FRACTIONS INTO BINARY

- Convert $1/3 = 0.3333\dots$ to binary

$$0.3333\dots \times 2 = 0.6666\dots \quad (0.0\dots)_2$$

$$0.6666\dots \times 2 = 1.3333\dots \quad (0.01\dots)_2$$

$$0.3333\dots \times 2 = 0.6666\dots \quad (0.010\dots)_2$$

$$0.6666\dots \times 2 = 1.3333\dots \quad (0.0101\dots)_2$$

- Therefore

$$\frac{1}{3} = (0.3333\dots)_{10} = (0.0101\dots)_2$$

SCIENTIFIC NOTATION IN BINARY

SCIENTIFIC NOTATION IN BINARY

- When written in binary scientific notation a rational number r has the form

$$r = \pm(1.f) \cdot 2^e$$

where:

SCIENTIFIC NOTATION IN BINARY

- When written in binary scientific notation a rational number r has the form

$$r = \pm(1.f) \cdot 2^e$$

where:

- f is an integer representing a fractional part

SCIENTIFIC NOTATION IN BINARY

- When written in binary scientific notation a rational number r has the form

$$r = \pm(1.f) \cdot 2^e$$

where:

- f is an integer representing a fractional part
- e is an integer power chosen such that

$$r = \pm(1.f) \times 2^e$$

SCIENTIFIC NOTATION IN BINARY

- When written in binary scientific notation a rational number r has the form

$$r = \pm(1.f) \cdot 2^e$$

where:

- f is an integer representing a fractional part
- e is an integer power chosen such that

$$r = \pm(1.f) \times 2^e$$

- Numbers of the form $1.f$ are said to be **normalized**

SCIENTIFIC NOTATION IN BINARY

- When written in binary scientific notation a rational number r has the form

$$r = \pm(1.f) \cdot 2^e$$

where:

- f is an integer representing a fractional part
- e is an integer power chosen such that

$$r = \pm(1.f) \times 2^e$$

- Numbers of the form $1.f$ are said to be **normalized**
- The 1 in $(1.f)$ is implicit and not explicitly written

BINARY FLOATING POINT NUMBERS

BINARY FLOATING POINT NUMBERS

- Floating point numbers are represented as a string of bits

BINARY FLOATING POINT NUMBERS

- Floating point numbers are represented as a string of bits
- The scheme used to represent a rational r is

BINARY FLOATING POINT NUMBERS

- Floating point numbers are represented as a string of bits
- The scheme used to represent a rational r is
 - A sign bit s , the leftmost, determines if $r < 0$ or $r \geq 0$

BINARY FLOATING POINT NUMBERS

- Floating point numbers are represented as a string of bits
- The scheme used to represent a rational r is
 - A sign bit s , the leftmost, determines if $r < 0$ or $r \geq 0$
 - Followed by n bits that represent an exponent e

BINARY FLOATING POINT NUMBERS

- Floating point numbers are represented as a string of bits
- The scheme used to represent a rational r is
 - A sign bit s , the leftmost, determines if $r < 0$ or $r \geq 0$
 - Followed by n bits that represent an exponent e
 - With the final m bits representing a fractional part f

BINARY FLOATING POINT NUMBERS

- Floating point numbers are represented as a string of bits
- The scheme used to represent a rational r is
 - A sign bit s , the leftmost, determines if $r < 0$ or $r \geq 0$
 - Followed by n bits that represent an exponent e
 - With the final m bits representing a fractional part f
- That is, a floating point number is parsed as

1 sign bit	n exponent bits	m fractional part bits
s	e	f

BINARY FLOATING POINT NUMBERS

BINARY FLOATING POINT NUMBERS

- Let's interpret the decimal meaning of the binary floating point number

BINARY FLOATING POINT NUMBERS

- Let's interpret the decimal meaning of the binary floating point number

0

111

1101

- The sign bit is $s = 0$

BINARY FLOATING POINT NUMBERS

- Let's interpret the decimal meaning of the binary floating point number

0

111

1101

- The sign bit is $s = 0$
 - A sign bit of 0 means the rational number is **positive**

BINARY FLOATING POINT NUMBERS

- Let's interpret the decimal meaning of the binary floating point number

0

111

1101

- The sign bit is $s = 0$
 - A sign bit of 0 means the rational number is **positive**
 - A sign bit of 1 means the rational number is **negative**

BINARY FLOATING POINT NUMBERS

- Let's interpret the decimal meaning of the binary floating point number

0

111

1101

- The sign bit is $s = 0$
 - A sign bit of 0 means the rational number is **positive**
 - A sign bit of 1 means the rational number is **negative**
- The fraction part is 1101

BINARY FLOATING POINT NUMBERS

- Let's interpret the decimal meaning of the binary floating point number

0

111

1101

- The sign bit is $s = 0$
 - A sign bit of 0 means the rational number is **positive**
 - A sign bit of 1 means the rational number is **negative**
- The fraction part is 1101
- Prepending the implicit 1, evaluate 1.1101 by Horner's method

1	1	1	0	1	
	2	6	14	28	
1	3	7	14	<table border="1"><tr><td>29</td></tr></table>	29
29					

$$1.1101 = \frac{29}{16} = 1 + \frac{13}{16}$$

BINARY FLOATING POINT NUMBERS

BINARY FLOATING POINT NUMBERS

- Binary floating point 0 111 1101 represents

$$+\frac{29}{16} \times 2^e = +\frac{29}{16} \times 2^{11}$$

BINARY FLOATING POINT NUMBERS

- Binary floating point 0 111 1101 represents

$$+\frac{29}{16} \times 2^e = +\frac{29}{16} \times 2^{111}$$

- The three bit exponent 111 needs to be interpreted as a decimal e

BINARY FLOATING POINT NUMBERS

- Binary floating point 0 111 1101 represents

$$+\frac{29}{16} \times 2^e = +\frac{29}{16} \times 2^{111}$$

- The three bit exponent 111 needs to be interpreted as a decimal e
- Floating point exponents are written in **biased** notation

BIASED NOTATION

BIASED NOTATION

- Biased notation presents an alternative to two's complement for representing integers (negative, zero, and positive whole numbers)

BIASED NOTATION

- Biased notation presents an alternative to two's complement for representing integers (negative, zero, and positive whole numbers)
- The idea is to use the natural numbers in the range

$$0 \leq k \leq N = \{0, 1, 2, \dots, N-1, N\}$$

to represent the range

$$-b \leq k \leq N-b = \{-b, 1-b, 2-b, \dots, +N-b-1, N-b\}$$

BIASED NOTATION

- Biased notation presents an alternative to two's complement for representing integers (negative, zero, and positive whole numbers)
- The idea is to use the natural numbers in the range

$$0 \leq k \leq N = \{0, 1, 2, \dots, N-1, N\}$$

to represent the range

$$-b \leq k \leq N-b = \{-b, 1-b, 2-b, \dots, +N-b-1, N-b\}$$

- That is, the number k stands for the value $k-b$

BIASED NOTATION

- Biased notation presents an alternative to two's complement for representing integers (negative, zero, and positive whole numbers)
- The idea is to use the natural numbers in the range

$$0 \leq k \leq N = \{0, 1, 2, \dots, N-1, N\}$$

to represent the range

$$-b \leq k \leq N-b = \{-b, 1-b, 2-b, \dots, +N-b-1, N-b\}$$

- That is, the number k stands for the value $k - b$
- The number b is called the **bias**

DECIMAL BIASED NOTATION

DECIMAL BIASED NOTATION

- In decimal, pretend that 3 digits are used to represent an exponent

DECIMAL BIASED NOTATION

- In decimal, pretend that 3 digits are used to represent an exponent
- That is, decimal strings from 000 to 999 encode an exponent

DECIMAL BIASED NOTATION

- In decimal, pretend that 3 digits are used to represent an exponent
- That is, decimal strings from 000 to 999 encode an exponent
- Select a **bias** b near the middle of this range, say $b = 500$

DECIMAL BIASED NOTATION

- In decimal, pretend that 3 digits are used to represent an exponent
- That is, decimal strings from 000 to 999 encode an exponent
- Select a **bias** b near the middle of this range, say $b = 500$
- Let $000 \leq k \leq 999$ and declare that k represents the value $k - b = k - 500$

DECIMAL BIASED NOTATION

DECIMAL BIASED NOTATION

- For example

Biased Representation: k	"Real" Value: $k - 500$
000	-500
007	-493
200	-300
454	-46
700	200
836	336
999	499

BINARY BIASED NOTATION

BINARY BIASED NOTATION

- In binary, pretend that 3 bits can be used to represent an exponent

BINARY BIASED NOTATION

- In binary, pretend that 3 bits can be used to represent an exponent
- That is, binary strings from 000 to 111 encode an exponent

BINARY BIASED NOTATION

- In binary, pretend that 3 bits can be used to represent an exponent
- That is, binary strings from 000 to 111 encode an exponent
- Select a **bias** b near the middle of this range, say
 $b = 4 = 100$

BINARY BIASED NOTATION

- In binary, pretend that 3 bits can be used to represent an exponent
- That is, binary strings from 000 to 111 encode an exponent
- Select a **bias** b near the middle of this range, say
 $b = 4 = 100$
- Let $000 \leq k \leq 111$ and declare that k represents the value
 $k - b = k - 4$

BINARY BIASED NOTATION

BINARY BIASED NOTATION

- For example

Biased Representation: k	"Real" Value: $k - 4$
$(000)_2 = 0$	-4
$(001)_2 = 1$	-3
$(010)_2 = 2$	-2
$(011)_2 = 3$	-1
$(100)_2 = 4$	0
$(101)_2 = 5$	1
$(110)_2 = 6$	2
$(111)_2 = 7$	3

COMPLETING THE EXAMPLE

COMPLETING THE EXAMPLE

- Recall, the binary floating point 0 111 1101 represents

$$+\frac{29}{16} \times 2^p = +\frac{29}{16} \times 2^{11}$$

COMPLETING THE EXAMPLE

- Recall, the binary floating point 0 111 1101 represents

$$+\frac{29}{16} \times 2^p = +\frac{29}{16} \times 2^{11}$$

- The three bit exponent 111 is a biased number with bias $b = 4$

COMPLETING THE EXAMPLE

- Recall, the binary floating point 0 111 1101 represents

$$+\frac{29}{16} \times 2^p = +\frac{29}{16} \times 2^{11}$$

- The three bit exponent 111 is a biased number with bias $b = 4$
- The exponent 111 is interpreted as representing

$$(111)_2 - 4 = 7 - 4 = 3$$

COMPLETING THE EXAMPLE

- Recall, the binary floating point 0 111 1101 represents

$$+\frac{29}{16} \times 2^p = +\frac{29}{16} \times 2^{111}$$

- The three bit exponent 111 is a biased number with bias $b = 4$
- The exponent 111 is interpreted as representing

$$(111)_2 - 4 = 7 - 4 = 3$$

- Therefore, the binary floating point 0 111 1101 represents

$$+\frac{29}{16} \times 2^3 = +\frac{29}{2}$$

ANOTHER EXAMPLE

ANOTHER EXAMPLE

- Interpret the binary floating point 1 010 0100 as a decimal rational number r

ANOTHER EXAMPLE

- Interpret the binary floating point 1 010 0100 as a decimal rational number r
- The sign bit is $s = 1$

ANOTHER EXAMPLE

- Interpret the binary floating point 1 010 0100 as a decimal rational number r
- The sign bit is $s = 1$
 - The rational number r is negative

ANOTHER EXAMPLE

- Interpret the binary floating point 1 010 0100 as a decimal rational number r
- The sign bit is $s = 1$
 - The rational number r is negative
- The exponent is $e = 010$

ANOTHER EXAMPLE

- Interpret the binary floating point 1 010 0100 as a decimal rational number r
- The sign bit is $s = 1$
 - The rational number r is negative
- The exponent is $e = 010$
 - With a bias of $b = 4$, the “real” value of e is

$$e = (010)_2 - 4 = 2 - 4 = -2$$

ANOTHER EXAMPLE

ANOTHER EXAMPLE

- The fractional part is $f = 0100$

ANOTHER EXAMPLE

- The fractional part is $f = 0100$
 - Appending a leading 1 and evaluating 1.0100 yields

1	0	1	0	0
	2	4	10	20
1	2	5	10	20

$$1.0100 = \frac{20}{16} = \frac{5}{4} = 1 + \frac{1}{4}$$

ANOTHER EXAMPLE

- The fractional part is $f = 0100$
 - Appending a leading 1 and evaluating 1.0100 yields

1	0	1	0	0
	2	4	10	20
1	2	5	10	20

$$1.0100 = \frac{20}{16} = \frac{5}{4} = 1 + \frac{1}{4}$$

- Therefore, the binary floating point 1 010 0100 represents

$$-\frac{5}{4} \times 2^{-2} = -\frac{5}{16}$$

AN EXAMPLE WITH 16 BITS

AN EXAMPLE WITH 16 BITS

- What is the decimal value of the binary floating point

1 010010 100110101

AN EXAMPLE WITH 16 BITS

- What is the decimal value of the binary floating point

1 010010 100110101

- The sign bit is $s = 1$

AN EXAMPLE WITH 16 BITS

- What is the decimal value of the binary floating point

1 010010 100110101

- The sign bit is $s = 1$
 - The decimal number is negative

AN EXAMPLE WITH 16 BITS

- What is the decimal value of the binary floating point

1 010010 100110101

- The sign bit is $s = 1$
 - The decimal number is negative
- The exponent is $e = 010010$, 6 bits long

AN EXAMPLE WITH 16 BITS

- What is the decimal value of the binary floating point

1 010010 100110101

- The sign bit is $s = 1$
 - The decimal number is negative
- The exponent is $e = 010010$, 6 bits long
 - With 6 bits, the range of unsigned integers is from 0 to $2^6 - 1 = 63$

AN EXAMPLE WITH 16 BITS

- What is the decimal value of the binary floating point

1 010010 100110101

- The sign bit is $s = 1$
 - The decimal number is negative
- The exponent is $e = 010010$, 6 bits long
 - With 6 bits, the range of unsigned integers is from 0 to $2^6 - 1 = 63$
 - Choose a bias in the middle of this range, say $b = 32$

AN EXAMPLE WITH 16 BITS

- What is the decimal value of the binary floating point

1 010010 100110101

- The sign bit is $s = 1$
 - The decimal number is negative
- The exponent is $e = 010010$, 6 bits long
 - With 6 bits, the range of unsigned integers is from 0 to $2^6 - 1 = 63$
 - Choose a bias in the middle of this range, say $b = 32$
 - With a bias of $b = 32$, the “real” value of e is

$$e = (010010)_2 - 32 = 18 - 32 = -14$$

AN EXAMPLE WITH 16 BITS

AN EXAMPLE WITH 16 BITS

- The fractional part is $f = 100110101$

AN EXAMPLE WITH 16 BITS

- The fractional part is $f = 100110101$
 - Appending a leading 1 and evaluating 1.100110101 yields

1	1	0	0	1	1	0	1	0	1
	2	6	12	24	50	102	204	410	820
1	3	6	12	25	51	102	205	410	821

$$1.100110101 = \frac{821}{512} = 1 + \frac{309}{512}$$

AN EXAMPLE WITH 16 BITS

- The fractional part is $f = 100110101$
 - Appending a leading 1 and evaluating 1.100110101 yields

1	1	0	0	1	1	0	1	0	1
	2	6	12	24	50	102	204	410	820
1	3	6	12	25	51	102	205	410	821

$$1.100110101 = \frac{821}{512} = 1 + \frac{309}{512}$$

- Therefore, the binary floating point $1\ 010010\ 100110101$ represents

$$-\frac{821}{512} \times 2^{-14} = -\frac{821}{2^{23}}$$

NORMALIZED AND DENORMALIZED FLOATING POINTS

NORMALIZED AND DENORMALIZED FLOATING POINTS

- A problem with our scheme is that we can not represent 0

NORMALIZED AND DENORMALIZED FLOATING POINTS

- A problem with our scheme is that we can not represent 0
- Normalized floating point numbers always have the form

$$r = \pm(1.f) \times 2^e$$

NORMALIZED AND DENORMALIZED FLOATING POINTS

- A problem with our scheme is that we can not represent 0
- Normalized floating point numbers always have the form

$$r = \pm(1.f) \times 2^e$$

- With the leading 1 in $1.f$ they can never equal 0

NORMALIZED AND DENORMALIZED FLOATING POINTS

- A problem with our scheme is that we can not represent 0
- Normalized floating point numbers always have the form

$$r = \pm(1.f) \times 2^e$$

- With the leading 1 in $1.f$ they can never equal 0
- A **denormalized** floating point has the form

$$r = \pm(0.f) \times 2^e$$

NORMALIZED AND DENORMALIZED FLOATING POINTS

- A problem with our scheme is that we can not represent 0
- Normalized floating point numbers always have the form

$$r = \pm(1.f) \times 2^e$$

- With the leading 1 in $1.f$ they can never equal 0
- A **denormalized** floating point has the form

$$r = \pm(0.f) \times 2^e$$

- There are reasons to allow denormalized floats

NORMALIZED AND DENORMALIZED FLOATING POINTS

- A problem with our scheme is that we can not represent 0
- Normalized floating point numbers always have the form

$$r = \pm(1.f) \times 2^e$$

- With the leading 1 in $1.f$ they can never equal 0
- A **denormalized** floating point has the form

$$r = \pm(0.f) \times 2^e$$

- There are reasons to allow denormalized floats
- We will declare the zero string 0000 0000... to be denormalized and represents 0

NORMALIZED AND DENORMALIZED FLOATING POINTS

- A problem with our scheme is that we can not represent 0
- Normalized floating point numbers always have the form

$$r = \pm(1.f) \times 2^e$$

- With the leading 1 in $1.f$ they can never equal 0
- A **denormalized** floating point has the form

$$r = \pm(0.f) \times 2^e$$

- There are reasons to allow denormalized floats
- We will declare the zero string 0000 0000... to be denormalized and represents 0
- All other strings are bits are normalized floats

DISTRIBUTION OF FLOATING POINT NUMBERS

DISTRIBUTION OF FLOATING POINT NUMBERS

- It is insightful to see how floating point numbers are distributed along a number line

DISTRIBUTION OF FLOATING POINT NUMBERS

- It is insightful to see how floating point numbers are distributed along a number line
- Consider the non-negative floats that can be represented using 9 bits with 1 sign bit, 3 exponent bits, and 5 fractional part bits

DISTRIBUTION OF FLOATING POINT NUMBERS

- It is insightful to see how floating point numbers are distributed along a number line
- Consider the non-negative floats that can be represented using 9 bits with 1 sign bit, 3 exponent bits, and 5 fractional part bits
- The biased exponents range from 000 to 111, that is from 0 to 7.

DISTRIBUTION OF FLOATING POINT NUMBERS

- It is insightful to see how floating point numbers are distributed along a number line
- Consider the non-negative floats that can be represented using 9 bits with 1 sign bit, 3 exponent bits, and 5 fractional part bits
- The biased exponents range from 000 to 111, that is from 0 to 7.
- Using a bias of $b = 4$ the “real” range is from -4 to 3

DISTRIBUTION OF FLOATING POINT NUMBERS

- It is insightful to see how floating point numbers are distributed along a number line
- Consider the non-negative floats that can be represented using 9 bits with 1 sign bit, 3 exponent bits, and 5 fractional part bits
- The biased exponents range from 000 to 111, that is from 0 to 7.
- Using a bias of $b = 4$ the “real” range is from -4 to 3
- The normalized values $1.f$ range from 1.00000 to 1.11111

DISTRIBUTION OF FLOATING POINT NUMBERS

- It is insightful to see how floating point numbers are distributed along a number line
- Consider the non-negative floats that can be represented using 9 bits with 1 sign bit, 3 exponent bits, and 5 fractional part bits
- The biased exponents range from 000 to 111, that is from 0 to 7.
- Using a bias of $b = 4$ the “real” range is from -4 to 3
- The normalized values $1.f$ range from 1.00000 to 1.11111
- Expressed in decimal the normalized values range from

$$1 \quad \text{to} \quad 1 + \frac{31}{32} = \frac{63}{32}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

DISTRIBUTION OF FLOATING POINT NUMBERS

- The normalized values ($1.f$) increment in steps of $1/32$ as the values increase

DISTRIBUTION OF FLOATING POINT NUMBERS

- The normalized values ($1.f$) increment in steps of $1/32$ as the values increase

DISTRIBUTION OF FLOATING POINT NUMBERS

- The normalized values ($1.f$) increment in steps of $1/32$ as the values increase

1.00000

DISTRIBUTION OF FLOATING POINT NUMBERS

- The normalized values ($1.f$) increment in steps of $1/32$ as the values increase

$$1.00000 = 1$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The normalized values ($1.f$) increment in steps of $1/32$ as the values increase

$$1.00000 = 1$$

$$1.00001$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The normalized values ($1.f$) increment in steps of $1/32$ as the values increase

$$1.00000 = 1$$

$$1.00001 = 1 + \frac{1}{32} = \frac{33}{32}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The normalized values ($1.f$) increment in steps of $1/32$ as the values increase

$$1.00000 = 1$$

$$1.00001 = 1 + \frac{1}{32} = \frac{33}{32}$$

$$1.00010$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The normalized values ($1.f$) increment in steps of $1/32$ as the values increase

$$1.00000 = 1$$

$$1.00001 = 1 + \frac{1}{32} = \frac{33}{32}$$

$$1.00010 = 1 + \frac{2}{32} = \frac{34}{32}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The normalized values ($1.f$) increment in steps of $1/32$ as the values increase

$$1.00000 = 1$$

$$1.00001 = 1 + \frac{1}{32} = \frac{33}{32}$$

$$1.00010 = 1 + \frac{2}{32} = \frac{34}{32}$$

$$1.00011$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The normalized values ($1.f$) increment in steps of $1/32$ as the values increase

$$1.00000 = 1$$

$$1.00001 = 1 + \frac{1}{32} = \frac{33}{32}$$

$$1.00010 = 1 + \frac{2}{32} = \frac{34}{32}$$

$$1.00011 = 1 + \frac{3}{32} = \frac{35}{32}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The normalized values ($1.f$) increment in steps of $1/32$ as the values increase

$$1.00000 = 1$$

$$1.00001 = 1 + \frac{1}{32} = \frac{33}{32}$$

$$1.00010 = 1 + \frac{2}{32} = \frac{34}{32}$$

$$1.00011 = 1 + \frac{3}{32} = \frac{35}{32}$$

⋮

DISTRIBUTION OF FLOATING POINT NUMBERS

- The normalized values ($1.f$) increment in steps of $1/32$ as the values increase

$$1.00000 = 1$$

$$1.00001 = 1 + \frac{1}{32} = \frac{33}{32}$$

$$1.00010 = 1 + \frac{2}{32} = \frac{34}{32}$$

$$1.00011 = 1 + \frac{3}{32} = \frac{35}{32}$$

$$\vdots = \vdots$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The normalized values ($1.f$) increment in steps of $1/32$ as the values increase

$$1.00000 = 1$$

$$1.00001 = 1 + \frac{1}{32} = \frac{33}{32}$$

$$1.00010 = 1 + \frac{2}{32} = \frac{34}{32}$$

$$1.00011 = 1 + \frac{3}{32} = \frac{35}{32}$$

$$\vdots = \vdots$$

$$1.11110$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The normalized values ($1.f$) increment in steps of $1/32$ as the values increase

$$1.00000 = 1$$

$$1.00001 = 1 + \frac{1}{32} = \frac{33}{32}$$

$$1.00010 = 1 + \frac{2}{32} = \frac{34}{32}$$

$$1.00011 = 1 + \frac{3}{32} = \frac{35}{32}$$

$$\vdots = \vdots$$

$$1.11110 = 1 + \frac{30}{32} = \frac{62}{32}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The normalized values ($1.f$) increment in steps of $1/32$ as the values increase

$$1.00000 = 1$$

$$1.00001 = 1 + \frac{1}{32} = \frac{33}{32}$$

$$1.00010 = 1 + \frac{2}{32} = \frac{34}{32}$$

$$1.00011 = 1 + \frac{3}{32} = \frac{35}{32}$$

$$\vdots = \vdots$$

$$1.11110 = 1 + \frac{30}{32} = \frac{62}{32}$$

$$1.11111$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The normalized values ($1.f$) increment in steps of $1/32$ as the values increase

$$1.00000 = 1$$

$$1.00001 = 1 + \frac{1}{32} = \frac{33}{32}$$

$$1.00010 = 1 + \frac{2}{32} = \frac{34}{32}$$

$$1.00011 = 1 + \frac{3}{32} = \frac{35}{32}$$

$$\vdots = \vdots$$

$$1.11110 = 1 + \frac{30}{32} = \frac{62}{32}$$

$$1.11111 = 1 + \frac{31}{32} = \frac{63}{32}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

DISTRIBUTION OF FLOATING POINT NUMBERS

- The smallest exponential scale is 2^{-4}

DISTRIBUTION OF FLOATING POINT NUMBERS

- The smallest exponential scale is 2^{-4}
- At this scale, the normalized values increment is steps of

$$\frac{1}{32} \times \frac{1}{16} = \frac{1}{512}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The smallest exponential scale is 2^{-4}
- At this scale, the normalized values increment is steps of

$$\frac{1}{32} \times \frac{1}{16} = \frac{1}{512}$$

- And at this smallest scale, the representable positive floats are

DISTRIBUTION OF FLOATING POINT NUMBERS

- The smallest exponential scale is 2^{-4}
- At this scale, the normalized values increment is steps of

$$\frac{1}{32} \times \frac{1}{16} = \frac{1}{512}$$

- And at this smallest scale, the representable positive floats are

DISTRIBUTION OF FLOATING POINT NUMBERS

- The smallest exponential scale is 2^{-4}
- At this scale, the normalized values increment is steps of

$$\frac{1}{32} \times \frac{1}{16} = \frac{1}{512}$$

- And at this smallest scale, the representable positive floats are

0 000 00001

DISTRIBUTION OF FLOATING POINT NUMBERS

- The smallest exponential scale is 2^{-4}
- At this scale, the normalized values increment is steps of

$$\frac{1}{32} \times \frac{1}{16} = \frac{1}{512}$$

- And at this smallest scale, the representable positive floats are

$$0\ 000\ 00001 = \frac{33}{512}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The smallest exponential scale is 2^{-4}
- At this scale, the normalized values increment is steps of

$$\frac{1}{32} \times \frac{1}{16} = \frac{1}{512}$$

- And at this smallest scale, the representable positive floats are

$$0\ 000\ 00001 = \frac{33}{512}$$

$$0\ 000\ 00010$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The smallest exponential scale is 2^{-4}
- At this scale, the normalized values increment is steps of

$$\frac{1}{32} \times \frac{1}{16} = \frac{1}{512}$$

- And at this smallest scale, the representable positive floats are

$$0\ 000\ 00001 = \frac{33}{512}$$

$$0\ 000\ 00010 = \frac{34}{512}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The smallest exponential scale is 2^{-4}
- At this scale, the normalized values increment is steps of

$$\frac{1}{32} \times \frac{1}{16} = \frac{1}{512}$$

- And at this smallest scale, the representable positive floats are

$$0\ 000\ 00001 = \frac{33}{512}$$

$$0\ 000\ 00010 = \frac{34}{512}$$

⋮

DISTRIBUTION OF FLOATING POINT NUMBERS

- The smallest exponential scale is 2^{-4}
- At this scale, the normalized values increment is steps of

$$\frac{1}{32} \times \frac{1}{16} = \frac{1}{512}$$

- And at this smallest scale, the representable positive floats are

$$0\ 000\ 00001 = \frac{33}{512}$$

$$0\ 000\ 00010 = \frac{34}{512}$$

$$\vdots = \vdots$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The smallest exponential scale is 2^{-4}
- At this scale, the normalized values increment is steps of

$$\frac{1}{32} \times \frac{1}{16} = \frac{1}{512}$$

- And at this smallest scale, the representable positive floats are

$$0\ 000\ 00001 = \frac{33}{512}$$

$$0\ 000\ 00010 = \frac{34}{512}$$

$$\vdots = \vdots$$

$$0\ 000\ 11110$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The smallest exponential scale is 2^{-4}
- At this scale, the normalized values increment is steps of

$$\frac{1}{32} \times \frac{1}{16} = \frac{1}{512}$$

- And at this smallest scale, the representable positive floats are

$$0\ 000\ 00001 = \frac{33}{512}$$

$$0\ 000\ 00010 = \frac{34}{512}$$

$$\vdots = \vdots$$

$$0\ 000\ 11110 = \frac{62}{512}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The smallest exponential scale is 2^{-4}
- At this scale, the normalized values increment is steps of

$$\frac{1}{32} \times \frac{1}{16} = \frac{1}{512}$$

- And at this smallest scale, the representable positive floats are

$$0\ 000\ 00001 = \frac{33}{512}$$

$$0\ 000\ 00010 = \frac{34}{512}$$

$$\vdots = \vdots$$

$$0\ 000\ 11110 = \frac{62}{512}$$

$$0\ 000\ 11111$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The smallest exponential scale is 2^{-4}
- At this scale, the normalized values increment is steps of

$$\frac{1}{32} \times \frac{1}{16} = \frac{1}{512}$$

- And at this smallest scale, the representable positive floats are

$$0\ 000\ 00001 = \frac{33}{512}$$

$$0\ 000\ 00010 = \frac{34}{512}$$

$$\vdots = \vdots$$

$$0\ 000\ 11110 = \frac{62}{512}$$

$$0\ 000\ 11111 = \frac{63}{512}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

DISTRIBUTION OF FLOATING POINT NUMBERS

- At this smallest exponential scale the floats range from

$$\frac{33}{512}$$

$$\frac{63}{512}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- At this smallest exponential scale the floats range from

$$\frac{33}{512} \qquad \frac{63}{512}$$

- That is, these float range from their smallest positive value, a little bigger than $1/16$, to almost $1/8$

DISTRIBUTION OF FLOATING POINT NUMBERS

- At this smallest exponential scale the floats range from

$$\frac{33}{512} \qquad \frac{63}{512}$$

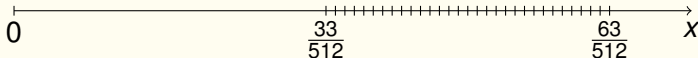
- That is, these float range from their smallest positive value, a little bigger than $1/16$, to almost $1/8$
- The representable floats in this range are “fairly” tightly bunched: Separated by steps of $1/512$

DISTRIBUTION OF FLOATING POINT NUMBERS

- There are $63 - 33 + 1 = 31$ floating point numbers between

$$\frac{33}{512} \qquad \frac{63}{512}$$

- There is a large gap in representable numbers from 0 to $33/512$



DISTRIBUTION OF FLOATING POINT NUMBERS

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-3}

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-3}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(8)} = \frac{1}{256}$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-3}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(8)} = \frac{1}{256}$
- And at this scale, the representable floats are

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-3}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(8)} = \frac{1}{256}$
- And at this scale, the representable floats are

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-3}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(8)} = \frac{1}{256}$
- And at this scale, the representable floats are

0 001 00000

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-3}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(8)} = \frac{1}{256}$
- And at this scale, the representable floats are

$$0\ 001\ 00000 = \frac{32}{256}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-3}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(8)} = \frac{1}{256}$
- And at this scale, the representable floats are

$$0\ 001\ 00000 = \frac{32}{256}$$

$$0\ 001\ 00001$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-3}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(8)} = \frac{1}{256}$
- And at this scale, the representable floats are

$$0\ 001\ 00000 = \frac{32}{256}$$

$$0\ 001\ 00001 = \frac{33}{256}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-3}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(8)} = \frac{1}{256}$
- And at this scale, the representable floats are

$$0\ 001\ 00000 = \frac{32}{256}$$

$$0\ 001\ 00001 = \frac{33}{256}$$

⋮

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-3}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(8)} = \frac{1}{256}$
- And at this scale, the representable floats are

$$0\ 001\ 00000 = \frac{32}{256}$$

$$0\ 001\ 00001 = \frac{33}{256}$$

$$\vdots = \vdots$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-3}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(8)} = \frac{1}{256}$
- And at this scale, the representable floats are

$$0\ 001\ 00000 = \frac{32}{256}$$

$$0\ 001\ 00001 = \frac{33}{256}$$

$$\vdots = \vdots$$

$$0\ 001\ 11110$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-3}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(8)} = \frac{1}{256}$
- And at this scale, the representable floats are

$$0\ 001\ 00000 = \frac{32}{256}$$

$$0\ 001\ 00001 = \frac{33}{256}$$

$$\vdots = \vdots$$

$$0\ 001\ 11110 = \frac{62}{256}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-3}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(8)} = \frac{1}{256}$
- And at this scale, the representable floats are

$$0\ 001\ 00000 = \frac{32}{256}$$

$$0\ 001\ 00001 = \frac{33}{256}$$

$$\vdots = \vdots$$

$$0\ 001\ 11110 = \frac{62}{256}$$

$$0\ 001\ 11111$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-3}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(8)} = \frac{1}{256}$
- And at this scale, the representable floats are

$$0\ 001\ 00000 = \frac{32}{256}$$

$$0\ 001\ 00001 = \frac{33}{256}$$

$$\vdots = \vdots$$

$$0\ 001\ 11110 = \frac{62}{256}$$

$$0\ 001\ 11111 = \frac{63}{256}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

DISTRIBUTION OF FLOATING POINT NUMBERS

- At this next smallest exponential scale the floats range from

$$\frac{32}{256}$$

$$\frac{63}{256}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- At this next smallest exponential scale the floats range from

$$\frac{32}{256}$$

$$\frac{63}{256}$$

- That is, 1/8 to almost 1/4

DISTRIBUTION OF FLOATING POINT NUMBERS

- At this next smallest exponential scale the floats range from

$$\frac{32}{256}$$

$$\frac{63}{256}$$

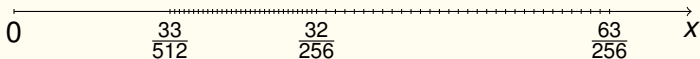
- That is, $1/8$ to almost $1/4$
- The representable floats in this range are less tightly bunched: Separated by steps of $1/256$

DISTRIBUTION OF FLOATING POINT NUMBERS

- There are $63 - 32 + 1 = 32$ floating point numbers between

$$\frac{32}{256}$$

$$\frac{63}{256}$$



DISTRIBUTION OF FLOATING POINT NUMBERS

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-2}

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-2}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(4)} = \frac{1}{128}$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-2}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(4)} = \frac{1}{128}$
- And at this scale, the representable floats are

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-2}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(4)} = \frac{1}{128}$
- And at this scale, the representable floats are

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-2}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(4)} = \frac{1}{128}$
- And at this scale, the representable floats are

0 010 00000

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-2}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(4)} = \frac{1}{128}$
- And at this scale, the representable floats are

$$0\ 010\ 00000 = \frac{32}{128}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-2}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(4)} = \frac{1}{128}$
- And at this scale, the representable floats are

$$0\ 010\ 00000 = \frac{32}{128}$$

$$0\ 010\ 00001$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-2}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(4)} = \frac{1}{128}$
- And at this scale, the representable floats are

$$0\ 010\ 00000 = \frac{32}{128}$$

$$0\ 010\ 00001 = \frac{33}{128}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-2}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(4)} = \frac{1}{128}$
- And at this scale, the representable floats are

$$0\ 010\ 00000 = \frac{32}{128}$$

$$0\ 010\ 00001 = \frac{33}{128}$$

$$0\ 010\ 00010$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-2}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(4)} = \frac{1}{128}$
- And at this scale, the representable floats are

$$0\ 010\ 00000 = \frac{32}{128}$$

$$0\ 010\ 00001 = \frac{33}{128}$$

$$0\ 010\ 00010 = \frac{34}{128}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-2}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(4)} = \frac{1}{128}$
- And at this scale, the representable floats are

$$0\ 010\ 00000 = \frac{32}{128}$$

$$0\ 010\ 00001 = \frac{33}{128}$$

$$0\ 010\ 00010 = \frac{34}{128}$$

⋮

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-2}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(4)} = \frac{1}{128}$
- And at this scale, the representable floats are

$$0\ 010\ 00000 = \frac{32}{128}$$

$$0\ 010\ 00001 = \frac{33}{128}$$

$$0\ 010\ 00010 = \frac{34}{128}$$

$$\vdots = \vdots$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-2}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(4)} = \frac{1}{128}$
- And at this scale, the representable floats are

$$0\ 010\ 00000 = \frac{32}{128}$$

$$0\ 010\ 00001 = \frac{33}{128}$$

$$0\ 010\ 00010 = \frac{34}{128}$$

$$\vdots = \vdots$$

$$0\ 010\ 11110$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-2}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(4)} = \frac{1}{128}$
- And at this scale, the representable floats are

$$0\ 010\ 00000 = \frac{32}{128}$$

$$0\ 010\ 00001 = \frac{33}{128}$$

$$0\ 010\ 00010 = \frac{34}{128}$$

$$\vdots = \vdots$$

$$0\ 010\ 11110 = \frac{62}{128}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-2}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(4)} = \frac{1}{128}$
- And at this scale, the representable floats are

$$0\ 010\ 00000 = \frac{32}{128}$$

$$0\ 010\ 00001 = \frac{33}{128}$$

$$0\ 010\ 00010 = \frac{34}{128}$$

$$\vdots = \vdots$$

$$0\ 010\ 11110 = \frac{62}{128}$$

$$0\ 010\ 11111$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The next smallest exponential scale is 2^{-2}
- At this scale, the normalized values increment is steps of $\frac{1}{(32)(4)} = \frac{1}{128}$
- And at this scale, the representable floats are

$$0\ 010\ 00000 = \frac{32}{128}$$

$$0\ 010\ 00001 = \frac{33}{128}$$

$$0\ 010\ 00010 = \frac{34}{128}$$

$$\vdots = \vdots$$

$$0\ 010\ 11110 = \frac{62}{128}$$

$$0\ 010\ 11111 = \frac{63}{128}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

DISTRIBUTION OF FLOATING POINT NUMBERS

- At exponential scale 2^{-2} the floats range from

$$\frac{32}{128}$$

$$\frac{63}{128}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- At exponential scale 2^{-2} the floats range from

$$\frac{32}{128}$$

$$\frac{63}{128}$$

- That is, from $1/4$ to almost $1/2$

DISTRIBUTION OF FLOATING POINT NUMBERS

- At exponential scale 2^{-2} the floats range from

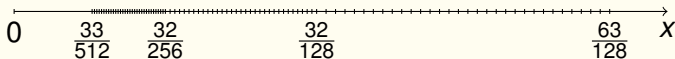
$$\frac{32}{128} \qquad \frac{63}{128}$$

- That is, from $1/4$ to almost $1/2$
- The representable floats in this range are still less tightly bunched: Separated by steps of $1/128$

DISTRIBUTION OF FLOATING POINT NUMBERS

- There are $63 - 32 + 1 = 32$ floating point numbers between

$$\frac{32}{128} \qquad \frac{63}{128}$$



DISTRIBUTION OF FLOATING POINT NUMBERS

DISTRIBUTION OF FLOATING POINT NUMBERS

- The largest exponential scale is 2^3

DISTRIBUTION OF FLOATING POINT NUMBERS

- The largest exponential scale is 2^3
- At this scale, the normalized values increment is steps of $\frac{8}{32} = \frac{1}{4}$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The largest exponential scale is 2^3
- At this scale, the normalized values increment is steps of $\frac{8}{32} = \frac{1}{4}$
- And at this scale, the representable floats are

DISTRIBUTION OF FLOATING POINT NUMBERS

- The largest exponential scale is 2^3
- At this scale, the normalized values increment is steps of $\frac{8}{32} = \frac{1}{4}$
- And at this scale, the representable floats are

DISTRIBUTION OF FLOATING POINT NUMBERS

- The largest exponential scale is 2^3
- At this scale, the normalized values increment is steps of $\frac{8}{32} = \frac{1}{4}$
- And at this scale, the representable floats are

0 111 00000

DISTRIBUTION OF FLOATING POINT NUMBERS

- The largest exponential scale is 2^3
- At this scale, the normalized values increment is steps of $\frac{8}{32} = \frac{1}{4}$
- And at this scale, the representable floats are

$$0\ 111\ 00000 = \frac{32}{4}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The largest exponential scale is 2^3
- At this scale, the normalized values increment is steps of $\frac{8}{32} = \frac{1}{4}$
- And at this scale, the representable floats are

$$0\ 111\ 00000 = \frac{32}{4}$$

$$0\ 111\ 00001$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The largest exponential scale is 2^3
- At this scale, the normalized values increment is steps of $\frac{8}{32} = \frac{1}{4}$
- And at this scale, the representable floats are

$$0\ 111\ 00000 = \frac{32}{4}$$

$$0\ 111\ 00001 = \frac{33}{4}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The largest exponential scale is 2^3
- At this scale, the normalized values increment is steps of $\frac{8}{32} = \frac{1}{4}$
- And at this scale, the representable floats are

$$0\ 111\ 00000 = \frac{32}{4}$$

$$0\ 111\ 00001 = \frac{33}{4}$$

$\vdots$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The largest exponential scale is 2^3
- At this scale, the normalized values increment is steps of $\frac{8}{32} = \frac{1}{4}$
- And at this scale, the representable floats are

$$0\ 111\ 00000 = \frac{32}{4}$$

$$0\ 111\ 00001 = \frac{33}{4}$$

$$\vdots = \vdots$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The largest exponential scale is 2^3
- At this scale, the normalized values increment is steps of $\frac{8}{32} = \frac{1}{4}$
- And at this scale, the representable floats are

$$0\ 111\ 00000 = \frac{32}{4}$$

$$0\ 111\ 00001 = \frac{33}{4}$$

$$\vdots = \vdots$$

$$0\ 111\ 11110$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The largest exponential scale is 2^3
- At this scale, the normalized values increment is steps of $\frac{8}{32} = \frac{1}{4}$
- And at this scale, the representable floats are

$$0\ 111\ 00000 = \frac{32}{4}$$

$$0\ 111\ 00001 = \frac{33}{4}$$

$$\vdots = \vdots$$

$$0\ 111\ 11110 = \frac{62}{4}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The largest exponential scale is 2^3
- At this scale, the normalized values increment is steps of $\frac{8}{32} = \frac{1}{4}$
- And at this scale, the representable floats are

$$0\ 111\ 00000 = \frac{32}{4}$$

$$0\ 111\ 00001 = \frac{33}{4}$$

$$\vdots = \vdots$$

$$0\ 111\ 11110 = \frac{62}{4}$$

$$0\ 111\ 11111$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The largest exponential scale is 2^3
- At this scale, the normalized values increment is steps of $\frac{8}{32} = \frac{1}{4}$
- And at this scale, the representable floats are

$$0\ 111\ 00000 = \frac{32}{4}$$

$$0\ 111\ 00001 = \frac{33}{4}$$

$$\vdots = \vdots$$

$$0\ 111\ 11110 = \frac{62}{4}$$

$$0\ 111\ 11111 = \frac{63}{4}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

DISTRIBUTION OF FLOATING POINT NUMBERS

- At the largest exponential scale the floats range from

$$\frac{32}{4}$$

$$\frac{63}{4}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- At the largest exponential scale the floats range from

$$\frac{32}{4}$$

$$\frac{63}{4}$$

- That is, from 8 to their largest possible value $63/4$

DISTRIBUTION OF FLOATING POINT NUMBERS

- At the largest exponential scale the floats range from

$$\frac{32}{4}$$

$$\frac{63}{4}$$

- That is, from 8 to their largest possible value $63/4$
- The representable floats in this range are not tightly packed: They are separated by steps of $1/4$

DISTRIBUTION OF FLOATING POINT NUMBERS

DISTRIBUTION OF FLOATING POINT NUMBERS

- The smallest positive float is

$$0\ 000\ 00001 = \frac{33}{32} \times 2^{-4} = \frac{33}{512}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- The smallest positive float is

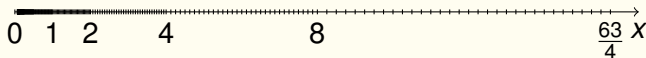
$$0\ 000\ 00001 = \frac{33}{32} \times 2^{-4} = \frac{33}{512}$$

- The largest positive float is

$$0\ 111\ 11111 = \frac{63}{32} \times 2^3 = \frac{63}{4}$$

DISTRIBUTION OF FLOATING POINT NUMBERS

- Here is a complete graph of the representable floating point numbers using this scheme



A SMALLER EXAMPLE

A SMALLER EXAMPLE

- Consider the floating point numbers that can be represented with

A SMALLER EXAMPLE

- Consider the floating point numbers that can be represented with
 - 1 sign bit

A SMALLER EXAMPLE

- Consider the floating point numbers that can be represented with
 - 1 sign bit
 - 2 exponent bits

A SMALLER EXAMPLE

- Consider the floating point numbers that can be represented with
 - 1 sign bit
 - 2 exponent bits
 - 3 fractional part bits

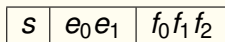
A SMALLER EXAMPLE

- Consider the floating point numbers that can be represented with
 - 1 sign bit
 - 2 exponent bits
 - 3 fractional part bits
- These float point numbers have the form

s	$e_0 e_1$	$f_0 f_1 f_2$
-----	-----------	---------------

A SMALLER EXAMPLE

- Consider the floating point numbers that can be represented with
 - 1 sign bit
 - 2 exponent bits
 - 3 fractional part bits
- These float point numbers have the form

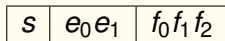


- 0 is represented by the denormalized string

0 00 000

A SMALLER EXAMPLE

- Consider the floating point numbers that can be represented with
 - 1 sign bit
 - 2 exponent bits
 - 3 fractional part bits
- These float point numbers have the form



- 0 is represented by the denormalized string

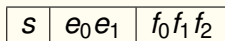
0 00 000

- The smallest positive float is represented by

0 00 001

A SMALLER EXAMPLE

- Consider the floating point numbers that can be represented with
 - 1 sign bit
 - 2 exponent bits
 - 3 fractional part bits
- These float point numbers have the form



- 0 is represented by the denormalized string

0 00 000

- The smallest positive float is represented by

0 00 001

- The largest positive float is represented by

0 11 111

A SMALLER EXAMPLE: THE BIASED EXPONENTS

A SMALLER EXAMPLE: THE BIASED EXPONENTS

- The 2-bit biased exponent ranges from 00 to 11

A SMALLER EXAMPLE: THE BIASED EXPONENTS

- The 2-bit biased exponent ranges from 00 to 11
- Choosing a bias of $b = 2$ gives a “real” or “unbiased” exponent range from -2 to 1

A SMALLER EXAMPLE: THE BIASED EXPONENTS

- The 2-bit biased exponent ranges from 00 to 11
- Choosing a bias of $b = 2$ gives a “real” or “unbiased” exponent range from -2 to 1
- And the normalized fractional parts ($1.f$) can be scaled by

$$2^{-2}, \quad 2^{-1}, \quad 2^0, \quad 2^1$$

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

1.000

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

$$1.000 = 1$$

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

$$1.000 = 1$$

$$1.001$$

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

$$1.000 = 1$$

$$1.001 = \frac{9}{8}$$

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

$$1.000 = 1$$

$$1.001 = \frac{9}{8}$$

$$1.010$$

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

$$1.000 = 1$$

$$1.001 = \frac{9}{8}$$

$$1.010 = \frac{10}{8}$$

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

$$1.000 = 1$$

$$1.001 = \frac{9}{8}$$

$$1.010 = \frac{10}{8}$$

$$1.011$$

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

$$1.000 = 1$$

$$1.001 = \frac{9}{8}$$

$$1.010 = \frac{10}{8}$$

$$1.011 = \frac{11}{8}$$

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

$$1.000 = 1$$

$$1.001 = \frac{9}{8}$$

$$1.010 = \frac{10}{8}$$

$$1.011 = \frac{11}{8}$$

$$1.100$$

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

$$1.000 = 1$$

$$1.001 = \frac{9}{8}$$

$$1.010 = \frac{10}{8}$$

$$1.011 = \frac{11}{8}$$

$$1.100 = \frac{12}{8}$$

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

$$1.000 = 1$$

$$1.001 = \frac{9}{8}$$

$$1.010 = \frac{10}{8}$$

$$1.011 = \frac{11}{8}$$

$$1.100 = \frac{12}{8}$$

$$1.101$$

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

$$1.000 = 1$$

$$1.001 = \frac{9}{8}$$

$$1.010 = \frac{10}{8}$$

$$1.011 = \frac{11}{8}$$

$$1.100 = \frac{12}{8}$$

$$1.101 = \frac{13}{8}$$

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

$$1.000 = 1$$

$$1.001 = \frac{9}{8}$$

$$1.010 = \frac{10}{8}$$

$$1.011 = \frac{11}{8}$$

$$1.100 = \frac{12}{8}$$

$$1.101 = \frac{13}{8}$$

$$1.110$$

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

$$1.000 = 1$$

$$1.001 = \frac{9}{8}$$

$$1.010 = \frac{10}{8}$$

$$1.011 = \frac{11}{8}$$

$$1.100 = \frac{12}{8}$$

$$1.101 = \frac{13}{8}$$

$$1.110 = \frac{14}{8}$$

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

$$1.000 = 1$$

$$1.001 = \frac{9}{8}$$

$$1.010 = \frac{10}{8}$$

$$1.011 = \frac{11}{8}$$

$$1.100 = \frac{12}{8}$$

$$1.101 = \frac{13}{8}$$

$$1.110 = \frac{14}{8}$$

$$1.111$$

A SMALLER EXAMPLE: THE NORMALIZED FRACTIONAL PARTS

- The normalized fractional parts ($1.f$) range from

1.000 to 1.111

- Convert the values in this range to decimal notation

$$1.000 = 1$$

$$1.010 = \frac{10}{8}$$

$$1.100 = \frac{12}{8}$$

$$1.110 = \frac{14}{8}$$

$$1.001 = \frac{9}{8}$$

$$1.011 = \frac{11}{8}$$

$$1.101 = \frac{13}{8}$$

$$1.111 = \frac{15}{8}$$

A SMALLER EXAMPLE: THE SMALLEST SCALE

A SMALLER EXAMPLE: THE SMALLEST SCALE

- At the smallest scale, 2^{-2} , the smallest representable float is

$$0\ 00\ 001 = +\frac{9}{8} \times 2^{-2} = \frac{9}{32}$$

A SMALLER EXAMPLE: THE SMALLEST SCALE

- At the smallest scale, 2^{-2} , the smallest representable float is

$$0\ 00\ 001 = +\frac{9}{8} \times 2^{-2} = \frac{9}{32}$$

- The other representable float point numbers at this scale are

A SMALLER EXAMPLE: THE SMALLEST SCALE

- At the smallest scale, 2^{-2} , the smallest representable float is

$$0\ 00\ 001 = +\frac{9}{8} \times 2^{-2} = \frac{9}{32}$$

- The other representable float point numbers at this scale are

A SMALLER EXAMPLE: THE SMALLEST SCALE

- At the smallest scale, 2^{-2} , the smallest representable float is

$$0.00001 = +\frac{9}{8} \times 2^{-2} = \frac{9}{32}$$

- The other representable float point numbers at this scale are

$$\frac{10}{8} \times 2^{-2}$$

A SMALLER EXAMPLE: THE SMALLEST SCALE

- At the smallest scale, 2^{-2} , the smallest representable float is

$$0.00001 = +\frac{9}{8} \times 2^{-2} = \frac{9}{32}$$

- The other representable float point numbers at this scale are

$$\frac{10}{8} \times 2^{-2} = \frac{10}{32}$$

A SMALLER EXAMPLE: THE SMALLEST SCALE

- At the smallest scale, 2^{-2} , the smallest representable float is

$$0\ 00\ 001 = +\frac{9}{8} \times 2^{-2} = \frac{9}{32}$$

- The other representable float point numbers at this scale are

$$\frac{10}{8} \times 2^{-2} = \frac{10}{32}$$

$$\frac{11}{8} \times 2^{-2}$$

A SMALLER EXAMPLE: THE SMALLEST SCALE

- At the smallest scale, 2^{-2} , the smallest representable float is

$$0.00001 = +\frac{9}{8} \times 2^{-2} = \frac{9}{32}$$

- The other representable float point numbers at this scale are

$$\frac{10}{8} \times 2^{-2} = \frac{10}{32}$$

$$\frac{11}{8} \times 2^{-2} = \frac{11}{32}$$

A SMALLER EXAMPLE: THE SMALLEST SCALE

- At the smallest scale, 2^{-2} , the smallest representable float is

$$0\ 00\ 001 = +\frac{9}{8} \times 2^{-2} = \frac{9}{32}$$

- The other representable float point numbers at this scale are

$$\frac{10}{8} \times 2^{-2} = \frac{10}{32}$$

$$\frac{12}{8} \times 2^{-2}$$

$$\frac{11}{8} \times 2^{-2} = \frac{11}{32}$$

A SMALLER EXAMPLE: THE SMALLEST SCALE

- At the smallest scale, 2^{-2} , the smallest representable float is

$$0\ 00\ 001 = +\frac{9}{8} \times 2^{-2} = \frac{9}{32}$$

- The other representable float point numbers at this scale are

$$\frac{10}{8} \times 2^{-2} = \frac{10}{32}$$

$$\frac{12}{8} \times 2^{-2} = \frac{12}{32}$$

$$\frac{11}{8} \times 2^{-2} = \frac{11}{32}$$

A SMALLER EXAMPLE: THE SMALLEST SCALE

- At the smallest scale, 2^{-2} , the smallest representable float is

$$0\ 00\ 001 = +\frac{9}{8} \times 2^{-2} = \frac{9}{32}$$

- The other representable float point numbers at this scale are

$$\frac{10}{8} \times 2^{-2} = \frac{10}{32}$$

$$\frac{12}{8} \times 2^{-2} = \frac{12}{32}$$

$$\frac{11}{8} \times 2^{-2} = \frac{11}{32}$$

$$\frac{13}{8} \times 2^{-2}$$

A SMALLER EXAMPLE: THE SMALLEST SCALE

- At the smallest scale, 2^{-2} , the smallest representable float is

$$0\ 00\ 001 = +\frac{9}{8} \times 2^{-2} = \frac{9}{32}$$

- The other representable float point numbers at this scale are

$$\frac{10}{8} \times 2^{-2} = \frac{10}{32}$$

$$\frac{12}{8} \times 2^{-2} = \frac{12}{32}$$

$$\frac{11}{8} \times 2^{-2} = \frac{11}{32}$$

$$\frac{13}{8} \times 2^{-2} = \frac{13}{32}$$

A SMALLER EXAMPLE: THE SMALLEST SCALE

- At the smallest scale, 2^{-2} , the smallest representable float is

$$0\ 00\ 001 = +\frac{9}{8} \times 2^{-2} = \frac{9}{32}$$

- The other representable float point numbers at this scale are

$$\frac{10}{8} \times 2^{-2} = \frac{10}{32}$$

$$\frac{12}{8} \times 2^{-2} = \frac{12}{32}$$

$$\frac{14}{8} \times 2^{-2}$$

$$\frac{11}{8} \times 2^{-2} = \frac{11}{32}$$

$$\frac{13}{8} \times 2^{-2} = \frac{13}{32}$$

A SMALLER EXAMPLE: THE SMALLEST SCALE

- At the smallest scale, 2^{-2} , the smallest representable float is

$$0\ 00\ 001 = +\frac{9}{8} \times 2^{-2} = \frac{9}{32}$$

- The other representable float point numbers at this scale are

$$\frac{10}{8} \times 2^{-2} = \frac{10}{32}$$

$$\frac{12}{8} \times 2^{-2} = \frac{12}{32}$$

$$\frac{14}{8} \times 2^{-2} = \frac{14}{32}$$

$$\frac{11}{8} \times 2^{-2} = \frac{11}{32}$$

$$\frac{13}{8} \times 2^{-2} = \frac{13}{32}$$

A SMALLER EXAMPLE: THE SMALLEST SCALE

- At the smallest scale, 2^{-2} , the smallest representable float is

$$0\ 00\ 001 = +\frac{9}{8} \times 2^{-2} = \frac{9}{32}$$

- The other representable float point numbers at this scale are

$$\frac{10}{8} \times 2^{-2} = \frac{10}{32}$$

$$\frac{12}{8} \times 2^{-2} = \frac{12}{32}$$

$$\frac{14}{8} \times 2^{-2} = \frac{14}{32}$$

$$\frac{11}{8} \times 2^{-2} = \frac{11}{32}$$

$$\frac{13}{8} \times 2^{-2} = \frac{13}{32}$$

$$\frac{15}{8} \times 2^{-2}$$

A SMALLER EXAMPLE: THE SMALLEST SCALE

- At the smallest scale, 2^{-2} , the smallest representable float is

$$0\ 00\ 001 = +\frac{9}{8} \times 2^{-2} = \frac{9}{32}$$

- The other representable float point numbers at this scale are

$$\frac{10}{8} \times 2^{-2} = \frac{10}{32}$$

$$\frac{12}{8} \times 2^{-2} = \frac{12}{32}$$

$$\frac{14}{8} \times 2^{-2} = \frac{14}{32}$$

$$\frac{11}{8} \times 2^{-2} = \frac{11}{32}$$

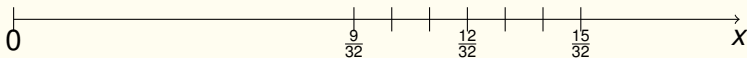
$$\frac{13}{8} \times 2^{-2} = \frac{13}{32}$$

$$\frac{15}{8} \times 2^{-2} = \frac{15}{32}$$

A SMALLER EXAMPLE: THE SMALLEST SCALE GRAPH

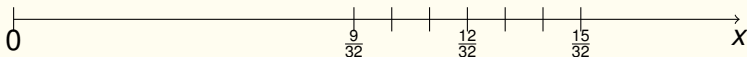
A SMALLER EXAMPLE: THE SMALLEST SCALE GRAPH

- A number line graph of these values is



A SMALLER EXAMPLE: THE SMALLEST SCALE GRAPH

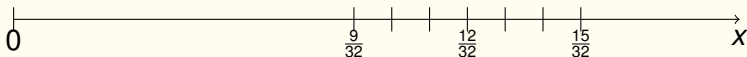
- A number line graph of these values is



- Any computed number x between 0 and $\frac{9}{32}$ will be rounded to either 0 or $\frac{9}{32}$

A SMALLER EXAMPLE: THE SMALLEST SCALE GRAPH

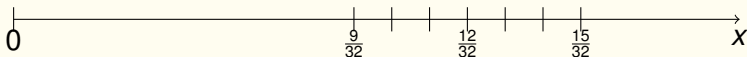
- A number line graph of these values is



- Any computed number x between 0 and $\frac{9}{32}$ will be rounded to either 0 or $\frac{9}{32}$
- Any computed x between representable float point numbers will be rounded

A SMALLER EXAMPLE: THE SMALLEST SCALE GRAPH

- A number line graph of these values is



- Any computed number x between 0 and $\frac{9}{32}$ will be rounded to either 0 or $\frac{9}{32}$
- Any computed x between representable float point numbers will be rounded
- These are called floating point **rounding errors**

A SMALLER EXAMPLE: THE LARGEST SCALE

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

- The other representable float point numbers at this scale are

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

- The other representable float point numbers at this scale are

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^1$$

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^1 = \frac{9}{4}$$

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^1 = \frac{9}{4}$$

$$\frac{10}{8} \times 2^1$$

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^1 = \frac{9}{4}$$

$$\frac{10}{8} \times 2^1 = \frac{10}{4}$$

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^1 = \frac{9}{4}$$
$$\frac{11}{8} \times 2^1$$

$$\frac{10}{8} \times 2^1 = \frac{10}{4}$$

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^1 &= \frac{9}{4} \\ \frac{11}{8} \times 2^1 &= \frac{11}{4}\end{aligned}$$

$$\frac{10}{8} \times 2^1 = \frac{10}{4}$$

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^1 &= \frac{9}{4} \\ \frac{11}{8} \times 2^1 &= \frac{11}{4}\end{aligned}$$

$$\begin{aligned}\frac{10}{8} \times 2^1 &= \frac{10}{4} \\ \frac{12}{8} \times 2^1 &= \frac{12}{4}\end{aligned}$$

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^1 &= \frac{9}{4} \\ \frac{11}{8} \times 2^1 &= \frac{11}{4}\end{aligned}$$

$$\begin{aligned}\frac{10}{8} \times 2^1 &= \frac{10}{4} \\ \frac{12}{8} \times 2^1 &= \frac{12}{4}\end{aligned}$$

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^1 &= \frac{9}{4} \\ \frac{11}{8} \times 2^1 &= \frac{11}{4} \\ \frac{13}{8} \times 2^1 &= \frac{13}{4}\end{aligned}$$

$$\begin{aligned}\frac{10}{8} \times 2^1 &= \frac{10}{4} \\ \frac{12}{8} \times 2^1 &= \frac{12}{4}\end{aligned}$$

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^1 &= \frac{9}{4} \\ \frac{11}{8} \times 2^1 &= \frac{11}{4} \\ \frac{13}{8} \times 2^1 &= \frac{13}{4}\end{aligned}$$

$$\begin{aligned}\frac{10}{8} \times 2^1 &= \frac{10}{4} \\ \frac{12}{8} \times 2^1 &= \frac{12}{4}\end{aligned}$$

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^1 &= \frac{9}{4} \\ \frac{11}{8} \times 2^1 &= \frac{11}{4} \\ \frac{13}{8} \times 2^1 &= \frac{13}{4}\end{aligned}$$

$$\begin{aligned}\frac{10}{8} \times 2^1 &= \frac{10}{4} \\ \frac{12}{8} \times 2^1 &= \frac{12}{4} \\ \frac{14}{8} \times 2^1 &= \frac{14}{4}\end{aligned}$$

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^1 &= \frac{9}{4} \\ \frac{11}{8} \times 2^1 &= \frac{11}{4} \\ \frac{13}{8} \times 2^1 &= \frac{13}{4}\end{aligned}$$

$$\begin{aligned}\frac{10}{8} \times 2^1 &= \frac{10}{4} \\ \frac{12}{8} \times 2^1 &= \frac{12}{4} \\ \frac{14}{8} \times 2^1 &= \frac{14}{4}\end{aligned}$$

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^1 &= \frac{9}{4} \\ \frac{11}{8} \times 2^1 &= \frac{11}{4} \\ \frac{13}{8} \times 2^1 &= \frac{13}{4} \\ \frac{15}{8} \times 2^1 &= \frac{15}{4}\end{aligned}$$

$$\begin{aligned}\frac{10}{8} \times 2^1 &= \frac{10}{4} \\ \frac{12}{8} \times 2^1 &= \frac{12}{4} \\ \frac{14}{8} \times 2^1 &= \frac{14}{4}\end{aligned}$$

A SMALLER EXAMPLE: THE LARGEST SCALE

- At the largest scale, 2^1 , the first representable float is

$$0\ 11\ 000 = +1 \times 2^1 = 2 = \frac{8}{4}$$

- The other representable float point numbers at this scale are

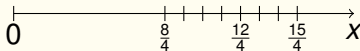
$$\begin{aligned}\frac{9}{8} \times 2^1 &= \frac{9}{4} \\ \frac{11}{8} \times 2^1 &= \frac{11}{4} \\ \frac{13}{8} \times 2^1 &= \frac{13}{4} \\ \frac{15}{8} \times 2^1 &= \frac{15}{4}\end{aligned}$$

$$\begin{aligned}\frac{10}{8} \times 2^1 &= \frac{10}{4} \\ \frac{12}{8} \times 2^1 &= \frac{12}{4} \\ \frac{14}{8} \times 2^1 &= \frac{14}{4}\end{aligned}$$

A SMALLER EXAMPLE: THE LARGEST SCALE GRAPH

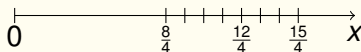
A SMALLER EXAMPLE: THE LARGEST SCALE GRAPH

- A number line graph of these values is



A SMALLER EXAMPLE: THE LARGEST SCALE GRAPH

- A number line graph of these values is



- Any computed number x between these representable float point numbers will be rounded-off

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

- The other representable float point numbers at this scale are

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

- The other representable float point numbers at this scale are

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^{-1}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^{-1} = \frac{9}{16}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^{-1} = \frac{9}{16}$$

$$\frac{10}{8} \times 2^{-1}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^{-1} = \frac{9}{16}$$

$$\frac{10}{8} \times 2^{-1} = \frac{10}{16}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^{-1} = \frac{9}{16}$$

$$\frac{11}{8} \times 2^{-1}$$

$$\frac{10}{8} \times 2^{-1} = \frac{10}{16}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^{-1} &= \frac{9}{16} \\ \frac{11}{8} \times 2^{-1} &= \frac{11}{16}\end{aligned}$$

$$\frac{10}{8} \times 2^{-1} = \frac{10}{16}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^{-1} = \frac{9}{16}$$

$$\frac{11}{8} \times 2^{-1} = \frac{11}{16}$$

$$\frac{10}{8} \times 2^{-1} = \frac{10}{16}$$

$$\frac{12}{8} \times 2^{-1}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^{-1} = \frac{9}{16}$$

$$\frac{11}{8} \times 2^{-1} = \frac{11}{16}$$

$$\frac{10}{8} \times 2^{-1} = \frac{10}{16}$$

$$\frac{12}{8} \times 2^{-1} = \frac{12}{16}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^{-1} = \frac{9}{16}$$

$$\frac{11}{8} \times 2^{-1} = \frac{11}{16}$$

$$\frac{13}{8} \times 2^{-1}$$

$$\frac{10}{8} \times 2^{-1} = \frac{10}{16}$$

$$\frac{12}{8} \times 2^{-1} = \frac{12}{16}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^{-1} = \frac{9}{16}$$

$$\frac{11}{8} \times 2^{-1} = \frac{11}{16}$$

$$\frac{13}{8} \times 2^{-1} = \frac{13}{16}$$

$$\frac{10}{8} \times 2^{-1} = \frac{10}{16}$$

$$\frac{12}{8} \times 2^{-1} = \frac{12}{16}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^{-1} = \frac{9}{16}$$

$$\frac{11}{8} \times 2^{-1} = \frac{11}{16}$$

$$\frac{13}{8} \times 2^{-1} = \frac{13}{16}$$

$$\frac{10}{8} \times 2^{-1} = \frac{10}{16}$$

$$\frac{12}{8} \times 2^{-1} = \frac{12}{16}$$

$$\frac{14}{8} \times 2^{-1}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^{-1} = \frac{9}{16}$$

$$\frac{11}{8} \times 2^{-1} = \frac{11}{16}$$

$$\frac{13}{8} \times 2^{-1} = \frac{13}{16}$$

$$\frac{10}{8} \times 2^{-1} = \frac{10}{16}$$

$$\frac{12}{8} \times 2^{-1} = \frac{12}{16}$$

$$\frac{14}{8} \times 2^{-1} = \frac{14}{16}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^{-1} = \frac{9}{16}$$

$$\frac{11}{8} \times 2^{-1} = \frac{11}{16}$$

$$\frac{13}{8} \times 2^{-1} = \frac{13}{16}$$

$$\frac{15}{8} \times 2^{-1}$$

$$\frac{10}{8} \times 2^{-1} = \frac{10}{16}$$

$$\frac{12}{8} \times 2^{-1} = \frac{12}{16}$$

$$\frac{14}{8} \times 2^{-1} = \frac{14}{16}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^{-1} , the first representable float is

$$0\ 01\ 000 = +1 \times 2^{-1} = \frac{1}{2} = \frac{8}{16}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^{-1} = \frac{9}{16}$$

$$\frac{11}{8} \times 2^{-1} = \frac{11}{16}$$

$$\frac{13}{8} \times 2^{-1} = \frac{13}{16}$$

$$\frac{15}{8} \times 2^{-1} = \frac{15}{16}$$

$$\frac{10}{8} \times 2^{-1} = \frac{10}{16}$$

$$\frac{12}{8} \times 2^{-1} = \frac{12}{16}$$

$$\frac{14}{8} \times 2^{-1} = \frac{14}{16}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

- The other representable float point numbers at this scale are

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

- The other representable float point numbers at this scale are

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^0$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^0 = \frac{9}{8}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^0 = \frac{9}{8}$$

$$\frac{10}{8} \times 2^0$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^0 = \frac{9}{8}$$

$$\frac{10}{8} \times 2^0 = \frac{10}{8}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

- The other representable float point numbers at this scale are

$$\frac{9}{8} \times 2^0 = \frac{9}{8}$$
$$\frac{11}{8} \times 2^0$$

$$\frac{10}{8} \times 2^0 = \frac{10}{8}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^0 &= \frac{9}{8} \\ \frac{11}{8} \times 2^0 &= \frac{11}{8}\end{aligned}$$

$$\frac{10}{8} \times 2^0 = \frac{10}{8}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^0 &= \frac{9}{8} \\ \frac{11}{8} \times 2^0 &= \frac{11}{8}\end{aligned}$$

$$\begin{aligned}\frac{10}{8} \times 2^0 &= \frac{10}{8} \\ \frac{12}{8} \times 2^0 &= \frac{12}{8}\end{aligned}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^0 &= \frac{9}{8} \\ \frac{11}{8} \times 2^0 &= \frac{11}{8}\end{aligned}$$

$$\begin{aligned}\frac{10}{8} \times 2^0 &= \frac{10}{8} \\ \frac{12}{8} \times 2^0 &= \frac{12}{8}\end{aligned}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^0 &= \frac{9}{8} \\ \frac{11}{8} \times 2^0 &= \frac{11}{8} \\ \frac{13}{8} \times 2^0 &= \frac{13}{8}\end{aligned}$$

$$\begin{aligned}\frac{10}{8} \times 2^0 &= \frac{10}{8} \\ \frac{12}{8} \times 2^0 &= \frac{12}{8}\end{aligned}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^0 &= \frac{9}{8} \\ \frac{11}{8} \times 2^0 &= \frac{11}{8} \\ \frac{13}{8} \times 2^0 &= \frac{13}{8}\end{aligned}$$

$$\begin{aligned}\frac{10}{8} \times 2^0 &= \frac{10}{8} \\ \frac{12}{8} \times 2^0 &= \frac{12}{8}\end{aligned}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^0 &= \frac{9}{8} \\ \frac{11}{8} \times 2^0 &= \frac{11}{8} \\ \frac{13}{8} \times 2^0 &= \frac{13}{8}\end{aligned}$$

$$\begin{aligned}\frac{10}{8} \times 2^0 &= \frac{10}{8} \\ \frac{12}{8} \times 2^0 &= \frac{12}{8} \\ \frac{14}{8} \times 2^0 &= \frac{14}{8}\end{aligned}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^0 &= \frac{9}{8} \\ \frac{11}{8} \times 2^0 &= \frac{11}{8} \\ \frac{13}{8} \times 2^0 &= \frac{13}{8}\end{aligned}$$

$$\begin{aligned}\frac{10}{8} \times 2^0 &= \frac{10}{8} \\ \frac{12}{8} \times 2^0 &= \frac{12}{8} \\ \frac{14}{8} \times 2^0 &= \frac{14}{8}\end{aligned}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

- The other representable float point numbers at this scale are

$$\begin{aligned}\frac{9}{8} \times 2^0 &= \frac{9}{8} \\ \frac{11}{8} \times 2^0 &= \frac{11}{8} \\ \frac{13}{8} \times 2^0 &= \frac{13}{8} \\ \frac{15}{8} \times 2^0 &= \frac{15}{8}\end{aligned}$$

$$\begin{aligned}\frac{10}{8} \times 2^0 &= \frac{10}{8} \\ \frac{12}{8} \times 2^0 &= \frac{12}{8} \\ \frac{14}{8} \times 2^0 &= \frac{14}{8}\end{aligned}$$

A SMALLER EXAMPLE: THE INTERMEDIATE SCALES

- At the intermediate scale, 2^0 , the first representable float is

$$0\ 10\ 000 = +1 \times 2^0 = 1 = \frac{8}{8}$$

- The other representable float point numbers at this scale are

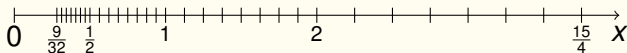
$$\begin{aligned}\frac{9}{8} \times 2^0 &= \frac{9}{8} \\ \frac{11}{8} \times 2^0 &= \frac{11}{8} \\ \frac{13}{8} \times 2^0 &= \frac{13}{8} \\ \frac{15}{8} \times 2^0 &= \frac{15}{8}\end{aligned}$$

$$\begin{aligned}\frac{10}{8} \times 2^0 &= \frac{10}{8} \\ \frac{12}{8} \times 2^0 &= \frac{12}{8} \\ \frac{14}{8} \times 2^0 &= \frac{14}{8}\end{aligned}$$

A SMALLER EXAMPLE: THE LINE GRAPH

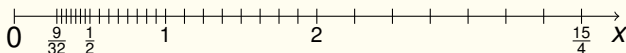
A SMALLER EXAMPLE: THE LINE GRAPH

- A number line graph for all of the non-negative representable float point numbers is



A SMALLER EXAMPLE: THE LINE GRAPH

- A number line graph for all of the non-negative representable float point numbers is



- Any computed number x between these representable float point numbers will be rounded-off

A SMALLER EXAMPLE: FLOATING POINT ERRORS

A SMALLER EXAMPLE: FLOATING POINT ERRORS

- Floating point arithmetic is not precise: Errors will be made

A SMALLER EXAMPLE: FLOATING POINT ERRORS

- Floating point arithmetic is not precise: Errors will be made
- A novice programmer mistake is to test a floating point number against 0

```
if (x == 0.0) ...
```

A SMALLER EXAMPLE: FLOATING POINT ERRORS

- Floating point arithmetic is not precise: Errors will be made
- A novice programmer mistake is to test a floating point number against 0

```
if (x == 0.0) ...
```

- Equivalently: Testing if two computed float point numbers are equal is a mistake

```
if (x == y) ...
```

A SMALLER EXAMPLE: FLOATING POINT ERRORS

- Floating point arithmetic is not precise: Errors will be made
- A novice programmer mistake is to test a floating point number against 0

```
if (x == 0.0) ...
```

- Equivalently: Testing if two computed float point numbers are equal is a mistake

```
if (x == y) ...
```

- Such a test can never be more precise than the **machine epsilon**

A SMALLER EXAMPLE: FLOATING POINT ERRORS

A SMALLER EXAMPLE: FLOATING POINT ERRORS

- The machine epsilon is the largest value $\varepsilon = 2^e$ such that

$$1.0 + \varepsilon = 1.0$$

A SMALLER EXAMPLE: FLOATING POINT ERRORS

- The machine epsilon is the largest value $\varepsilon = 2^e$ such that

$$1.0 + \varepsilon = 1.0$$

- For our small example

$$\varepsilon = \frac{1}{16}$$

A SMALLER EXAMPLE: FLOATING POINT ERRORS

- The machine epsilon is the largest value $\varepsilon = 2^e$ such that

$$1.0 + \varepsilon = 1.0$$

- For our small example

$$\varepsilon = \frac{1}{16}$$

- Note that $1 + \frac{1}{8} = \frac{9}{8}$ is a representable float

A SMALLER EXAMPLE: FLOATING POINT ERRORS

- The machine epsilon is the largest value $\varepsilon = 2^e$ such that

$$1.0 + \varepsilon = 1.0$$

- For our small example

$$\varepsilon = \frac{1}{16}$$

- Note that $1 + \frac{1}{8} = \frac{9}{8}$ is a representable float
- But $1 + \frac{1}{16} = \frac{17}{16}$ is not representable, and the sum will be rounded to 1

A SMALLER EXAMPLE: FLOATING POINT ERRORS

A SMALLER EXAMPLE: FLOATING POINT ERRORS

- Floating point arithmetic does not obey common arithmetic principles

A SMALLER EXAMPLE: FLOATING POINT ERRORS

- Floating point arithmetic does not obey common arithmetic principles
- For example, floating point arithmetic is not associative

A SMALLER EXAMPLE: FLOATING POINT ERRORS

- Floating point arithmetic does not obey common arithmetic principles
- For example, floating point arithmetic is not associative
 - The sum

$$1 + \left(\frac{5}{16} + \frac{5}{16} \right) = 1 + \left(\frac{10}{16} \right) = \frac{26}{16}$$

is computed correctly: All values are representable floating point numbers

A SMALLER EXAMPLE: FLOATING POINT ERRORS

- Floating point arithmetic does not obey common arithmetic principles
- For example, floating point arithmetic is not associative
 - The sum

$$1 + \left(\frac{5}{16} + \frac{5}{16} \right) = 1 + \left(\frac{10}{16} \right) = \frac{26}{16}$$

is computed correctly: All values are representable floating point numbers

- However, the sum

$$\left(1 + \frac{5}{16} \right) + \frac{5}{16} = \left(\frac{20}{16} \right) + \frac{5}{16} = \frac{24}{16}$$

is *not* computed correctly

A SMALLER EXAMPLE: FLOATING POINT ERRORS

- Floating point arithmetic does not obey common arithmetic principles
- For example, floating point arithmetic is not associative
 - The sum

$$1 + \left(\frac{5}{16} + \frac{5}{16} \right) = 1 + \left(\frac{10}{16} \right) = \frac{26}{16}$$

is computed correctly: All values are representable floating point numbers

- However, the sum

$$\left(1 + \frac{5}{16} \right) + \frac{5}{16} = \left(\frac{20}{16} \right) + \frac{5}{16} = \frac{24}{16}$$

is *not* computed correctly

- $1 + \frac{5}{16} = \frac{21}{16}$ is not representable and rounded to $\frac{20}{16}$

A SMALLER EXAMPLE: FLOATING POINT ERRORS

- Floating point arithmetic does not obey common arithmetic principles
- For example, floating point arithmetic is not associative
 - The sum

$$1 + \left(\frac{5}{16} + \frac{5}{16} \right) = 1 + \left(\frac{10}{16} \right) = \frac{26}{16}$$

is computed correctly: All values are representable floating point numbers

- However, the sum

$$\left(1 + \frac{5}{16} \right) + \frac{5}{16} = \left(\frac{20}{16} \right) + \frac{5}{16} = \frac{24}{16}$$

is *not* computed correctly

- $1 + \frac{5}{16} = \frac{21}{16}$ is not representable and rounded to $\frac{20}{16}$
- $\frac{20}{16} + \frac{5}{16} = \frac{25}{16}$ is not representable and rounded to $\frac{24}{16}$

IEEE 754 FLOATING POINT STANDARD

IEEE 754 FLOATING POINT STANDARD

- The IEEE 754 Floating Point Standard provides a model to make errors minimal and provide consistency across programming languages and computer architectures

IEEE 754 FLOATING POINT STANDARD

- The IEEE 754 Floating Point Standard provides a model to make errors minimal and provide consistency across programming languages and computer architectures
- A good summary by David Goldberg is

What Every Computer Scientist Should Know About
Floating Point Arithmetic

IEEE 754 FLOATING POINT STANDARD

- The IEEE 754 Floating Point Standard provides a model to make errors minimal and provide consistency across programming languages and computer architectures
- A good summary by David Goldberg is

What Every Computer Scientist Should Know About
Floating Point Arithmetic

- The IEEE standard defines four different precisions

IEEE 754 FLOATING POINT STANDARD

- The IEEE 754 Floating Point Standard provides a model to make errors minimal and provide consistency across programming languages and computer architectures
- A good summary by David Goldberg is

What Every Computer Scientist Should Know About Floating Point Arithmetic

- The IEEE standard defines four different precisions
- Most hardware provides for single (type `float`) and double precision (type `double`)

IEEE SINGLE PRECISION

IEEE SINGLE PRECISION

- The format parameters for single precision are

IEEE SINGLE PRECISION

- The format parameters for single precision are
 - Word length: 32 bits

IEEE SINGLE PRECISION

- The format parameters for single precision are
 - Word length: 32 bits
 - 1 sign bit

IEEE SINGLE PRECISION

- The format parameters for single precision are
 - Word length: 32 bits
 - 1 sign bit
 - 8 exponent bits

IEEE SINGLE PRECISION

- The format parameters for single precision are
 - Word length: 32 bits
 - 1 sign bit
 - 8 exponent bits
 - 23 fractional part bits

IEEE SINGLE PRECISION

- The format parameters for single precision are
 - Word length: 32 bits
 - 1 sign bit
 - 8 exponent bits
 - 23 fractional part bits
- The exponent bias is 127

IEEE SINGLE PRECISION

- The format parameters for single precision are
 - Word length: 32 bits
 - 1 sign bit
 - 8 exponent bits
 - 23 fractional part bits
- The exponent bias is 127
 - The unbiased exponent range is from -127 to 128

IEEE SINGLE PRECISION

- The format parameters for single precision are
 - Word length: 32 bits
 - 1 sign bit
 - 8 exponent bits
 - 23 fractional part bits
- The exponent bias is 127
 - The unbiased exponent range is from -127 to 128
 - But the exponents -127 and 128 are used for special qualities

IEEE SINGLE PRECISION

- The format parameters for single precision are
 - Word length: 32 bits
 - 1 sign bit
 - 8 exponent bits
 - 23 fractional part bits
- The exponent bias is 127
 - The unbiased exponent range is from -127 to 128
 - But the exponents -127 and 128 are used for special qualities
 - The practical range of exponents is from $e_{\min} = -126$ to $e_{\max} = 127$

IEEE SINGLE PRECISION

- The format parameters for single precision are
 - Word length: 32 bits
 - 1 sign bit
 - 8 exponent bits
 - 23 fractional part bits
- The exponent bias is 127
 - The unbiased exponent range is from -127 to 128
 - But the exponents -127 and 128 are used for special qualities
 - The practical range of exponents is from $e_{\min} = -126$ to $e_{\max} = 127$
 - Since $2^{10} \approx 10^3$, floats can scale from about

$$2^{-126} = (2^{10})^{-12.6} \approx (10^3)^{-12.6} \approx 10^{-38}$$

to

$$2^{127} = (2^{10})^{12.7} \approx (10^3)^{12.7} \approx 10^{38}$$

IEEE SINGLE PRECISION

IEEE SINGLE PRECISION

- Because of the hidden bit in $(1.f)$, there are 24 bits of precision in `floats`

IEEE SINGLE PRECISION

- Because of the hidden bit in $(1.f)$, there are 24 bits of precision in `floats`
- Since $2^{10} \approx 10^3$, 24 bits of precision gives about 7 digits of decimal precision

IEEE SINGLE PRECISION

- Because of the hidden bit in $(1.f)$, there are 24 bits of precision in `floats`
- Since $2^{10} \approx 10^3$, 24 bits of precision gives about 7 digits of decimal precision
- That is

$$2^{24} = (2^{10})^{2.4} \approx (10^3)^{2.4} \approx 10^7$$

IEEE SINGLE PRECISION

- Because of the hidden bit in $(1.f)$, there are 24 bits of precision in `floats`
- Since $2^{10} \approx 10^3$, 24 bits of precision gives about 7 digits of decimal precision

- That is

$$2^{24} = (2^{10})^{2.4} \approx (10^3)^{2.4} \approx 10^7$$

- At the smallest scale, `floats` are separated by increments of

$$2^{-23} \times 2^{-126} = 2^{-149}$$

IEEE SINGLE PRECISION

- Because of the hidden bit in $(1.f)$, there are 24 bits of precision in `floats`
- Since $2^{10} \approx 10^3$, 24 bits of precision gives about 7 digits of decimal precision

- That is

$$2^{24} = (2^{10})^{2.4} \approx (10^3)^{2.4} \approx 10^7$$

- At the smallest scale, `floats` are separated by increments of

$$2^{-23} \times 2^{-126} = 2^{-149}$$

- At the largest scale, `floats` are separated by increments of

$$2^{-23} \times 2^{127} = 2^{104}$$

IEEE DOUBLE PRECISION

IEEE DOUBLE PRECISION

- The format parameters for double precision are

IEEE DOUBLE PRECISION

- The format parameters for double precision are
 - Word length: 64 bits

IEEE DOUBLE PRECISION

- The format parameters for double precision are
 - Word length: 64 bits
 - 1 sign bit

IEEE DOUBLE PRECISION

- The format parameters for double precision are
 - Word length: 64 bits
 - 1 sign bit
 - 11 exponent bits

IEEE DOUBLE PRECISION

- The format parameters for double precision are
 - Word length: 64 bits
 - 1 sign bit
 - 11 exponent bits
 - 52 fractional part bits

IEEE DOUBLE PRECISION

- The format parameters for double precision are
 - Word length: 64 bits
 - 1 sign bit
 - 11 exponent bits
 - 52 fractional part bits
- The exponent bias is 1023

IEEE DOUBLE PRECISION

- The format parameters for double precision are
 - Word length: 64 bits
 - 1 sign bit
 - 11 exponent bits
 - 52 fractional part bits
- The exponent bias is 1023
 - The unbiased exponent range is from -1023 to 1024

IEEE DOUBLE PRECISION

- The format parameters for double precision are
 - Word length: 64 bits
 - 1 sign bit
 - 11 exponent bits
 - 52 fractional part bits
- The exponent bias is 1023
 - The unbiased exponent range is from -1023 to 1024
 - The exponents -1023 and 1024 are used for special qualities

IEEE DOUBLE PRECISION

- The format parameters for double precision are
 - Word length: 64 bits
 - 1 sign bit
 - 11 exponent bits
 - 52 fractional part bits
- The exponent bias is 1023
 - The unbiased exponent range is from -1023 to 1024
 - The exponents -1023 and 1024 are used for special qualities
 - The practical range of exponents is from $e_{\min} = -1022$ to $e_{\max} = 1023$

IEEE DOUBLE PRECISION

- The format parameters for double precision are
 - Word length: 64 bits
 - 1 sign bit
 - 11 exponent bits
 - 52 fractional part bits
- The exponent bias is 1023
 - The unbiased exponent range is from -1023 to 1024
 - The exponents -1023 and 1024 are used for special qualities
 - The practical range of exponents is from $e_{\min} = -1022$ to $e_{\max} = 1023$
 - Since $2^{10} \approx 10^3$, `doubles` can scale from about

$$2^{-1022} = (2^{10})^{-102.2} \approx (10^3)^{-102.2} \approx 10^{-306}$$

to

$$2^{1023} = (2^{10})^{102.3} \approx (10^3)^{102.3} \approx 10^{309}$$

IEEE DOUBLE PRECISION

IEEE DOUBLE PRECISION

- Because of the hidden bit in $(1.f)$, there are 53 bits of precision in `doubles`

IEEE DOUBLE PRECISION

- Because of the hidden bit in $(1.f)$, there are 53 bits of precision in `doubles`
- Since $2^{10} \approx 10^3$, 53 bits of precision gives about 16 digits of decimal precision

IEEE DOUBLE PRECISION

- Because of the hidden bit in $(1.f)$, there are 53 bits of precision in `doubles`
- Since $2^{10} \approx 10^3$, 53 bits of precision gives about 16 digits of decimal precision
- That is

$$2^{53} = (2^{10})^{5.3} \approx (10^3)^{5.3} \approx 10^{16}$$

IEEE DOUBLE PRECISION

- Because of the hidden bit in $(1.f)$, there are 53 bits of precision in `doubles`
- Since $2^{10} \approx 10^3$, 53 bits of precision gives about 16 digits of decimal precision

- That is

$$2^{53} = (2^{10})^{5.3} \approx (10^3)^{5.3} \approx 10^{16}$$

- At the smallest scale, `doubles` are separated by increments of

$$2^{-52} \times 2^{-1022} = 2^{-1074}$$

IEEE DOUBLE PRECISION

- Because of the hidden bit in $(1.f)$, there are 53 bits of precision in `doubles`
- Since $2^{10} \approx 10^3$, 53 bits of precision gives about 16 digits of decimal precision

- That is

$$2^{53} = (2^{10})^{5.3} \approx (10^3)^{5.3} \approx 10^{16}$$

- At the smallest scale, `doubles` are separated by increments of

$$2^{-52} \times 2^{-1022} = 2^{-1074}$$

- At the largest scale, `doubles` are separated by increments of

$$2^{-52} \times 2^{1023} = 2^{971}$$

IEEE SPECIAL QUALITIES

IEEE SPECIAL QUALITIES

- When the exponent e is at its extremes, special qualities of the number are defined

IEEE SPECIAL QUALITIES

- When the exponent e is at its extremes, special qualities of the number are defined
 - Recall, for single precision the exponent e can range from $e_{\min} - 1 = -127$ to $e_{\max} + 1 = 128$

IEEE SPECIAL QUALITIES

- When the exponent e is at its extremes, special qualities of the number are defined
 - Recall, for single precision the exponent e can range from $e_{\min} - 1 = -127$ to $e_{\max} + 1 = 128$
 - For double precision the exponent can range from $e_{\min} - 1 = -1023$ to $e_{\max} + 1 = 1024$

- When the exponent e is at its extremes, special qualities of the number are defined
 - Recall, for single precision the exponent e can range from $e_{\min} - 1 = -127$ to $e_{\max} + 1 = 128$
 - For double precision the exponent can range from $e_{\min} - 1 = -1023$ to $e_{\max} + 1 = 1024$
- When the exponent takes on its most negative value

- When the exponent e is at its extremes, special qualities of the number are defined
 - Recall, for single precision the exponent e can range from $e_{\min} - 1 = -127$ to $e_{\max} + 1 = 128$
 - For double precision the exponent can range from $e_{\min} - 1 = -1023$ to $e_{\max} + 1 = 1024$
- When the exponent takes on its most negative value
 - The float represents 0 if the fractional part f is 0

- When the exponent e is at its extremes, special qualities of the number are defined
 - Recall, for single precision the exponent e can range from $e_{\min} - 1 = -127$ to $e_{\max} + 1 = 128$
 - For double precision the exponent can range from $e_{\min} - 1 = -1023$ to $e_{\max} + 1 = 1024$
- When the exponent takes on its most negative value
 - The float represents 0 if the fractional part f is 0
 - The float represents the denormalized number $\pm(0.f) \times 2^{e_{\min}}$ if $f \neq 0$

- When the exponent e is at its extremes, special qualities of the number are defined
 - Recall, for single precision the exponent e can range from $e_{\min} - 1 = -127$ to $e_{\max} + 1 = 128$
 - For double precision the exponent can range from $e_{\min} - 1 = -1023$ to $e_{\max} + 1 = 1024$
- When the exponent takes on its most negative value
 - The float represents 0 if the fractional part f is 0
 - The float represents the denormalized number $\pm(0.f) \times 2^{e_{\min}}$ if $f \neq 0$
- When the exponent takes on its most positive value

- When the exponent e is at its extremes, special qualities of the number are defined
 - Recall, for single precision the exponent e can range from $e_{\min} - 1 = -127$ to $e_{\max} + 1 = 128$
 - For double precision the exponent can range from $e_{\min} - 1 = -1023$ to $e_{\max} + 1 = 1024$
- When the exponent takes on its most negative value
 - The float represents 0 if the fractional part f is 0
 - The float represents the denormalized number $\pm(0.f) \times 2^{e_{\min}}$ if $f \neq 0$
- When the exponent takes on its most positive value
 - The float represents infinity (∞) if the fractional part f is 0

- When the exponent e is at its extremes, special qualities of the number are defined
 - Recall, for single precision the exponent e can range from $e_{\min} - 1 = -127$ to $e_{\max} + 1 = 128$
 - For double precision the exponent can range from $e_{\min} - 1 = -1023$ to $e_{\max} + 1 = 1024$
- When the exponent takes on its most negative value
 - The float represents 0 if the fractional part f is 0
 - The float represents the denormalized number $\pm(0.f) \times 2^{e_{\min}}$ if $f \neq 0$
- When the exponent takes on its most positive value
 - The float represents infinity (∞) if the fractional part f is 0
 - The float represents “Not a Number” (NaN) if $f \neq 0$

HOW FINE IS FLOATING-POINT ARITHMETIC?

HOW FINE IS FLOATING-POINT ARITHMETIC?

- Here Nick Trefethen's answer to the density of floating point numbers

HOW FINE IS FLOATING-POINT ARITHMETIC?

- Here Nick Trefethen's answer to the density of floating point numbers
- According to molecular physics, there are approximately 3×10^8 molecules per meter in a gas at atmospheric conditions

HOW FINE IS FLOATING-POINT ARITHMETIC?

- Here Nick Trefethen's answer to the density of floating point numbers
- According to molecular physics, there are approximately 3×10^8 molecules per meter in a gas at atmospheric conditions
 - Essentially the cube root of Avogadro's number: 6.02×10^{23}

HOW FINE IS FLOATING-POINT ARITHMETIC?

- Here Nick Trefethen's answer to the density of floating point numbers
- According to molecular physics, there are approximately 3×10^8 molecules per meter in a gas at atmospheric conditions
 - Essentially the cube root of Avogadro's number: 6.02×10^{23}
- The circumference of the earth is 4×10^7 meters,

HOW FINE IS FLOATING-POINT ARITHMETIC?

HOW FINE IS FLOATING-POINT ARITHMETIC?

- So in a circle around the earth, there are around $(3 \times 10^8) \times (4 \times 10^7) \approx 10^{16}$ molecules

HOW FINE IS FLOATING-POINT ARITHMETIC?

- So in a circle around the earth, there are around $(3 \times 10^8) \times (4 \times 10^7) \approx 10^{16}$ molecules
- In IEEE double precision arithmetic, there are 2^{52} numbers between 1 and 2, which is also about 10^{16}

HOW FINE IS FLOATING-POINT ARITHMETIC?

- So in a circle around the earth, there are around $(3 \times 10^8) \times (4 \times 10^7) \approx 10^{16}$ molecules
- In IEEE double precision arithmetic, there are 2^{52} numbers between 1 and 2, which is also about 10^{16}
- So if we put a giant circle around the earth with a distance coordinate ranging from 1 to 2, the spacing of the floating-point numbers along the circle will be about the same as the spacing of the air molecules.

