MSiA-413 Introduction to Databases and Information Retrieval

Lecture 2
Fixed-point and Floating-point Representations

Instructor: Nikos Hardavellas

Slides adapted from Steve Tarzia

Last week we talked about Integers

- Integers can be stored in a base-two positional notation in binary
- Addition and subtraction follow the familiar mechanics
- Learned some tricks (eg., $2^{10} = 1024 \approx 1000$)
- Signed integers use two's complement representation for negatives
  - Two's complement makes subtraction just as easy as addition
  - Positive numbers are represented in the same way whether you're using a signed or unsigned data type, but
  - Small negatives and huge positives can be confused if you misinterpret the type

A few more things about integers

- Multiplication: two's complement works magically here too
- Positive division works as expected
- "*Sign extension*:" when increasing the "bit size" of a negative number, add leading ones
  - Eg., -2 is `1110` as a 4-bit signed integer and `11111110` in 8 bits
- Computers typically use 32 or 64 bit integers

Any questions on last week's material?

Integers are great for **counting**, but sometimes we need to **measure** fractional quantities

Binary numbers can have "decimal" places, too

- $\texttt{0.1111111111}_{\text{two}}$ is slightly smaller than 1
- $\texttt{0.0000000001}_{\text{two}}$ is slightly larger than 0
- $\texttt{0.1}_{\text{two}}$ is one half

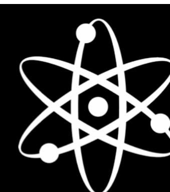
- $\texttt{10.101}_{\text{two}} = 1 \times 2^1 + 0 \times 2^0 + 1 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3}$
  $= 2 + 0 + 1/2 + 0 + 1/8 = 2 = \texttt{2.625}_{\text{ten}}$

How shall we represent fractional number in the computer?

Fixed point: Integers 2.0

- Simplest solution is to just stick an implicit **binary (radix) point** somewhere (We don't call it a decimal point because we're not in base ten)
- Examples of fixed point numbers in base ten:
  - Represent the cost of a purchase with an integer number of cents
    - The cost of a sandwich is 625 cents ($6.25)
  - Represent the distance between cities by counting the hundredths of a mile
    - Evanston is 1321 hundredths of a mile from Chicago (13.21 miles)
    - and 79,543 hundredths of a mile from Philadelphia

Fixed point example in 16 bits



Let's store the chemical elements' atomic weights

- Smallest value (hydrogen) is 1.00784
- Largest value (uranium) is 238.02891
- Negative values are not possible
- We can reserve 8 bits for the fractional part and 8 bits for the part > 1
- In this particular binary fixed point representation, the weight of uranium is:
  1110111000000111_2 Remember that the radix point is implicit. This represents the value
  $11101110.00000111_2 = 238_{10} = 238.02734375_{10}$
  (We had to round off, so this is not precisely accurate)
- And the weight of hydrogen is:
  `0000000100000010,`
  i.e., `00000001.00000010` = 1 = 1.0078125

Fixed point is simple & efficient but it has its limitations

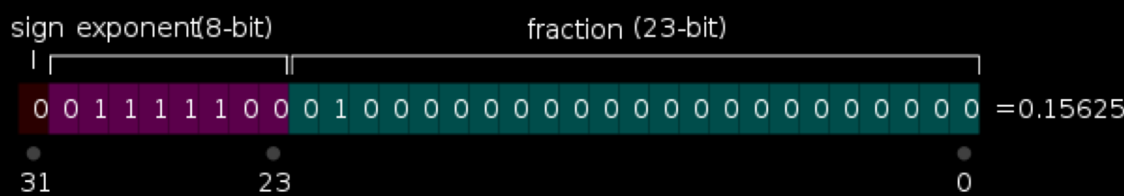
- Range is very limited
  - Multiplication overflows easily – can double the number of bits
    - E.g., if working in 32-bits, then we can only multiply 16-bit values without overflow
  - Division **underflows** easily (small values are rounded to zero)
- Precision varies across the range:
  - Small numbers have few significant figures

Floating point

- Based on scientific notation:
  - $10{,}340 = 1.034 \times 10^{4}$
  - $0.00424 = 4.24 \times 10^{-3}$
- Gives a compact representation of extreme values:
  - $1{,}000{,}000{,}000{,}000{,}000{,}000{,}000{,}000 = 1.0 \times 10^{24}$
  - $0.000\,000\,000\,000\,000\,000\,000\,001 = 1.0 \times 10^{-24}$
- In binary:
  - $100010_{two} = 1.0001_{two} \times 2^{5}{}_{ten} = 1.0001 \times 10^{101}{}_{two}$
  - $0.00101_{two} = 1.01_{two} \times 2^{-3}{}_{ten} = 1.01 \times 10^{-11}{}_{two}$

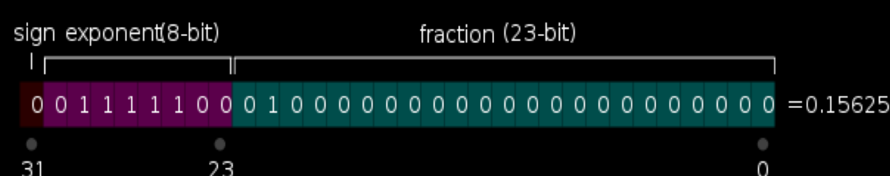
Representing floating point in bits

- $0.15625_{ten} = 0.00101_{two} = 1.01_{two} \times 2^{-3}{}_{ten} = 1.01 \times 10^{-11}{}_{two}$
- Three essential parts are the **sign**, **fraction**, & **exponent**
  - Notice that the first significant figure is always “1” so we don’t have to store it
- In the mid 1980s, the IEEE standardized the floating point representation of 32 and 64 bit numbers:
  - The exponent has a sign too, but the standard says to add a “bias” of 127



64-bit floating point

- Similar to 32-bit, but we have more precision in the fraction and larger exponents are possible
- 32-bit is called **single precision** and 64-bit is called **double precision**
- Double precision can represent larger, smaller, and more precise numbers



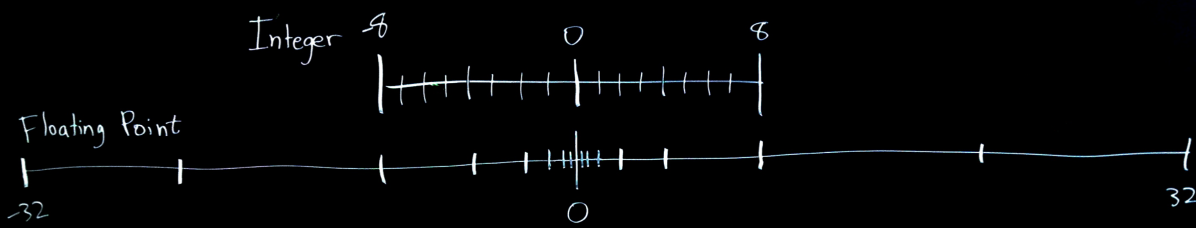
A few special floats

- The IEEE standard allows for a few special values to be stored
  - Positive and negative zero (remember that we normally start with an implied "1")
    - All exponent bits set to zeros
  - Positive and negative infinity (the result of divide by zero)
  - Not a number (the result of zero divided by zero)
    - These all have the exponent bits set to all ones

The Flexibility and Flaws of Floats

- A 32-bit signed integer can represent all the whole numbers between -2,147,483,648 and 2,147,483,647
- A 32-bit floating point number can be as large as $\pm 3.402823 \times 10^{38}$ = 340,282,300,000,000,000,000,000,000,000,000,000,000
- or as tiny as $5.8774718 \times 10^{-39}$ = 0.000 000 000 000 000 000 000 000 000 000 000 000 005 877 471 8
- But, single-precision floats have only 24 bits of precision:
  - Can only precisely store integers up to 2^{24} = 16,777,216
- Floats can store larger numbers than integers of the same bit-length, but with less precision because 8 bits are set aside for the exponent

Floats just distributed the same number of values differently – with exponential spacing



Know when to use integers, floating point, and fixed point

- When **counting** or labelling things, always use integers
- When **measuring** physical quantities, usually use floating point
  - May use fixed point if speed/simplicity is more important than accuracy
- If your machine does not support floating point (eg., a toaster):
  - Use fixed point representation for fractional quantities
- If rounding is desired then use fixed point (but carefully)
  - U.S. currency values usually should be rounded to the nearest cent
- Use 64-bit integers when you need values > 2 billion
- Floating point rules of thumb:
  - Single precision gives ~7 decimal digits of precision
  - Double precision gives ~16 decimal digits of precision

One more point about fractions in binary: Base ten decimals usually have to be rounded

- We all know that 1/3 cannot be represented exactly in decimal
  - That's because 10^x not divisible by 3 (for any integer x)
- Similarly, 1/10 cannot be represented exactly in binary
  - Because 2^x is not divisible by 10 (for any integer x)
- In general, a rational number ***a/b*** can be *exactly* represented in binary only if ***b*** is a power of 2
  - Otherwise, there is some rounding error
- Most fractions cannot be stored exactly with a finite number of bits
  - Actually, this is also true in decimal!
- So, always expect small rounding errors when working in floating point

How do computers work with floats?

- It's complicated and slow!
- Have to manipulate both the fraction and the exponent
- Addition is no longer simple