CS 520
Integer to Float Conversion

Convert Integer to Float – i2f

- Integer Value: $3456789A_{16}$ to IEEE single-precision

--> 0011 0100 0101 0110 0111 1000 1001

--> 0011 0100 0101 0110 0111 1000 1001 1010. x 2^0

--> 00 1.10100010101100111100010011010 x 2^{29}

Significand can only be 23 digits so we need to truncate (in this case**) extra bits

Leave it out 10100010101100111100010

Actual exponent = Stored exponent – 127 --> Stored exponent = 29 + 127 = 156

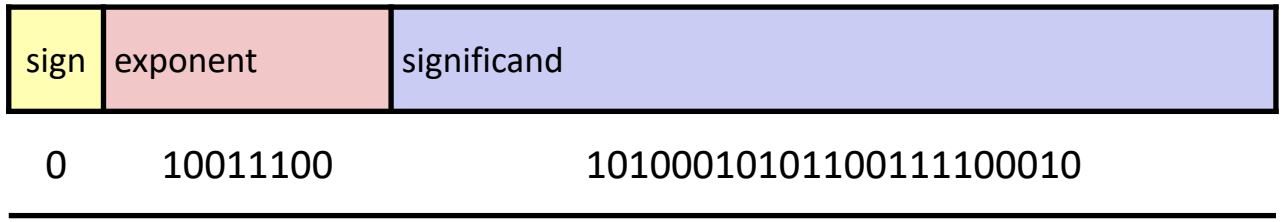
Exponent = 10011100_2

Number was positive

--> 01001110010100010101100111100 10

--> 0100 1110 0101 0001 0101 1100 1111 0010

--> 4 E 5 1 5 9 E 2



**truncate or round?

Convert Integer to Float - Rounding

- From the previous example:
- Significand can only be 23 bits so truncated the last six:

$$1.10100010101100111100010\underline{01}\ \underline{1010} \times 2^{29}$$

Guard Bits

Summarized as Sticky Bit

If all bits are zero it is 0

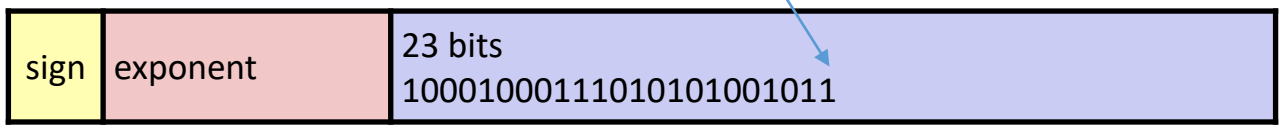
Otherwise it is 1

Convert Integer to Float - Rounding

- ***Round** to nearest, ties to **even** – **rounds** to the nearest value; if the number falls midway, it is **rounded** to the nearest value with an **even** least significant digit; this is the default for binary **floating point** and the recommended default for decimal*

Guard Bits (Two highest bits to be discarded)	Sticky Bit (Summarizes all other discarded bits)	Action
00	0 or 1	Truncate
01	0 or 1	Truncate
10 10	0 1	Round to even* Add one (to the LSB)
11	0 or 1	Add one (to the LSB)

*If the LSB is 0 leave it alone, if the LSB
Is odd, add 1 to it and make it even



Convert Integer to Float - Rounding

- Truncate
  - Just discard the bits
- Round to even (you will end up with an even number)
  - If low bit in significand is 1, add 1
  - Else do nothing
- Add one
  - Add 1 to significand

- Keep in mind that adding 1 might need re-normalization if all the bits in the significand are 1 which will carry out the last 1 !
- 1.1111….1 +1 = 10.000…0 --> shift it right and adjust exponent!

Convert Integer to Float - Rounding

- Try on your own
- 0x345678A0 --> 4E515IE2

  guard bits: 10

  sticky bit: 0

  round to even

  low bit is 0 (i.e. already even), so nothing done

Convert Integer to Float - Rounding

- Try on your own
- 0x345678E0 --> 4E5159E4

  guard bits: 10

  sticky bit: 0

  round to even

  low bit is 1 (i.e. odd), add 1 to that position

Convert Integer to Float - Rounding

- Try on your own
- 0x345678B0 --> 4E5159E3

  guard bits: 10

  sticky bit: 1

  add 1 to low bit position

Convert Integer to Float - Rounding

- Try on your own
- 0x345678B0 --> 4E5159E3

  guard bits: 11

  sticky bit: 0

  add 1 to low bit position