UNM SCHOOL of ENGINEERING
Department of Computer Science

CS341
CDR(5):
Conclusions on CDR of int and FP

Lecture #6 - part 2

Prof. Soraya Abad-Mota, PhD

Conclusions about FP representation

- FP in C
- Puzzles
- Summary

Floating Point in C (sec. 2.4.6)

- C Guarantees Two Levels (read/understand p.125 bullets)
  - **`float`** single precision
  - **`double`** double precision
- Conversions/Casting
  - Casting between **`int`**, **`float`**, and **`double`** changes bit representation
  - **`double`**/**`float`** → **`int`**
    - Truncates fractional part
    - Like rounding toward zero
    - Not defined when out of range or NaN: Generally sets to TMin
  - **`int`** → **`double`**
    - Exact conversion, as long as **`int`** has ≤ 53 bit word size
  - **`int`** → **`float`**
    - Will round according to rounding mode

Floating Point Puzzles (2.54) (on your own)

- For each of the following C expressions, either:
  - Argue that it is true for all argument values, or
  - explain why not true

```
int x = …;
float f = …;
double d = …;
```

Assume neither
d nor **f** is NaN

```
1. x == (int)(float) x
2. x == (int)(double) x
3. f == (float)(double) f
4. d == (double)(float) d
5. f == -(-f);
6. 2/3 == 2/3.0
7. d < 0.0      ⇒   ((d*2) < 0.0)
8. d > f        ⇒   -f > -d
9. d * d >= 0.0
10.(d+f)-d == f
```

Summary

- IEEE Floating Point has clear mathematical properties
  (standard makes independent of HW or SW)
- Represents numbers of form M x 2^E
- One can reason about operations independent of implementation
  - As if computed with perfect precision and then rounded
- Not the same as real arithmetic
  - Violates associativity/distributivity
  - Makes life difficult for compilers & serious numerical applications programmers