# Floating Point

Ned Nedialkov

McMaster University
Canada

SE 3F03
March 2013

# Outline

Standards

FP registers

Data registers

Control register

Tag register

# Standards

- IEEE 754, 1985

  http://en.wikipedia.org/wiki/IEEE_floating_point
- IEEE 754, 2008

  http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4610935

- Main data formats: single and double
- FP numbers are of the form

$$\pm 1.b_1 \cdots b_m \times 2^e,$$

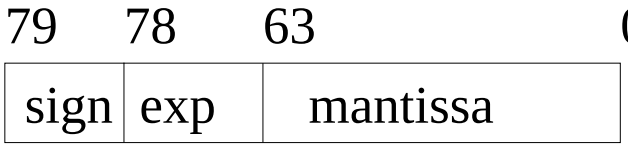
where $m = 23$ for single precision, and $m = 52$ for double precision
$e$ is exponent

# FP registers

- Data registers (FP data)
- Control registers (control of FP operations)
- Status register (status of FP operations)
- Pointer register (exception handlers)

# Data registers

- 8 FP data registers of the form




- Sign is 1 (negative), 0 positive
- Exponent is 15 bits
- Mantissa is 64 bits
- The stored FP numbers are of the form

$$\pm 1.b_1 \cdots b_{64} \times 2^{\text{exp}}$$

- This is a normalized number: the first digit is 1
- Denormalized numbers: later

- The data registers are organized as a stack, e.g.

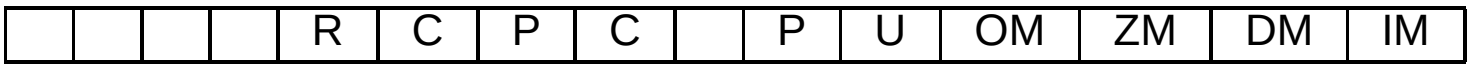
  ST5
  ST4
  ST3
  ST2
  ST1
  ST0 top of the stack (TOS)
  ST7
  ST6

# Control register

| | | | | R | C | P | C | | P | U | OM | ZM | DM | IM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

- RC rounding control
  - 00 neares
  - 01 towards $+\infty$
  - 10 towards $-\infty$
  - 11 truncate, towards 0
- PC precision control
  - 00 single
  - 01 not used
  - 10 double
  - extended

# Control register




- Exception masks
  - PM precision
  - UM underflow
  - OM overflow
  - ZM divide by 0
  - IM invalid
- When set, an exception does not occur

# Status register

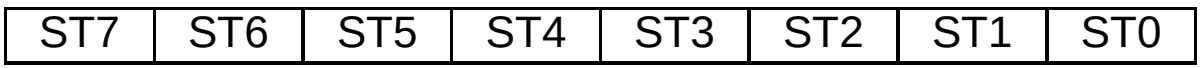
| B | C3 | T | O | S | C2 | C1 | C0 | ES | SF | P | U | O | Z | D | I |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

- B busy flag, for compatibility with 8087
- TOS top of stack
  - Flags
  - 

| FPU flag | CPU flag |
|---|---|
| C0 | CF |
| C2 | PF |
| C3 | ZF |

  - *C*1 stack overflow/underflow
  - ES error status
  - SF stack fault
  - The remaining are as in the control register

# Tag register




- Holds information about the content of a FP register
  - 00 valid
  - 01 zero
  - 10 special (invalid, $\infty$ or denormalized)
  - 11 empty