CSE4210
Architecture and Hardware for DSP

Floating Point Numbers

FP Numbers

- Fixed point representation has a small *Dynamic Range* (ratio of max to min, non-zero, numbers reprehensible)
- FPN +/- 0.M x $\beta^{+/-e}$
- 6 attributes define a FPN (some might be implicit)

FP Numbers

- Consider 4 digits, compare between fixed and floating point number. Also consider only positive numbers

1-to-9999	0.01 to $0.99*10^{99}$
Dynamic range $\approx 10^4$	Dynamic range 10^{101}

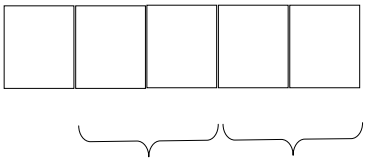
FPN

- The fraction is an unsigned number called *mantissa*, also called *significand* if not a fraction.
- The exponent is represented by a characteristic (excess representation).
- The number is *normalized* if the MSD of the mantissa is not a zero.

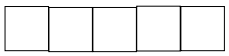
FPN

- Bias is added to the exponent to get the characteristic
- Biased is 0.5* β^s, where s is the number of bits in the exponent.
- Although there is no unique representation of zero, it is usually represented as all 0's

FPN



FPN



Unnormalized

- Consider all decimal

Digits 0 51 78→0.78 * 10^{1} = +7.8

Digits 0 52 07→0.07 * 10^{2} = +7.0

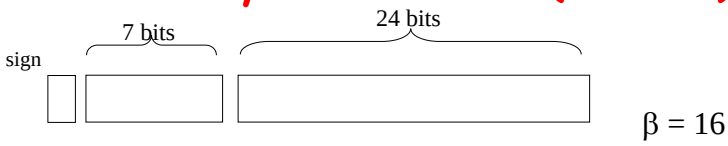
Digits 0 47 12→0.78 * 10^{-3} = 0.00012

Digits 1 51 78→-0.78 * 10^{1} = -7.8

Digits 0 52 00→0.00 * 10^{2} = zero

Digits 0 00 00→0.00 * 10^{0} = zero

IBM System 370 (short)



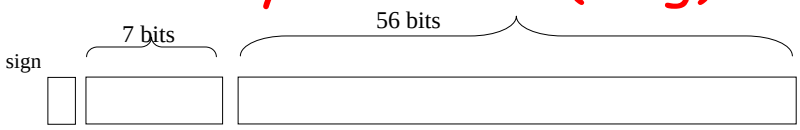
β = 16

24-bit mantissa

Characteristic (excess 64)

What is the max (min) representable number in this format?

IBM System 370 (long)



β = 16

56-bit mantissa

Characteristic (excess 64)

What is the max (min) representable number in this format?

FP Addition

- The two exponent must be equal.
- Shift the mantissa of the smallest operand to the right, and decrease exponent by one until the 2 exponents are equal.
- After alignment, the 2 mantissas are added (subtracted).
- The resulting number with its exponent is normalized (postnormalization).

FP Multiplication

- Mantissas are multiplied and exponent added,
- If 0.5 <= $|m_1 * m_2|$ < =1 do nothing
- IF 0.25 <= $|m_1 * m_2|$<0.5 shift to the left one place and decrease exponent by one
- If euther is zero, set normalized zero
- If overflow, set to max or min number

FP Division

- Divide mantissa and subtract exponent.
- Normalize if necessary
- If the dividend is zero, set to infinite
- If both are zero, NaN
- If overflow, set appropriately

Problems in FP Computations

$A = 0.100000 \times 16^{1}$

$B = 0.FFFFFF \times 16^{0}$

$A = 0.100000 \times 16^{1}$

$B = 0.0FFFFF \times 16^{1}$

$= 0.000001 \times 16^{1} = 0.1 \times 16^{-4}$

$A = 0.1000000000 \times 16^{1}$

$B = 0.FFFFFF0000 \times 16^{0}$

$A = 0.10000000 \times 16^{1}$

$B = 0.0FFFFFF \times 16^{1}$

$= 0.0000001 \times 16^{1}$

EROR = $0.1 \times 16^{-4} - 0.1 \times 16^{-5}$ =0.F x 16-5

Problems in FP Computations

$A = 0.100000 \times 16^{1}$

$B = 0.100000 \times 16^{-10}$

$A = 0.100000000000 \times 16^{1}$

$B = 0.000000000001 \times 16^{1}$

$= 0.100000 \times 16^{1}$

$A = 0.1000000000 \times 16^{1}$

$B = 0.FFFFFF0000 \times 16^{0}$

$A = 0.10000000 \times 16^{1}$

$B = 0.0FFFFFF \times 16^{1}$

$= 0.0000001 \times 16^{1}$

EROR = A+B = A and B is not zero ?

Problems in FP Computations

- An easy solution for this problem is guard digits
- Guard digits are extra digits appended to the right of the mantissa to hold intermediate results.

IEEE 754 FP

	Single	Double
Word length	32 bits	64 bits
Sign	1-bit	1-bit
Biased exp	8 bits	11 bits
Significand	(1)+23 bits	(1)+52 bits
Bias	127	1023
Precision	$2^{-32}=10^{-7}$	$2^{-52} = 10^{-16}$

$$X = (-1)^S * 2^{E-bias} * (1.f)$$

E=0,255 reserved

IEEE 754 FP

E	S	E	Significand	Interpretation
0	0	0	0	+zero
1	1	0	0	-Zero
2	0/1	0	Not 0	±Denormalized number
….				
127	0	0	0	+infinite
128	1	255	0	- infinite
129	X	255	Not 0	NaN

York University CSE

IEEE 754 FP

- The standard recommends but does not require the use of an extended format for temporary results.
- In this case, the leading hidden bit appears in the format

IEEE 754 FP

- Rounding
  - RN: Unbiased rounding to nearest (tie? Round to even)
  - RZ: Rounding towards zero
  - RM: Rounding towards minus infinity
  - RP: Rounding towards plus infinite

IEEE 754 FP

- Examples:

Exceptions in IEEE 754

- Invalid operation
- Overflow
- Division by zero
- Underflow
- Inexact result.

York University CSE