EE 486 : lecture 2 , the floating point numbers

M. J. Flynn

FPN: basic idea

- m ε M, machine numbers; r ε R, result numbers; R includes all M
- A FPN is +/- 0 . M x $\beta^{+/- e}$
- 1 2 3 4 5 6
- 6 attributes define a FPN

FPN: basic terms

- m ε M, machine numbers; r ε R, result numbers (R includes M and more)
- Expressed as six implicit or explicit parameters:
  - Sign of the number (explicit: usually S+M)
  - Radix point (implicit: fraction or significand)
  - Mantissa or significand(explicit;length implicit)
  - Radix (implicit)
  - Exponent and sign (usually excess coded)

FPN, more on the terms

- Exponent is usually represented in form of characteristic (or an excess code)
- Characteristic = exp + bias; where the bias is “usually” ½ β^n, n is the number of exp digits (bits)
- Why excess code?
  - Simplifies compares
  - Representation of zero is all zeros (same as integers)
  - Zero has minimum absolute exponent

FPN, more on terms

- Sign of the number: s_f is usually in S + M where a (–0) is not allowed
- Radix: only 2 or 16 are is use, $\beta = 2$ provides greater representational capacity, β=16 makes shifting easier and faster.
- Radix point: 0.1 or 1.0 with β =2; either immediately before or after the MSD of the fraction.
- Mantissa, fraction, significand synonyms

More FPN concepts

- Maximum representable number, max
  max =$\beta^{emax} M_{max} = \beta^{emax} (1 - \beta^{-m})$, m is mantissa bits
- Minimum representable (non 0) number,
  min = $\beta^{emin} M_{min}$
- Range is [min, max]
- Precision is β^{-m} same as ulp (unit in the last place)
- Gap is β^{exp} x (ulp).

Normalized nos., over/under flows and hidden “1”

- Normalized no. has non 0 msd; unnormalized in anything else; denormalized is unnormal with min exp
- Overflow: result > max… then either set *o* flag or set result to infinity representation.
- Underflow: min > result > 0.. Then either set *u* flag or set result to zero.
- Hidden 1: if $\beta = 2$ then msb of mantissa =1 so imply it (1).xxx or 0.(1)xxx

Rounding

- Rounding is the mapping of result number r εR, to an adjacent machine number m εM.
- Various types:
  - RZ(r) truncation of r to m
  - RN (r) round to nearest, up in a tie or round to nearest even
  - RP (r) round to positive infinity
  - RM (r) round to minus infinity

Four classic FPN systems

- Generic binary, classic 36b old style *fpns*, ca. 1952- 1990.
- Hex, $\beta = 16$, main frames
- Cray, quick and “dirty” binary, still useful for signal processing applications.
- IEEE standard binary, the most complicated and, by now, the most widely used

Generic binary

1	8	27
s_f	exp	mantissa

36

- Bias = $2^8/2 = 128$
- Max = 2^{127} (1- 2^{-27}), Min = 2^{-128} (2^{-1})
- Precision = 2^{-27}
- Round is RZ(r) only, no hidden 1
- Set *o* flag on overflow, set r to 0 on underflow.

Mainframes (S 360 ca 1963)

1	7	24
s_f	char	mantissa

32

Mantissa has 6 hex digits; note a normalized number may have leading digit 0001. The exponent (char) uses a binary radix.

So bias is $2^7/2 = 64$, emax is 127-64 = 63 and emin = -64

Then Max = 16^{63} (1 – 16^{-6}) and Min = 16^{-64} (16^{-1})

Set o flag on overflows and either set u flag or set r to 0 on underflows

Mainframe FPN, more

- In addition to single (32b), there’s double (64b) arranged 1+7+56(14digits) and quad (128b)
- Rounding is RZ(r) or RN (r)
- Implementations must use guard digit otherwise 1 x X is not = X (as 1 is 0001)
- Newer mainframes (S 390) offer both the old format and IEEE.

IEEE format, the basics

- Single: (32b) s+e+m bit layout is 1+7+23 Double: (64b) 1+ 11 +52
- Extended: (44b) 1+11+31 and (80b) 1+15+63 for register only operations; also quad
- β = 2 with S + M representation
- Hidden 1 to the left of radix point (1).xxx
- Exponent bias $(2^e/2)$ -1 not $2^e/2$

IEEE format more basics

- Four rounding modes RZ(r), RN(r)..round to nearest even, an unbiased round, RP(r), RM(r)
- Denormalize when exp is min and corresponding significand is less than min.
- Special indicators or allowance for +/- infinity, +/- 0, denormal, NAN..not a number
- Signed zero !

Guard bits and rounding

- Before post normal the result should be computed to

1	1.	L	G	R	S

Before post-normal, several leading “1”s can proceed the binary point G is first guard digit, round bit, R is the second guard bit and S is the sticky bit (the OR of all lower order bits).

Guard bits and rounding

- The rounding of the result adds A after post normalization

1.	L	G	R	S

First must combine R and S and then add A to the G bit

The action, A, is derived from L, G and S

Action table for RN

L	G	S	Action	A
X	0	0	Exact result, no round	0
X	0	1	Inexact result, but significand already rounded correctly	0
0	1	0	Tie case with even L, no rounding	0
1	1	0	Tie case with odd L, round to even	1
X	1	1	Round to nearest	1

IEEE reserved operands

S, sign	Biased exp	Significand	Meaning
0	0	0	+ zero
1	0	0	- zero
0/1	0	Not 0	+/- a denormalized number
0	255	0	+ infinity
1	255	0	- infinity
X	255	Not 0	NAN not a number