Assignment 6

Sonntag, 22. Juli 2018 20:32

Q format	Data type from <stdint.h>	Number of sign bits	Number of integer bits	Number of fractional bits	Total amount of bits	Minimum value	Maximum value
Q1.13	int16_t	3	Ø	13	16	-1	0.9998
U Q2.14	int16_t	Ø	2	14	16	Ø	3.9999
Signed Q0.15	int16_t	1	Ø	15	16	-1	0.9999
Q4.4 + Q4.4	int16_t	8	4	4	16	-16	15.9375
Q7 * Q15	int32_t	10	Ø	22	32	-1	0.9999
UQ1.2 + Q2.2	int8_t	4	2	2	8	-4	3.7500

- Notation
  - Q*n (implicit)* or Q*m.n (explicit)*
  - *n*... Number of fractional bits
  - *m*... Number of integer bits
    - inclusive (or exclusive) sign bit
    - U Qm.n ... unsigned designator
- Value range
  - Unsigned: $[0, (2^m - 2^{-n})]$
    - $m \geq 0$
  - Signed: $[-(2^{m-1}), (2^{m-1} - 2^{-n})]$
    - $m \geq 1$ (sign bit included)
- Adding
  - $Qm_1.n_1 + Qm_2.n_2 = Qm.n$
  - Fraction *n*: $\max(n_1, n_2)$
  - Integer *m*: $\max(m_1, m_2)+1$
- Multiplying
  - $Qm_1.n_1 * Qm_2.n_2 = Qm.n$
  - Fraction *n*: $n_1 + n_2$
  - Integer *m*: $m_1 + m_2$
    - Signed*Signed: $m_1 + m_2 - 1$ (extra sign bit)
- Convert from float
  - $q = f \cdot 2^n$
  - Problem
    - Truncation or rounding?
- Convert to float
  - $f = q \cdot 2^{(-n)}$

- Q4.4 + Q4.4 = Q5.4
- UQ1.2 + Q2.2 = Q3.2
- Q7 · Q15 = Q22

- Calculate the result by hand
  - 2.125 [Q3.3] + (-1.5) [Q1.2]
  - -0.125 [Q1.4] * 0.725 [Q1.4]

- 2.125 [Q3.3] + (-1.5) [Q1.2]

Q3.3 + Q1.2 = Q4.3 ⇒ S III FFF

2.125 ---> 0.125 · 2 = 0.250 | 0
0.250 · 2 = 0.500 | 0
0.500 · 2 = 1.000 | 1

10b

= 0010.001

-1.5 ---> 0.500 · 2 = 1 | 1
0 | 0
0 | 0

01b

= 0001.100

0010.001
1110.100
0000.101

5

$2^{-3} \cdot 5 = 0.625$

0001.100
1110.011
+1
1110.100

- -0.125 [Q1.4] · 0.725 [Q1.4]

Q1.4 · Q1.4 = Q1.8 ⇒ S FFFF FFFF

-0.125 ---> 0.125 · 2 = 0.250 | 0
0.250 · 2 = 0.500 | 0
0.500 · 2 = 1.000 | 1

0b

0.725 ---> 0.725 · 2 = 1.450 | 1
0.450 · 2 = 0.900 | 0
0.900 · 2 = 1.800 | 1

0b

-0.125 ⟶ 0.125 · 2 = 0.250 | 0
↓ 0.250 · 2 = 0.500 | 0
Ob 0.500 · 2 = 1.000 | 1

= 0.0010
1.1101
+1
1.1110

0.725 ⟶ 0.725 · 2 = 1.450 | 1
↓ 0.450 · 2 = 0.900 | 0
Ob 0.900 · 2 = 1.800 | 1
0.800 · 2 = 1.600 | 1

= 0.1011

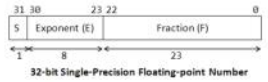
1.1110 · 0.1011
0 0000
1 1111 0
0 0000 00
1 1111 110
1 1111 1110
1 1 111
1 1 11111
1 1110 1010

- What is the best minimum Q-format and resulting error to store
  - 3.141592

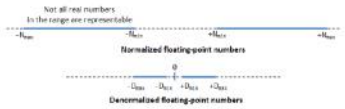
Leider nicht berechnet

- Proof by hand that for IEEE 754 binary32 floating point format
  - the largest positive number representable is roughly 3.4e+38
  - the smallest positive number representable is roughly 1.4e-45

- Single precision
  - 1 sign bit (S)
  - 8 exponent bits (E)
  - 23 fraction bits (F)
    - 22 effective fraction bits



- Normalized form
  - $(-1)^S \cdot 2^{(E-127)} \cdot 1.F$
- Denormalized form
  $(-1)^S \cdot 2^{(-126)} \cdot 0.F$



S EEEE EEEE FFFF FFFF FFFF FFFF FFFF FFF
1111 1110 1111 1111 1111 1111 1111 1111
254 8388607

max = $(-1)^S \cdot 2^{(E-127)} \cdot 1.F$ => normalized Form
= $(-1)^0 \cdot 2^{(254-127)} \cdot (1 + 8388607 \cdot 2^{-23})$
= $3{,}4028 \cdot 10^{38}$

S EEEE EEEE FFFF FFFF FFFF FFFF FFFF FFF
0000 0000 0000 0000 0000 0000 0000 0001
0 1

max = $(-1)^S \cdot 2^{(-126)} \cdot 0.F$ => denormalized Form
= $(-1)^0 \cdot 2^{(-126)} \cdot (1 \cdot 2^{-23})$
= $1{,}40129 \cdot 10^{-45}$

- `0.1f + 0.2f != 0.3f`

0.1f:

0.1 · 2 = 0.2	0
0.2 · 2 = 0.4	0
0.4 · 2 = 0.8	0
0.8 · 2 = 1.6	1

Shift um n=4 $\Rightarrow 2^{-4}$

0.6 · 2 = 1.2	1
0.2 · 2 = 0.4	0
0.4 · 2 = 0.8	0
0.8 · 2 = 1.6	1

$2^{-n} = 2^{E-127}$

$2^{-4} = 2^{E-127} \quad |\log_2$

$-4 = E - 127 \quad |+127$

$E = 123$

S	EEEE EEEE	FFFF FFFF FFFF FFFF FFFF FFF
0	0111 1011	1001 1001 1001 1001 1001 101

123 — 5033165

Value = $(-1)^S \cdot 2^{(E-127)} \cdot 1.F$

Value = $(-1)^0 \cdot 2^{(123-127)} \cdot (1 + 5033165 \cdot 2^{-23}) = 0.100000000149$

0.2f:

0.2 · 2 = 0.4	0
0.4 · 2 = 0.8	0
0.8 · 2 = 1.6	1
0.6 · 2 = 1.2	1

Shift um n=3 $\Rightarrow 2^{-3}$

$2^{n} = 2^{E-127}$

$2^{-3} = 2^{E-127} \quad |\log_2$

$-3 = E - 127 \quad |+127$

$E = 124$

S	EEEE EEEE	FFFF FFFF FFFF FFFF FFFF FFF
0	0111 1100	1001 1001 1001 1001 1001 101

124 — 5033165

Value = $(-1)^0 \cdot 2^{(124-127)} \cdot (1 + 5033165 \cdot 2^{-23}) = 0.200000000298$

0.3f:

0.3 · 2 = 0.6	0
0.6 · 2 = 1.2	1
0.2 · 2 = 0.4	0
0.4 · 2 = 0.8	0

Shift um n=2 $\Rightarrow 2^{-2}$

0.8 · 2 = 1.6	1
0.6 · 2 = 1.2	1
0.2 · 2 = 0.4	0
0.4 · 2 = 0.8	0

$2^{-2} = 2^{(E-127)}$

$E = 125$

S	EEEE EEEE	FFFF FFFF FFFF FFFF FFFF FFF
0	0111 1101	0011 0011 0011 0011 0011 010

125 — 1677722

Value = $(-1)^0 \cdot 2^{(125-127)} \cdot (1 + 1677722 \cdot 2^{-23}) = 0.300000001192$

$0.100000000149 + 0.200000000298 \; != \; 0.300000001192$ ✓

$0.1f + 0.2f \; != \; 0.3f$ ✓