Lesson 5, continued ...

## Unsigned fixed-point expansions

We saw (earlier) in Lesson 5 that if $x = \frac{p}{q}$ for $p, q \in \mathbb{Z}$ where $q \neq 0$ and $q = 2^j$ for some nonnegative $j \in \mathbb{Z}$, then $x$ yields a finite binary expansion with

$$x = \underbrace{b_i\, b_{i-1} \cdots b_2\, b_1\, b_0}_{\ell\text{-bit leading part}} \cdot \underbrace{b_{-1}\, b_{-2} \cdots b_{-f}}_{f\text{-bit fractional part}}$$

**Example 5.12** Let $x = \frac{101}{16} = \frac{64 + 32 + 4 + 1}{16}$ where $p = 101$ and $q = 16 = 2^4$

$$= 4 + 2 + \frac{1}{4} + \frac{1}{16}$$

$$= 2^2 + 2^1 + 2^{-2} + 2^{-4}$$

$$= (110.0101)_2$$

$$= \underbrace{b_2\, b_1\, b_0}_{\ell = 3 = i+1} \cdot \underbrace{b_{-1}\, b_{-2}\, b_{-3}\, b_{-4}}_{f = 4 \Rightarrow m = \ell + f = 7}$$

## Example 5.13

$$\text{Let} \quad x = \frac{731}{64} = \frac{512 + 128 + 64 + 16 + 8 + 2 + 1}{64}$$

$$= \frac{2^9 + 2^7 + 2^6 + 2^4 + 2^3 + 2^1 + 2^0}{2^6}$$

$$= 2^3 + 2^1 + 2^0 + 2^{-2} + 2^{-3} + 2^{-5} + 2^{-6}$$

$$= 1011.011011$$

$$= \underbrace{b_3\, b_2\, b_1\, b_0}_{} . \underbrace{b_{-1}\, b_{-2}\, b_{-3}\, b_{-4}\, b_{-5}\, b_{-6}}_{}$$

$\ell = 4 = i + 1$ and $f = 6$ $\Rightarrow m = \ell + f = 10$

For the sake of developing deep intuition about floating-point numbers, let's introduce a toy, fixed-point data type

$$x = \text{ufixed8}_{2(\ell,f)}(B) = \text{ufixed8}_{2(\ell,f)}(B_7 B_6 B_5 B_4 B_3 B_2 B_1 B_0)$$

This notation encodes many ideas:

- ufixed = unsigned, fixed-point number
- ufixed8 = number encoded using 8 individual digits
- $\text{ufixed8}_{2(\ell,f)}$ = this data encoding will be in binary (base 2) with $\ell$-bits dedicated to the leading part of our binary expansion and $f$-bits set aside for the fractional part where

  $m = 8 = \ell + f$ and $\ell, f \in \mathbb{Z}$ with $\ell, f \geq 0$

- Input $B = B_7 B_6 B_5 B_4 B_3 B_2 B_1 B_0$ is an raw, uninterpreted 8-bit binary word.
- output $x \in \mathbb{Q}$ is a type 1 rational number expressed as a finite binary expansion.

**Example 5.14**

Let $B = B_7 B_6 B_5 B_4 B_3 B_2 B_1 B_0 = 0110\ 0101$ (digit 8, digit 7, digit 6, digit 5, digit 4, digit 3, digit 2, digit 1)

Then if we set $\ell = 4$ and $f = 4$, we see

$$X = ufixed8_{2(4,4)}(B_7 B_6 B_5 B_4 B_3 B_2 B_1 B_0)$$

$$= ufixed8_{2(4,4)}(0110\ 0101)$$

$$= (0110.0101)_2$$

$$= \underbrace{b_3 b_2 b_1 b_0}_{\text{leading part}} . \underbrace{b_{-1} b_{-2} b_{-3} b_{-4}}_{\text{fractional part}}$$

leading part: $\ell = i+1 = 4$

fractional part: $f = 4$

where
$B_7 = 0 = b_3$
$B_6 = 1 = b_2$
$B_5 = 1 = b_1$
$B_4 = 0 = b_0$
$B_3 = 0 = b_{-1}$
$B_2 = 1 = b_{-2}$
$B_1 = 0 = b_{-3}$
$B_0 = 1 = b_{-4}$

$$= b_3 \cdot 2^3 + b_2 \cdot 2^2 + \cdots + b_{-4} \cdot 2^{-4}$$

$$= \sum_{K=-4}^{3} b_K \cdot 2^K$$

Example 5.14, cont...

Of course, we can convert $x = \text{ufixed8}_{2(4,4)}(0110\,0101)$ into decimal form by the appropriate operations

$$x = (0110.0101)_2$$

$$= 2^2 + 2^1 + 2^{-2} + 2^{-4}$$

$$= 4 + 2 + \frac{1}{4} + \frac{1}{16} = \frac{64+32+4+1}{16} = \left(\frac{101}{16}\right)_{10}$$

$$= 4 + 2 + 0.25 + 0.0625$$

$$= (6.3125)_{10}$$

Notice if we change the values of $\ell$ and $f$, the exact same raw, uninterpreted 8-bit binary word $B = 0110\,0101$ yields a different interpreted value:

$$y = \text{ufixed8}_{2(6,2)}(0110\,0101)$$

$$= (0110\,01.01)_2$$

$$= 2^4 + 2^3 + 2^0 + 2^{-2} = 16 + 8 + \frac{1}{4} = (24.25)_{10}$$

This unsigned, fixed-point data structure suggests some very interesting features for storing and encoding finite binary expansions.

Let's take a look at an analogous data structure:

$$x = \text{ufixed4}_{2(\ell,f)}(B) = \text{ufixed4}_{2(\ell,f)}(B_3\,B_2\,B_1\,B_0)$$

$x$: interpreted binary value

$B_3\,B_2\,B_1\,B_0$: raw, uninterpreted 4-bit binary word

In this case, let's study all possible combinations of values for $\ell = i+1$ and $f$ such that $\ell + f = 4$. These include

Case 1: $\ell = 4$, $f = 0$

Case 2: $\ell = 3$, $f = 1$

Case 3: $\ell = 2$, $f = 2$

Case 4: $\ell = 1$, $f = 3$

Case 5: $\ell = 0$, $f = 4$

In the next five pages, let's consider all possible ways to interpret a raw uninterpreted 4-bit binary word as a finite fixed-point binary expansion using our ufixed4 data class.

## Example 5.15

We begin by interpreting our $m = 4$-bit raw, uninterpreted binary word $B = B_3 B_2 B_1 B_0$ as an unsigned, fixed-point number

$$x = \text{ufixed4}_{2(\ell,f)}(B_3 B_2 B_1 B_0) \quad \text{w/ leading part } \ell = 4, \text{ fractional part } f = 0$$

$$= \text{ufixed4}_{2(4,0)}(B_3 B_2 B_1 B_0)$$

$$= \underbrace{b_3 b_2 b_1 b_0}_{\text{leading part}} . \overbrace{0}^{\text{implicit fractional part since } f=0}$$

with $B_3 = b_3$, $B_2 = b_2$, $B_1 = b_1$, $B_0 = b_0$

$$= b_3 \cdot 2^3 + b_2 \cdot 2^2 + b_1 \cdot 2^1 + b_0 \cdot 2^0$$

$$= \sum_{j=0}^{3} B_j \cdot 2^{j-0} = \sum_{K=0}^{3} b_K \cdot 2^K \quad \text{where } b_K = B_K \text{ \& } j = K.$$

## Example 5.15, continued...

Let's take a look at the entire table of values associated with $ufixed4_{2(4,0)}(B)$ for all possible 4-bit binary words $B = B_3 B_2 B_1 B_0$:

implicit binary point since $f = 0$

| raw, uninterpreted binary word | interpreted binary value | decimal value | fractional $p/q$ |
|---|---|---|---|
| 0000 | $(0000.0)_2$ | 0.0 | 0/1 |
| 0001 | $(0001.0)_2$ | 1.0 | 1/1 |
| 0010 | $(0010.0)_2$ | 2.0 | 2/1 |
| 0011 | $(0011.0)_2$ | 3.0 | 3/1 |
| 0100 | 0100.0 | 4.0 | 4/1 |
| 0101 | 0101.0 | 5.0 | 5/1 |
| 0110 | 0110.0 | 6.0 | 6/1 |
| 0111 | 0111.0 | 7.0 | 7/1 |
| 1000 | 1000.0 | 8.0 | 8/1 |
| 1001 | 1001.0 | 9.0 | 9/1 |
| 1010 | 1010.0 | 10.0 | 10/1 |
| 1011 | 1011.0 | 11.0 | 11/1 |
| 1100 | 1100.0 | 12.0 | 12/1 |
| 1101 | 1101.0 | 13.0 | 13/1 |
| 1110 | 1110.0 | 14.0 | 14/1 |
| 1111 | 1111.0 | 15.0 | 15/1 |

Range for $x$: $0 \leq x \leq 2^4 - 1$

$$\Rightarrow 0 \leq x \leq 2^4 - \frac{1}{2^0}$$

$$\Rightarrow 0 \leq x \leq 2^\ell - \frac{1}{2^f}$$

Example 5.15, continued...

$x = ufixed4_{2(3,1)}(B) = ufixed4_{2(3,1)}(B_3 B_2 B_1 B_0)$ $\quad l = 3,\ f = 1$

$\Rightarrow i+1 = 3,\ f = 1$

$\Rightarrow i = 2,\ f = 1$

$= \underbrace{b_2 b_1 b_0}_{\text{leading part}} . \underbrace{b_{-1}}_{\text{fractional part}}$

$= b_2 \cdot 2^2 + b_1 \cdot 2^1 + b_0 \cdot 2^0 + b_{-1} \cdot 2^{-1}$

$= b_2 \cdot 2^2 + b_1 \cdot 2^1 + b_0 \cdot 2^0 + \frac{b_{-1}}{2}$

$= \sum_{j=0}^{3} B_j \cdot 2^{j-1}$

$= \sum_{K=-1}^{2} b_K \cdot 2^K$ with $b_K = B_{K+1}$ and $K = j-1$

| raw, uninterpreted binary word | interpreted binary value | decimal value | fractional p/q |
|---|---|---|---|
| 0000 | 000.0 | 0.0 | 0/2 |
| 0001 | 000.1 | 0.5 | 1/2 |
| 0010 | 001.0 | 1.0 | 2/2 |
| 0011 | 001.1 | 1.5 | 3/2 |
| 0100 | 010.0 | 2.0 | 4/2 |
| 0101 | 010.1 | 2.5 | 5/2 |
| 0110 | 011.0 | 3.0 | 6/2 |
| 0111 | 011.1 | 3.5 | 7/2 |
| 1000 | 100.0 | 4.0 | 8/2 |
| 1001 | 100.1 | 4.5 | 9/2 |
| 1010 | 101.0 | 5.0 | 10/2 |
| 1011 | 101.1 | 5.5 | 11/2 |
| 1100 | 110.0 | 6.0 | 12/2 |
| 1101 | 110.1 | 6.5 | 13/2 |
| 1110 | 111.0 | 7.0 | 14/2 |
| 1111 | 111.1 | 7.5 | 15/2 |

range of $x$ : $\quad 0 \le x \le 2^3 - \frac{1}{2}$

$\Rightarrow \quad 0 \le x \le 2^l - \frac{1}{2^f}$

**Example 5.15, continued..**

$X = \text{ufixed4}_{2(2,2)}(B) = \text{ufixed4}_{2(2,2)}(B_3 B_2 B_1 B_0)$ $\quad \ell = i+1 = 2,\ f = 2,\ m = 4$

$\Rightarrow \quad i = 1,\ f = 2$

$$= b_1 b_0 . b_{-1} b_{-2}$$

$$= b_1 \cdot 2^1 + b_0 \cdot 2^0 + b_{-1} \cdot 2^{-1} + b_{-2} \cdot 2^{-2}$$

$$= b_1 \cdot 2^1 + b_0 \cdot 2^0 + \frac{b_{-1}}{2} + \frac{b_{-2}}{4}$$

$$= \sum_{j=0}^{3} B_j \cdot 2^{j-2}$$

$$= \sum_{K=-2}^{1} b_K \cdot 2^K \quad \text{with } b_K = B_{K+2} \text{ and } K = j-2$$

| raw, uninterpreted binary word | interpreted binary value | decimal value | fractional p/q |
|---|---|---|---|
| 0000 | 00.00 | 0.00 | 0/4 |
| 0001 | 00.01 | 0.25 | 1/4 |
| 0010 | 00.10 | 0.50 | 2/4 |
| 0011 | 00.11 | 0.75 | 3/4 |
| 0100 | 01.00 | 1.00 | 4/4 |
| 0101 | 01.01 | 1.25 | 5/4 |
| 0110 | 01.10 | 1.50 | 6/4 |
| 0111 | 01.11 | 1.75 | 7/4 |
| 1000 | 10.00 | 2.00 | 8/4 |
| 1001 | 10.01 | 2.25 | 9/4 |
| 1010 | 10.10 | 2.50 | 10/4 |
| 1011 | 10.11 | 2.75 | 11/4 |
| 1100 | 11.00 | 3.00 | 12/4 |
| 1101 | 11.01 | 3.25 | 13/4 |
| 1110 | 11.10 | 3.50 | 14/4 |
| 1111 | 11.11 | 3.75 | 15/4 |

range: $0 \le x \le 2^2 - \frac{1}{4} = 2^2 - \frac{1}{2^2} = 2^\ell - \frac{1}{2^f} = 2^{i+1} - \frac{1}{2^f}$

Example 5.15, continued ...

$X = ufixed4_{2(1,3)}(B) = ufixed4_{2(1,3)}(B_3 B_2 B_1 B_0)$ $\quad l = i+1 = 1,\ f = 3,\ m = 4$

$\Rightarrow i = 0,\ f = 3,\ m = i+1+f$

$$= b_0 . b_{-1} b_{-2} b_{-3}$$

$$= b_0 \cdot 2^0 + \frac{b_{-1}}{2^1} + \frac{b_{-2}}{2^{-2}} + \frac{b_{-3}}{2^{-3}}$$

$$= \sum_{j=0}^{3} B_j \cdot 2^{j-3}$$

$$= \sum_{K=-3}^{0} b_K \cdot 2^K \quad \text{with} \quad b_K = B_{K+3} \quad \text{and} \quad K = j-3$$

| raw, uninterpreted binary word | interpreted binary value | decimal value | fractional p/q |
|---|---|---|---|
| 0000 | 0.000 | 0.000 | 0/8 |
| 0001 | 0.001 | 0.125 | 1/8 |
| 0010 | 0.010 | 0.250 | 2/8 |
| 0011 | 0.011 | 0.375 | 3/8 |
| 0100 | 0.100 | 0.500 | 4/8 |
| 0101 | 0.101 | 0.625 | 5/8 |
| 0110 | 0.110 | 0.750 | 6/8 |
| 0111 | 0.111 | 0.875 | 7/8 |
| 1000 | 1.000 | 1.000 | 8/8 |
| 1001 | 1.001 | 1.125 | 9/8 |
| 1010 | 1.010 | 1.250 | 10/8 |
| 1011 | 1.011 | 1.375 | 11/8 |
| 1100 | 1.100 | 1.500 | 12/8 |
| 1101 | 1.101 | 1.625 | 13/8 |
| 1110 | 1.110 | 1.750 | 14/8 |
| 1111 | 1.111 | 1.875 | 15/8 |

Range: $0 \le x \le 2^1 - \frac{1}{8} = 2^1 - \frac{1}{2^3} = 2^l - \frac{1}{2^f} = 2^{i+1} - \frac{1}{2^f}$

**Example 5.15, continued ...** Let's try case 5 w/ $\ell = 0$ and $f = 4$.

$$x = \text{ufixed4}_{2(0,4)}(B_3 B_2 B_1 B_0)$$

$\ell = 0$ , $f = 4$ , $m = \ell + f = 4$

(no value for i)

$$= 0.\ b_{-1}\ b_{-2}\ b_{-3}\ b_{-4}$$

implicit leading part since $\ell = 0$

fractional part occupies all 4-bits

$$= \frac{b_{-1}}{2^1} + \frac{b_{-2}}{2^2} + \frac{b_{-3}}{2^3} + \frac{b_{-4}}{2^4}$$

$$= \sum_{j=0}^{3} B_j \cdot 2^{j-4}$$

$$= \sum_{K=-4}^{-1} b_K \cdot 2^K \quad \text{with} \quad b_K = B_{K+4} \quad \text{and} \quad K = j-4$$

| raw, uninterpreted binary word | interpreted binary value | decimal value | fractional p/q |
|---|---|---|---|
| 0000 | 0.0000 | 0.0000 | 0/16 |
| 0001 | 0.0001 | 0.0625 | 1/16 |
| 0010 | 0.0010 | 0.1250 | 2/16 |
| 0011 | 0.0011 | 0.1875 | 3/16 |
| 0100 | 0.0100 | 0.2500 | 4/16 |
| 0101 | 0.0101 | 0.3125 | 5/16 |
| 0110 | 0.0110 | 0.3750 | 6/16 |
| 0111 | 0.0111 | 0.4375 | 7/16 |
| 1000 | 0.1000 | 0.5000 | 8/16 |
| 1001 | 0.1001 | 0.5625 | 9/16 |
| 1010 | 0.1010 | 0.6250 | 10/16 |
| 1011 | 0.1011 | 0.6875 | 11/16 |
| 1100 | 0.1100 | 0.7500 | 12/16 |
| 1101 | 0.1101 | 0.8125 | 13/16 |
| 1110 | 0.1110 | 0.8750 | 14/16 |
| 1111 | 0.1111 | 0.9375 | 15/16 |

range: $0 \le x \le 1 - \frac{1}{16} \Rightarrow 0 \le x \le 2^0 - \frac{1}{2^4} \Rightarrow 0 \le x \le 2^\ell - \frac{1}{2^f}$

Now that we've done a complete study of all five cases for nonnegative integer values of $\ell$ and $f$ such that

$$\ell + f = m = 4$$

Let's use our hard-fought intuition to make some general observations. In particular, if

$$x = ufixed M_{2(\ell,f)} (B_{m-1} B_{m-2} \cdots B_2 B_1 B_0)$$

then we can write the interpreted value of the finite binary expansion for $x$ as

$$x = \begin{cases} b_i b_{i-1} \cdots b_2 b_1 b_0 . b_{-1} b_{-2} \cdots b_{-f} & \text{if } \ell = i+1 > 0 \\ 0 . b_{-1} b_{-2} \cdots b_{-(f-1)} b_{-f} & \text{if } \ell = 0 \end{cases}$$

where the total number of bit $m = \ell + f$ and

$\ell$ = # bits in the leading part of $x$

$f$ = # bits in the fractional part of $x$

Given this $\mathrm{ufixedm}_{2(\ell,f)}(B)$ data structure, we can analyze this mathematically as follows:

$$X = \mathrm{ufixedm}_{2(\ell,f)}(B_{m-1}\, B_{m-2} \cdots B_2\, B_1\, B_0)$$

$$= b_i\, b_{i-1} \cdots b_2\, b_1\, b_0 \,.\, b_{-1}\, b_{-2} \cdots b_{-f} \qquad \text{assuming } \ell > 0 \Rightarrow i \geq 0$$

$$= b_i \cdot 2^i + b_{i-1} \cdot 2^{i-1} + \cdots + b_1 \cdot 2^1 + b_0 \cdot 2^0 + b_{-1} \cdot 2^{-1} + \cdots + b_{-f} \cdot 2^{-f}$$

$$= \sum_{k=-f}^{i} b_k \cdot 2^k$$

$$= \sum_{j=0}^{m-1} B_j\, 2^{j-f} \qquad \text{where } b_k = B_{k+f} \text{ and } k = j-f$$

With this in mind, we get an immediate approximation on the range of type I rational $x \in \mathbb{Q}$ that can be encoded in our $\mathrm{ufixedm}_{2(\ell,f)}(B)$ data class using a raw, uninterpreted m-bit binary string.

In particular, we see that

- range of $x$: $0 \leq x \leq 2^{\ell} - \frac{1}{2^{f}}$

Moreover, we know that our word length for $x$ is

$$m = \ell + f \quad \text{bits}$$

# The $ufixed8_{2(\ell,f)}(B)$ Data Type

interpreted binary or decimal value

$$X = ufixed8_{2(\ell,f)}(B)$$

$$= ufixed8_{2(\ell,f)}(\underbrace{B_7\ B_6\ B_5\ B_4\ B_3\ B_2\ B_1\ B_0}_{\text{Raw, uninterpreted 8-bit binary word}})$$

$$= \underbrace{b_i\ b_{i-1}\ \cdots\ b_1\ b_0}_{\text{leading part of } x \text{ has } (i+1)=\ell \text{ bits}} . \underbrace{b_{-1}\ \cdots\ b_{-f}}_{\text{fractional part of } x \text{ has } f\text{-bits}} \quad \text{where } m = \ell + f \text{ and } \ell = i+1 > 0$$

In this situation, we have nine separate cases to consider for combination pairs of values for $\ell$ and $f$ w/ $m = 8 = \ell + f$

case 1: $\ell = 8$, $f = 0$

Case 2: $\ell = 7$, $f = 1$

case 3: $\ell = 6$, $f = 2$

Case 4: $\ell = 5$, $f = 3$

Case 5: $\ell = 4$, $f = 4$

case 6: $\ell = 3$, $f = 5$

case 7: $\ell = 2$, $f = 6$

case 8: $\ell = 1$, $f = 7$

case 9: $\ell = 0$, $f = 8$

In each of these nine cases, we have $2^8 = 256$ different possible values for our raw uninterpreted 8-bit word

$$B = B_7\ B_6\ B_5\ B_4\ B_3\ B_2\ B_1\ B_0$$

To get deep intuition on how this data type encodes type 1 $x \in \mathbb{Q}$ that yield finite binary expansions, lets run through all possible combinations and explore this data encoding scheme. Once we've done so, our goal will be to generalize our interpretation of this work to the ufixed $m_{2(\ell,f)}(B)$ class.

Remember that the ultimate goal of all this work with fixed-point data is to develop a deeply intuitive an intimate relationship with the fundamental ideas that form the foundations for IEEE 754 floating-point format.