MATHSXtra

Unit 45

Fixed and floating point notation

Objectives

On completion of this unit you should understand:

1. Fractional binary notation.
2. Fixed point binary numbers
3. Floating point binary numbers.

Cook, M. and Rimmer, H. (1993), Stafford College,
Earl Street, Stafford, ST16 2QR

Fractional binary numbers

When we considered the change from denary to binary numbers in an earlier unit, we did this by writing the headings above each column.

2^8	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
256	128	64	32	16	8	4	2	1

Notice that each number is double the previous one as you move to the left.
1, 2, 4, 8, 16.....and so on.
To convert the denary number 35 into a binary number, we need 32 + 2 + 1 and so we obtain, $1ØØØ11_2$.

This same system applies to numbers less than one. We can continue our headings to the right of the decimal point.

2^0	.	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}.....

2^{-1} is the same as $1/2^1$ or $^1/_2$ or 0.5.
2^{-2} is the same as $1/2^2$ or $^1/_4$ or 0.25.
2^{-3} is the same as $1/2^3$ or $^1/_8$ or 0.125.

We can therefore consider our headings as follows, starting with one and moving to the right of the decimal point.

	2^0	.	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}.....
	1	.	$^1/_2$	$^1/_4$	$^1/_8$	$^1/_{16}$	$^1/_{32}$	$^1/_{64}$.....
or	1	.	0.5	0.25	0.125	0.0625	0.03125	0.015625....

Notice that each number is half the previous one as you move to the right.

Study these examples.

Example 1
Convert the binary number Ø.1Ø1Ø1Ø to a denary number.

We set up our headings as follows with the number in position. We shall use the fraction headings this time.

2^0	.	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}.....
1	.	$^1/_2$	$^1/_4$	$^1/_8$	$^1/_{16}$	$^1/_{32}$	$^1/_{64}$.....
Ø	.	1	Ø	1	Ø	1	Ø

The number Ø.1Ø1Ø1Ø is therefore, $^1/_2 + ^1/_8 + ^1/_{32} = ^{21}/_{32}$ or 0.65625.

Example 2
Convert the binary number 11.Ø11Ø to a denary number.

We set up our headings as follows with the number in position. We shall use the decimal headings this time.

2^1	2^0	.	2^{-1}	2^{-2}	2^{-3}	2^{-4}
2	1	.	0.5	0.25	0.125	0.0625
1	1	.	Ø	1	1	Ø

The binary number 11.ØØ1Ø1 is therefore,
2 + 1 + 0.25 + 0.125 = 3.375 as a mixed decimal or $3^3/_8$ as a mixed fraction in the denary system.

Try this exercise.

Exercise A
Convert the following binary numbers to denary numbers giving your answers as,
a) mixed fractions,
b) mixed decimals.

1. 11.1Ø1
2. 1.1Ø111
3. 1Ø.Ø1Ø11
4. 1Ø1.1Ø11Ø1Ø
5. 11Ø1.11
6. 1ØØ.Ø1Ø1Ø1Ø
7. 1ØØ.1ØØ1
8. 11.11
9. Ø.11Ø1
10. 1Ø1.Ø11

Check your answers with those at the end of the unit.

Converting decimals to binary numbers

Consider these examples.

Example 3
Convert 3.75_{10} to a binary number.

We set up our headings as follows with the number in position. We use the decimal headings.

2^1	2^0	.	2^{-1}	2^{-2}	2^{-3}	2^{-4}
2	**1**	**.**	**0.5**	**0.25**	**0.125**	**0.0625**

3_{10} is written as 11_2. We now need to convert the decimal part of the number.
0.75 is made up of 0.5 + 0.25. so $3.75_{10} = 11.11_2$.

Example 4
Convert 8.4375_{10} to a binary number.

We set up our headings as follows with the number in position. We use the decimal headings.

2^3	2^2	2^1	2^0	.	2^{-1}	2^{-2}	2^{-3}	2^{-4}
8	**4**	**2**	**1**	**.**	**0.5**	**0.25**	**0.125**	**0.0625**

8_{10} is written as $1ØØØ_2$. We now need to convert the decimal part of the number.
0.4375 is made up of 0.25 + 0.125. + 0.0625 so $8_{10} = 1ØØØ.Ø111_2$.

It is not always possible to find an exact binary number to replace all decimals. For example if we had to find a binary number for 0.4374, we would need to use 0.25 + 0.125 and then 0.0624. There is no binary digit for 0.0624 and so the nearest would be to use the binary digit for 0.0625.

Try the following exercise.

Exercise B
Convert the following numbers to binary form.

1. 2.0625
2. 5.3125
3. 0.5625
4. 3.25
5. 0.75
6. 0.8125

Check your answers with those at the end of the unit.

Fixed point binary numbers

Just as calculators have a fixed display of eight or ten numbers, computer registers for holding numbers are a fixed length, so compromises need to be made.

There are two major types of fixed representation. These are integer or fractional. We are going to consider an 8-bit register.

In this system, the first digit tells us whether the number is positive or negative.

A number starting with a Ø is positive.
A number starting with a 1 is negative.

The maximum positive number which can be stored using 8 bits must be Ø1111111. Remember that the Ø tells us that the number is positive. This number is +127.

The smallest positive number in an 8-bit register is therefore ØØØØØØØ1. This is +1.

Zero is considered to be neither positive nor negative.

The negative numbers in an 8-bit register range from -1 to -128. In binary form these are written as 11111111 and 1ØØØØØØØ.

The following examples show how negative numbers are written in an 8-bit binary system, using the quick method shown in the Binary, Octal and Hexadecimal unit.

Study these examples.

Example 5

Using eight bits, and the quick method, find the binary number to represent -3_{10}.

Using eight bits, the number 3_{10} is converted to the binary number $ØØØØØØ11_2$.

Starting at the right hand side, rewrite the number up to and including the first 1.

$ØØØØØØ11_2$ write **1**

For all the remaining digits, the Ø's become 1 and the 1's become Ø.

111111Ø1.

In this 8-bit system, -3_{10} is written $111111Ø1_2$.

Example 6
Using the quick method find the binary number to represent -20_{10} in an 8-bit register.

The number 20_{10}, is written as $ØØØ1Ø1ØØ_2$.
Starting at the right hand side, rewrite the number up to and including the first 1.
$ØØØ1Ø1ØØ_2$ write **1ØØ**
For all the remaining digits, the Ø's become 1 and the 1's become Ø.
111Ø11ØØ.

-20_{10} is equal to the binary number $111Ø11ØØ_2$.

Try this exercise.

Exercise C
Using an 8-bit register, find the binary equivalent of each of the following denary numbers.
1. 27
2. 111
3. 8
4. 15
5. 39
6. -10
7. -18
8. -35
9. -101
10. -69

Check your answers with those at the end of the unit.

Floating point representation of binary numbers

Numbers can be split into two parts. The first part is called a **mantissa** and the second part is called an **exponent**.
Consider the standard form number 1.789×10^6.
1.789 is the mantissa and 10^6 is the exponent.
A similar method can be employed for binary numbers. We shall consider a 16-bit register. This can be split into two parts. The first part carries the mantissa and the second part carries the exponent.

Consider these examples.

Example 7
A 16-bit register is split into two parts. Ten bits are to be used for the mantissa and six bits are to be used for the exponent. Convert the binary number, Ø.ØØ111ØØ11 | ØØØ11Ø into denary form.

Ø.ØØ111ØØ11 is the mantissa and is using ten bits. This is a positive number because the first digit is Ø.
ØØØ11Ø is the exponent and is using six bits. The exponent is also positive because it begins with a Ø.

First consider the exponent. ØØØ11Ø is equal to 6 in denary form.
This means that the point in the mantissa is to be moved six places to the right. ØØØ1110.Ø11
We can now convert this binary number to denary form.
ØØØ1110.Ø11 = 8 + 4 + 2 + $^1/_4$ + $^1/_8$ = $14^3/_8$ or 14.375.

Example 8
A 16-bit register is split into two parts. Twelve bits are to be used for the mantissa and four bits are to be used for the exponent. Convert the binary number, Ø.Ø1Ø111ØØ111 | Ø111 into denary form.

Ø.Ø1Ø111ØØ111 is the mantissa and is using twelve bits. This is a positive number because the first digit is Ø.

Ø111 is the exponent and is using four bits.

First consider the exponent. Ø111 is equal to 7 in denary form.
This means that the point in the mantissa is to be moved seven places to the right. ØØ1Ø1110.Ø111
We can now convert this binary number to denary form.
ØØ1Ø1110.Ø111
= 32 + 8 + 4 + 2 + 0.25 + 0.125 + 0.0625 = 46.4375 or $46^7/_{16}$.

Try this exercise.

Exercise D
Convert the following positive 16-bit binary numbers to denary form as mixed fractions or decimals.

1. Ø.Ø1Ø111ØØ1 | ØØØØ11
2. Ø.ØØ1Ø11111Ø | ØØ1ØØØ
3. Ø.Ø1Ø1111Ø111 | Ø11Ø
4. Ø.ØØØ11ØØØ111 | Ø11Ø
5. Ø.ØØØ1110Ø11 | ØØ111
6. Ø.1110Ø11 | ØØØØØ11Ø
7. Ø.ØØ111ØØ1Ø1 | ØØ1ØØ
8. Ø.Ø1Ø110ØØØØØ | Ø111
9. Ø.Ø1Ø1Ø1ØØ111 | Ø1Ø1
10. Ø.ØØØ111ØØØØØ | Ø1ØØ

Check your answers with those at the end of the unit.

Floating point representation of negative binary numbers

We shall now consider binary numbers in a 16-bit register, where the first digit is 1. This means that the number is negative.

Consider these examples.

Example 9
Convert 1.Ø1Ø11111Ø | ØØØØ111 to denary form.

Consider the exponent, ØØØØ111 is equal to 7.
Now consider the mantissa, 1.Ø1Ø11111Ø.
Move the decimal point seven places to the right, 1Ø1Ø1111.Ø
The first digit is 1 so the number is negative. Using the two's complement quick method from the Binary Octal and Hexadecimal System unit, we can convert this number. Starting at the right hand side, rewrite the number up to and including the first 1.
1Ø1Ø1111.Ø, write, **1.Ø.**
For all the remaining digits the Ø's become 1 and the 1's become Ø as follows, **Ø1Ø1ØØØ1**.Ø. This is equal to 81

The binary number 1.Ø1Ø11111Ø | ØØØØ111 is equal to -81 in denary form.

Example 10

Convert 1.ØØØ1Ø111ØØ| ØØØ11 to denary form.

Consider the exponent.
ØØØ11 is equal to 3.
Now consider the mantissa, 1.ØØØ1Ø111ØØ.
Move the decimal point three places to the right, 1ØØØ.1Ø111ØØ.
The first digit is 1 so the number is negative. Using the two's complement quick method from the Binary Octal and Hexadecimal System unit, we can convert this number. Starting at the right hand side, rewrite the number up to and including the first 1.
1ØØØ.1Ø111**ØØ**, write, **1ØØ**.
For all the remaining digits the Ø's become 1 and the 1's become Ø as follows, **Ø111.Ø1ØØ**1ØØ.
This is equal to 7 + 0.25 + 0.03125 = 7.28125 or $7^{9}/_{32}$.
The binary number,
1.ØØØ1Ø111ØØ| ØØØ11 is equal to the denary number -7.28125.

Try this exercise.

Exercise E

Convert the following negative 16-bit binary numbers to denary form. Give your answers as mixed fractions or decimals.

1. 1.Ø1Ø111ØØ1 | ØØØØ11
2. 1.11Ø1Ø1111 | ØØØ111
3. 1.Ø1Ø1111Ø111 | Ø11Ø
4. 1.ØØØ11ØØØ111 | ØØ1Ø
5. 1.ØØ111ØØ11 | ØØØ111
6. 1.1110111 | ØØØØØ11Ø
7. 1.ØØ111ØØ1Ø1 | ØØ1ØØ
8. 1.Ø1Ø11ØØØØØØ | Ø111
9. 1.Ø1Ø1Ø1ØØØØØ | Ø1Ø1
10. 1.ØØØ111ØØØØØ | Ø1ØØ

Check your answers with those at the end of the unit.

Floating point representation when the exponent is negative

We know that the exponent is negative if it begins with the number 1.

Ø.ØØ1Ø111ØØ| 1ØØØ11 is a positive number but the exponent is negative. The mantissa is positive because it begins with a Ø. This means that the number itself is positive.
The exponent is negative, so we would need to move the decimal point on the matissa to the **left**

Consider this example.

Example 11
Convert Ø.ØØ1Ø111ØØ| 111Ø11 to denary form.

Consider the exponent.
111Ø11 is negative. It begins with a 1.
Using the quick method we can write the two's complement, ØØØ1Ø1. This is the denary number 5, so 111Ø11 is equal to -5 in the denary system.

Now consider the mantissa, Ø.ØØ1Ø111ØØ.
Move the decimal point five places to the **left**, Ø.ØØØØØØØ1Ø111ØØ.

This is equal to $^{1}/_{256} + {}^{1}/_{1024} + {}^{1}/_{2048} + {}^{1}/_{4096} + = {}^{23}/_{4096}$.
or
0.00390625 + 0.000976562 + 0.000488281 + 0.00024414 = 0.00561523.
The binary number,
Ø.ØØ1Ø111ØØ| 111Ø11 is equal to the denary number $^{23}/_{4096}$ or 0.00561523.

Try this exercise.

Exercise F
Convert the following binary numbers to denary form. Give your answers as fractions or decimals.

1. Ø.Ø1Ø1Ø11ØØ| 1111Ø1
2. Ø.Ø1Ø1Ø1ØØØØ| 11Ø1Ø
3. Ø.Ø1Ø1Ø11ØØØØ| 11Ø1
4. Ø.Ø111Ø11ØØ| 11111Ø
5. Ø.Ø111ØØØ11ØØ| 1111

Check your answers with those at the end of the unit.

Floating point representation when the exponent and the mantissa are negative

1.ØØØ1Ø111ØØ| 1ØØØ1 has a negative mantissa and a negative exponent. The number is negative. Because the exponent is negative we need to move the decimal point of the mantissa to the **left**.

Example 12
Convert 1.ØØ1Ø11111| 1111Ø1 to denary form.

Consider the exponent.
1111Ø1 is negative. It begins with a 1.
Using the quick method we can write the two's complement, ØØØØ11. This is the denary number 3, so 1111Ø1 is equal to -3 in the denary system.

Now consider the mantissa, 1.ØØ1Ø11111. We can write the two's complement of this number using the quick method. Ø.11Ø1ØØØØ1
Move the decimal point three places to the **left**, Ø.ØØØ11Ø1ØØØØ1.

This is equal to $^{1}/_{16} + {}^{1}/_{32} + {}^{1}/_{128} + {}^{1}/_{4096} = {}^{417}/_{4096}$.
The binary number,
1.ØØ1Ø11111| 1111Ø1 is equal to the denary number $-{}^{417}/_{4096}$ or -0.1018066.

Try this exercise.

Exercise G
Convert the following binary numbers to denary form. Give your answers as fractions or decimals.

1. 1.11Ø1ØØØØØ| 1111Ø1
2. 1.Ø1Ø1Ø1ØØØØ| 11Ø11
3. 1.Ø1Ø1Ø11ØØØØ| 11Ø1
4. 1.Ø1111Ø11ØØ| 1111Ø
5. 1.Ø111Ø1111ØØ| 1111

Check your answers with those at the end of the unit.

Answers

Exercise A

1.	a)	$3^{5}/_{8}$	b)	3.625
2.	a)	$1^{23}/_{32}$	b)	1.71875
3.	a)	$2^{11}/_{32}$	b)	2.34375
4.	a)	$5^{45}/_{64}$	b)	5.703125
5.	a)	$13^{3}/_{4}$	b)	13.75
6.	a)	$4^{21}/_{64}$	b)	4.328125
7.	a)	$4^{9}/_{16}$	b)	4.5625
8.	a)	$3^{3}/_{4}$	b)	3.75
9.	a)	$^{13}/_{16}$	b)	0.8125
10.	a)	$5^{3}/_{8}$	b)	5.375

Exercise B

1. 1Ø.ØØØ1
2. 1Ø1.Ø1Ø1
3. Ø.1ØØ1
4. 11.Ø1
5. Ø.11
6. Ø.11Ø1

Exercise C

1. ØØØ11Ø11
2. Ø11Ø1111
3. ØØØØ1ØØØ
4. ØØØØ1111
5. ØØ1ØØ111
6. 1111Ø11Ø
7. 111Ø1111Ø
8. 11Ø1111Ø1
9. 1ØØ11Ø11
10. 1Ø111Ø11

Exercise D

1. $2^{57}/_{64}$ or 2.890625
2. 47
3. $23^{23}/_{32}$ or 23.71875
4. $6^{7}/_{32}$ or 6.21875
5. $14^{3}/_{8}$ or 14.375
6. $57^{1}/_{2}$ or 57.5
7. $3^{37}/_{64}$ or 3.578125
8. 44
9. $10^{39}/_{64}$ or 10.609375
10. $1^{3}/_{4}$ or 1.75

Exercise E

1. $-5^{7}/_{64}$ or -5.109375
2. $-20^{1}/_{4}$ or -20.25
3. $-40^{9}/_{32}$ or -40 28125
4. $-3^{313}/_{512}$ or -3.6113281
5. $-99^{1}/_{4}$ or -99.25
6. $-4^{1}/_{2}$ or -4.5
7. $-12^{27}/_{64}$ or -12.421875
8. -84
9. $-21^{1}/_{2}$ or -21.5
10. $-14^{1}/_{4}$ or -14.25

Exercise F

1. $^{43}/_{1024}$ or 0.0419921
2. $^{21}/_{4096}$ or 0.0051269
3. $^{43}/_{1024}$ or 0.0419921
4. $^{59}/_{512}$ or 0.1152343
5. $^{227}/_{1024}$ or 0.2216796

Exercise G

1. $-^{3}/_{128}$ or -0.0234375
2. $-^{43}/_{2048}$ or -0.020996
3. $-^{85}/_{1024}$ or -0.0830078
4. $-^{133}/_{1024}$ or -0.1298828
5. $-^{273}/_{1024}$ or -0.2666015