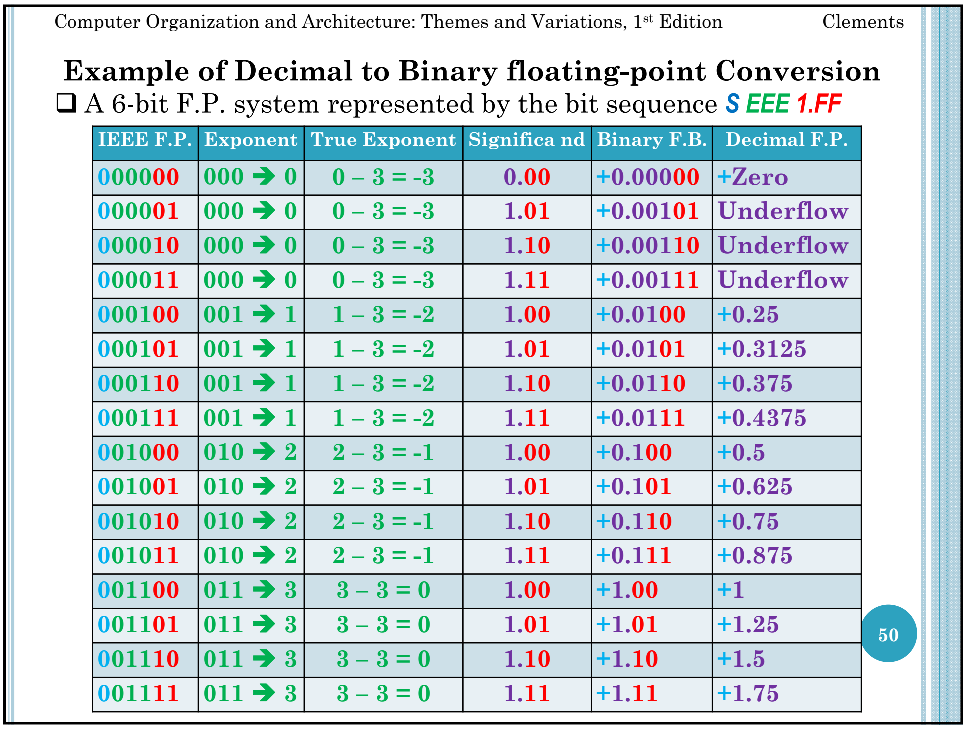
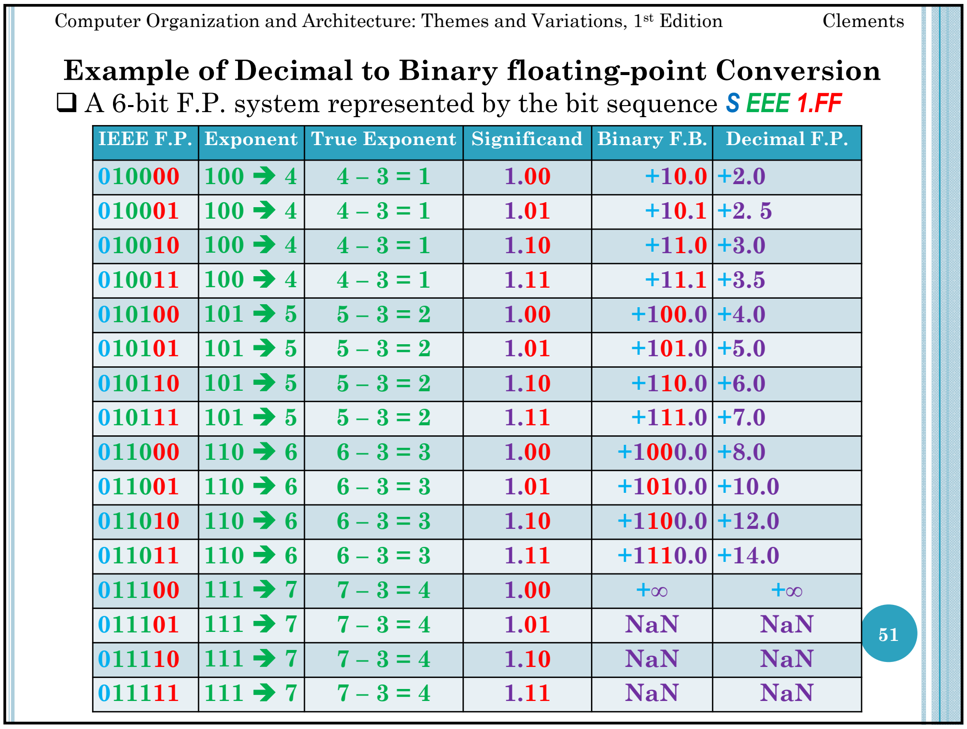


Example of Decimal to Binary floating-point Conversion

- ❑ A 6-bit F.P. system represented by the bit sequence *S EEE 1.FF*

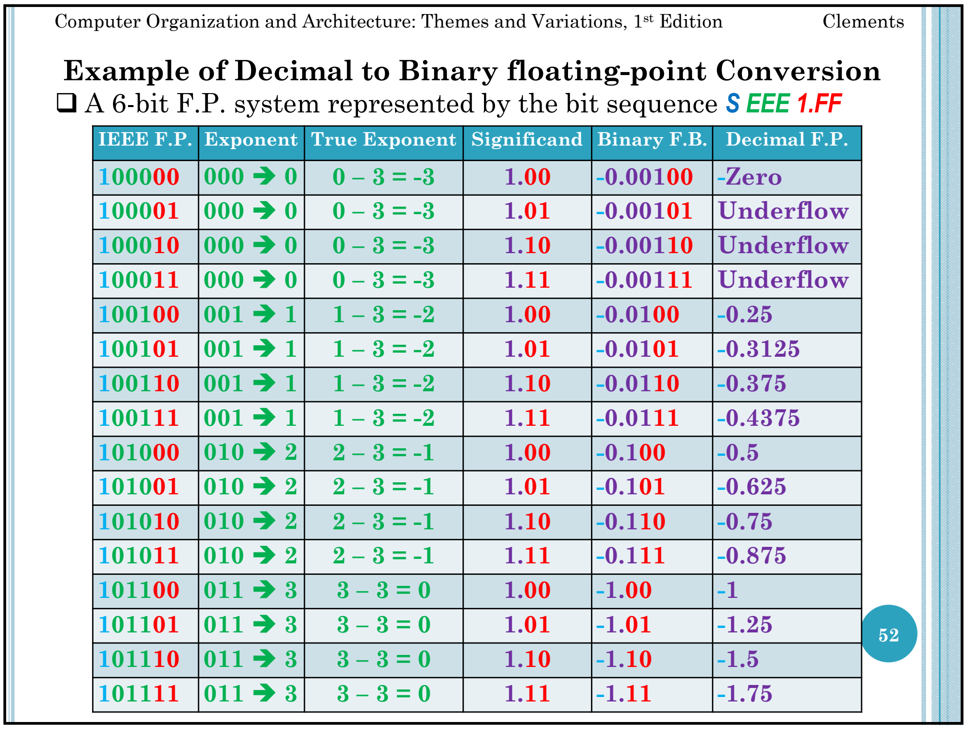
IEEE F.P.	Exponent	True Exponent	Significa nd	Binary F.B.	Decimal F.P.
000000	000 → 0	0 – 3 = -3	0.00	+0.00000	+Zero
000001	000 → 0	0 – 3 = -3	1.01	+0.00101	Underflow
000010	000 → 0	0 – 3 = -3	1.10	+0.00110	Underflow
000011	000 → 0	0 – 3 = -3	1.11	+0.00111	Underflow
000100	001 → 1	1 – 3 = -2	1.00	+0.0100	+0.25
000101	001 → 1	1 – 3 = -2	1.01	+0.0101	+0.3125
000110	001 → 1	1 – 3 = -2	1.10	+0.0110	+0.375
000111	001 → 1	1 – 3 = -2	1.11	+0.0111	+0.4375
001000	010 → 2	2 – 3 = -1	1.00	+0.100	+0.5
001001	010 → 2	2 – 3 = -1	1.01	+0.101	+0.625
001010	010 → 2	2 – 3 = -1	1.10	+0.110	+0.75
001011	010 → 2	2 – 3 = -1	1.11	+0.111	+0.875
001100	011 → 3	3 – 3 = 0	1.00	+1.00	+1
001101	011 → 3	3 – 3 = 0	1.01	+1.01	+1.25
001110	011 → 3	3 – 3 = 0	1.10	+1.10	+1.5
001111	011 → 3	3 – 3 = 0	1.11	+1.11	+1.75



Example of Decimal to Binary floating-point Conversion

- ❑ A 6-bit F.P. system represented by the bit sequence *S EEE 1.FF*

IEEE F.P.	Exponent	True Exponent	Significand	Binary F.B.	Decimal F.P.
010000	100 → 4	4 – 3 = 1	1.00	+10.0	+2.0
010001	100 → 4	4 – 3 = 1	1.01	+10.1	+2. 5
010010	100 → 4	4 – 3 = 1	1.10	+11.0	+3.0
010011	100 → 4	4 – 3 = 1	1.11	+11.1	+3.5
010100	101 → 5	5 – 3 = 2	1.00	+100.0	+4.0
010101	101 → 5	5 – 3 = 2	1.01	+101.0	+5.0
010110	101 → 5	5 – 3 = 2	1.10	+110.0	+6.0
010111	101 → 5	5 – 3 = 2	1.11	+111.0	+7.0
011000	110 → 6	6 – 3 = 3	1.00	+1000.0	+8.0
011001	110 → 6	6 – 3 = 3	1.01	+1010.0	+10.0
011010	110 → 6	6 – 3 = 3	1.10	+1100.0	+12.0
011011	110 → 6	6 – 3 = 3	1.11	+1110.0	+14.0
011100	111 → 7	7 – 3 = 4	1.00	+∞	+∞
011101	111 → 7	7 – 3 = 4	1.01	NaN	NaN
011110	111 → 7	7 – 3 = 4	1.10	NaN	NaN
011111	111 → 7	7 – 3 = 4	1.11	NaN	NaN



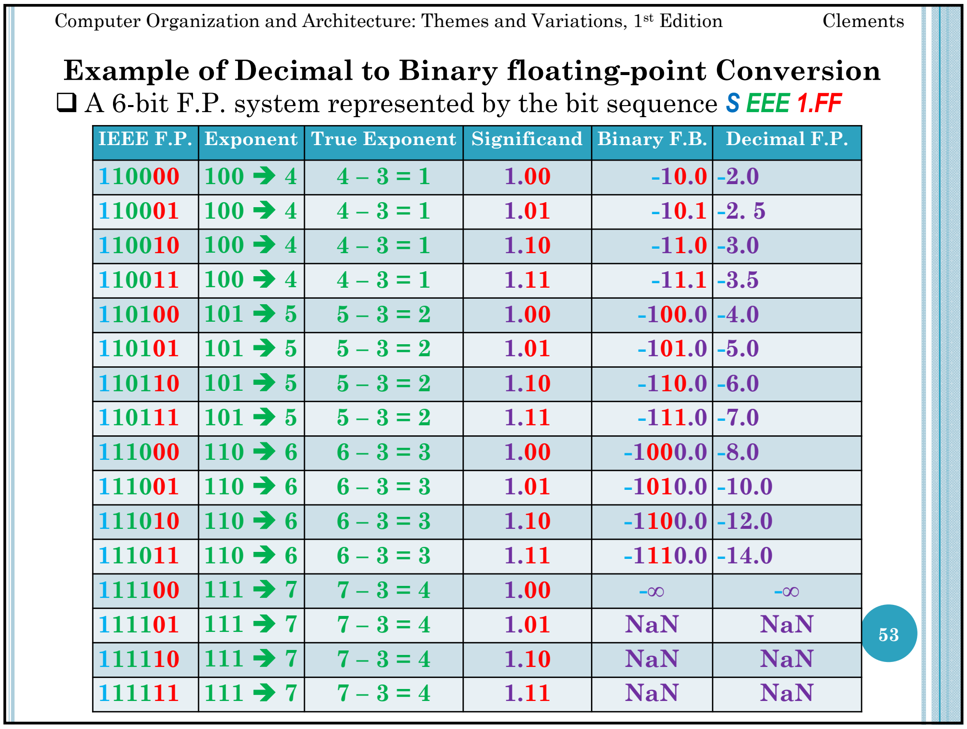
Computer Organization and Architecture: Themes and Variations, 1st Edition Clements

Example of Decimal to Binary floating-point Conversion

- ❑ A 6-bit F.P. system represented by the bit sequence *S EEE 1.FF*

IEEE F.P.	Exponent	True Exponent	Significand	Binary F.B.	Decimal F.P.
100000	000 ➔ 0	0 – 3 = -3	1.00	-0.00100	-Zero
100001	000 ➔ 0	0 – 3 = -3	1.01	-0.00101	Underflow
100010	000 ➔ 0	0 – 3 = -3	1.10	-0.00110	Underflow
100011	000 ➔ 0	0 – 3 = -3	1.11	-0.00111	Underflow
100100	001 ➔ 1	1 – 3 = -2	1.00	-0.0100	-0.25
100101	001 ➔ 1	1 – 3 = -2	1.01	-0.0101	-0.3125
100110	001 ➔ 1	1 – 3 = -2	1.10	-0.0110	-0.375
100111	001 ➔ 1	1 – 3 = -2	1.11	-0.0111	-0.4375
101000	010 ➔ 2	2 – 3 = -1	1.00	-0.100	-0.5
101001	010 ➔ 2	2 – 3 = -1	1.01	-0.101	-0.625
101010	010 ➔ 2	2 – 3 = -1	1.10	-0.110	-0.75
101011	010 ➔ 2	2 – 3 = -1	1.11	-0.111	-0.875
101100	011 ➔ 3	3 – 3 = 0	1.00	-1.00	-1
101101	011 ➔ 3	3 – 3 = 0	1.01	-1.01	-1.25
101110	011 ➔ 3	3 – 3 = 0	1.10	-1.10	-1.5
101111	011 ➔ 3	3 – 3 = 0	1.11	-1.11	-1.75

52



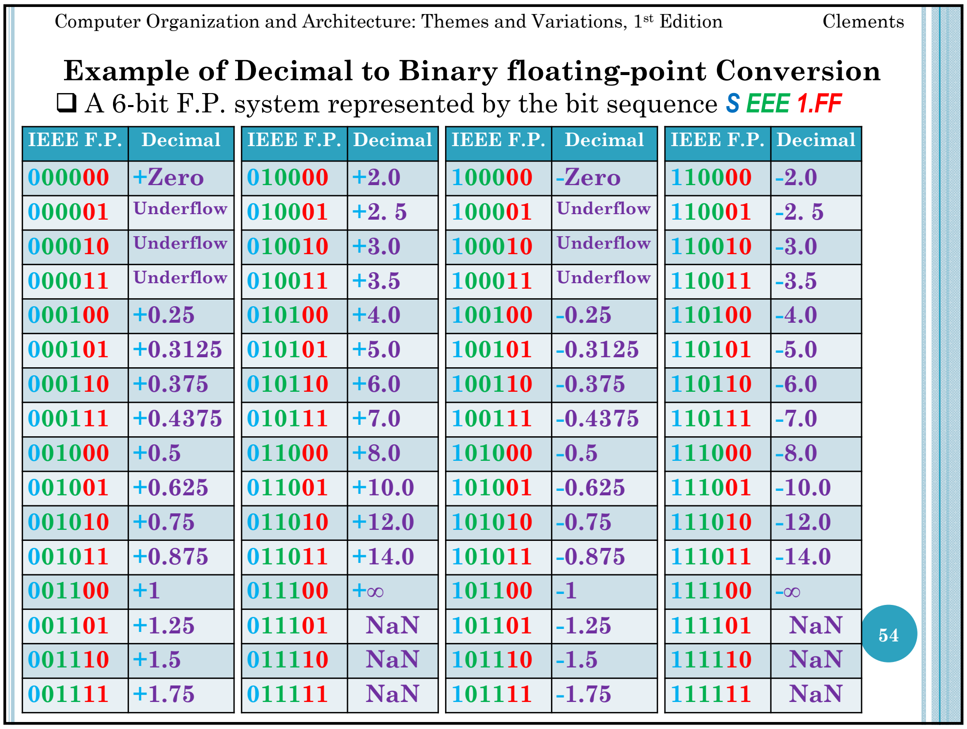
Computer Organization and Architecture: Themes and Variations, 1st Edition Clements

Example of Decimal to Binary floating-point Conversion

- ❑ A 6-bit F.P. system represented by the bit sequence *S EEE 1.FF*

IEEE F.P.	Exponent	True Exponent	Significand	Binary F.B.	Decimal F.P.
110000	100 ➔ 4	4 – 3 = 1	1.00	-10.0	-2.0
110001	100 ➔ 4	4 – 3 = 1	1.01	-10.1	-2. 5
110010	100 ➔ 4	4 – 3 = 1	1.10	-11.0	-3.0
110011	100 ➔ 4	4 – 3 = 1	1.11	-11.1	-3.5
110100	101 ➔ 5	5 – 3 = 2	1.00	-100.0	-4.0
110101	101 ➔ 5	5 – 3 = 2	1.01	-101.0	-5.0
110110	101 ➔ 5	5 – 3 = 2	1.10	-110.0	-6.0
110111	101 ➔ 5	5 – 3 = 2	1.11	-111.0	-7.0
111000	110 ➔ 6	6 – 3 = 3	1.00	-1000.0	-8.0
111001	110 ➔ 6	6 – 3 = 3	1.01	-1010.0	-10.0
111010	110 ➔ 6	6 – 3 = 3	1.10	-1100.0	-12.0
111011	110 ➔ 6	6 – 3 = 3	1.11	-1110.0	-14.0
111100	111 ➔ 7	7 – 3 = 4	1.00	-∞	-∞
111101	111 ➔ 7	7 – 3 = 4	1.01	NaN	NaN
111110	111 ➔ 7	7 – 3 = 4	1.10	NaN	NaN
111111	111 ➔ 7	7 – 3 = 4	1.11	NaN	NaN

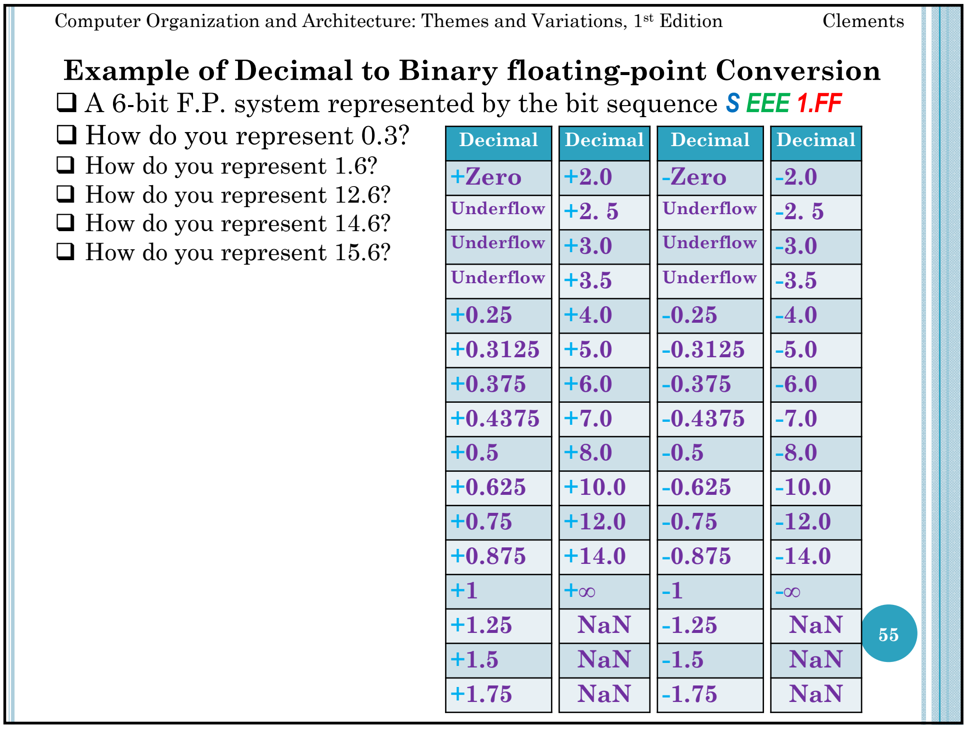
53



Example of Decimal to Binary floating-point Conversion

- ❑ A 6-bit F.P. system represented by the bit sequence *S EEE 1.FF*

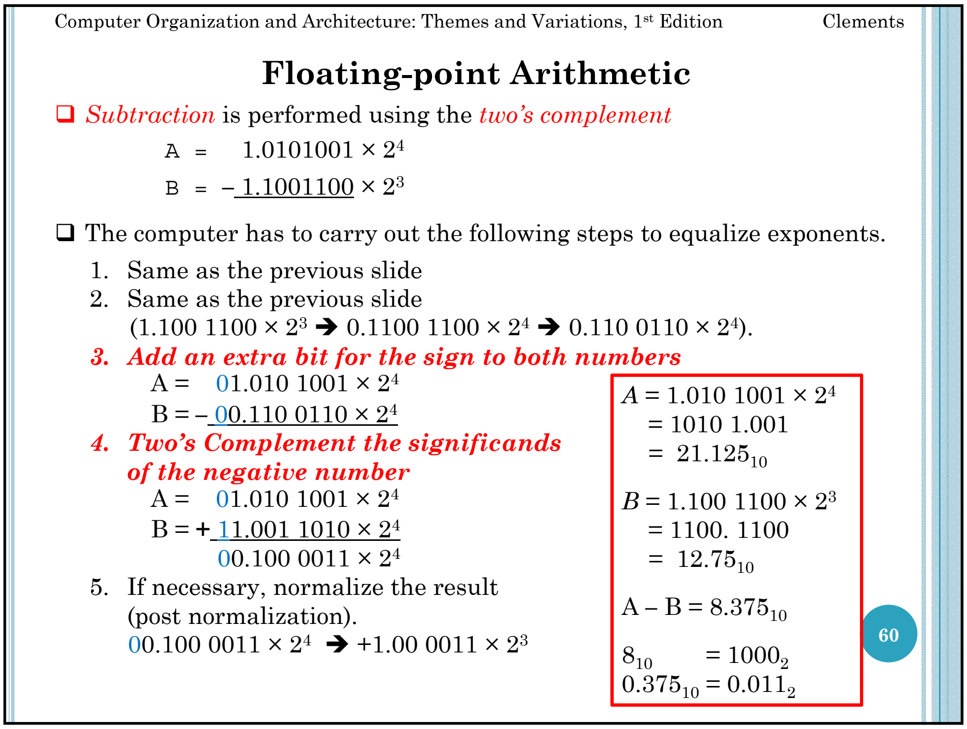
IEEE F.P.	Decimal	IEEE F.P.	Decimal	IEEE F.P.	Decimal	IEEE F.P.	Decimal
000000	+Zero	010000	+2.0	100000	-Zero	110000	-2.0
000001	Underflow	010001	+2. 5	100001	Underflow	110001	-2. 5
000010	Underflow	010010	+3.0	100010	Underflow	110010	-3.0
000011	Underflow	010011	+3.5	100011	Underflow	110011	-3.5
000100	+0.25	010100	+4.0	100100	-0.25	110100	-4.0
000101	+0.3125	010101	+5.0	100101	-0.3125	110101	-5.0
000110	+0.375	010110	+6.0	100110	-0.375	110110	-6.0
000111	+0.4375	010111	+7.0	100111	-0.4375	110111	-7.0
001000	+0.5	011000	+8.0	101000	-0.5	111000	-8.0
001001	+0.625	011001	+10.0	101001	-0.625	111001	-10.0
001010	+0.75	011010	+12.0	101010	-0.75	111010	-12.0
001011	+0.875	011011	+14.0	101011	-0.875	111011	-14.0
001100	+1	011100	+∞	101100	-1	111100	-∞
001101	+1.25	011101	NaN	101101	-1.25	111101	NaN
001110	+1.5	011110	NaN	101110	-1.5	111110	NaN
001111	+1.75	011111	NaN	101111	-1.75	111111	NaN



Example of Decimal to Binary floating-point Conversion

- ❑ A 6-bit F.P. system represented by the bit sequence *S EEE 1.FF*
- ❑ How do you represent 0.3?
- ❑ How do you represent 1.6?
- ❑ How do you represent 12.6?
- ❑ How do you represent 14.6?
- ❑ How do you represent 15.6?

Decimal	Decimal	Decimal	Decimal
+Zero	+2.0	-Zero	-2.0
Underflow	+2. 5	Underflow	-2. 5
Underflow	+3.0	Underflow	-3.0
Underflow	+3.5	Underflow	-3.5
+0.25	+4.0	-0.25	-4.0
+0.3125	+5.0	-0.3125	-5.0
+0.375	+6.0	-0.375	-6.0
+0.4375	+7.0	-0.4375	-7.0
+0.5	+8.0	-0.5	-8.0
+0.625	+10.0	-0.625	-10.0
+0.75	+12.0	-0.75	-12.0
+0.875	+14.0	-0.875	-14.0
+1	+∞	-1	-∞
+1.25	NaN	-1.25	NaN
+1.5	NaN	-1.5	NaN
+1.75	NaN	-1.75	NaN



Floating-point Arithmetic

- ❑ *Subtraction* is performed using the *two's complement*

$$A = 1.0101001 \times 2^4$$
$$B = -\underline{1.1001100} \times 2^3$$

- ❑ The computer has to carry out the following steps to equalize exponents.

1. Same as the previous slide
2. Same as the previous slide
   (1.100 1100 × 2^3 ➔ 0.1100 1100 × 2^4 ➔ 0.110 0110 × 2^4).
3. ***Add an extra bit for the sign to both numbers***

   $A = 01.010\ 1001 \times 2^4$

   $B = -\underline{00.110\ 0110 \times 2^4}$
4. ***Two's Complement the significands of the negative number***

   $A = 01.010\ 1001 \times 2^4$

   $B = +\underline{11.001\ 1010 \times 2^4}$

   $00.100\ 0011 \times 2^4$
5. If necessary, normalize the result (post normalization).

   $00.100\ 0011 \times 2^4$ ➔ $+1.00\ 0011 \times 2^3$

$A = 1.010\ 1001 \times 2^4$
$= 1010\ 1.001$
$= 21.125_{10}$

$B = 1.100\ 1100 \times 2^3$
$= 1100.\ 1100$
$= 12.75_{10}$

$A - B = 8.375_{10}$

$8_{10} = 1000_2$
$0.375_{10} = 0.011_2$